Modeling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives

Gemma Boleda*
Universitat Pompeu Fabra

Sabine Schulte im Walde**
University of Stuttgart

Toni Badia†
Universitat Pompeu Fabra

*We present a study on the automatic acquisition of semantic classes for Catalan adjectives from distributional and morphological information, with particular emphasis on polysemous adjectives. The aim is to distinguish and characterize broad classes, such as qualitative (*gran *'big') and relational (*pulmonar *'pulmonary') adjectives, as well as to identify polysemous adjectives such as* econòmic *('economic | cheap'). We specifically aim at modeling regular polysemy, that is, types of sense alternations that are shared across lemmata. To date, both semantic classes for adjectives and regular polysemy have only been sparsely addressed in empirical computational linguistics.*

Two main specific questions are tackled in this article. First, what is an adequate broad semantic classification for adjectives? We provide empirical support for the qualitative and relational classes as defined in theoretical work, and uncover one type of adjective that has not received enough attention, namely, the event-related class. Second, how is regular polysemy best modeled in computational terms? We present two models, and argue that the second one, which models regular polysemy in terms of simultaneous membership to multiple basic classes, is both theoretically and empirically more adequate than the first one, which attempts to identify independent polysemous classes. Our best classifier achieves 69.1% accuracy, against a 51% baseline.

1. Introduction

Adjectives are one of the most elusive parts of speech with respect to meaning. For example, it is very difficult to establish a broad classification of adjectives into semantic classes, analogous to a broad ontological classification of nouns (Raskin and Nirenburg

* Department of Translation and Language Sciences, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain. E-mail: gemma.boleda@upf.edu.
** E-mail: schulte@ims.uni-stuttgart.de.
† E-mail: toni.badia@upf.edu.

Submission received: 10 December 2008; revised submission received: 16 July 2011; accepted for publication: 5 September 2011. Part of the work reported in this article was done while the first author was a postdoctoral scholar at U. Politècnica de Catalunya and a visiting researcher at U. Stuttgart.

1998). This article tackles precisely this task, that is, the *semantic classification of adjectives*, for Catalan. We aim at automatically inducing the semantic class for an adjective given its linguistic properties, as extracted from corpora and other resources.

The acquisition of semantic classes has been widely studied for verbs (Dorr and Jones 1996; McCarthy 2000; Korhonen, Krymolowski, and Marx 2003; Lapata and Brew 2004; Schulte im Walde 2006; Joanis, Stevenson, and James 2008) and, to a lesser extent, for nouns (Hindle 1990; Pereira, Tishby, and Lee 1993), but, with very few exceptions (Bohnet, Klatt, and Wanner 2002; Carvalho and Ranchhod 2003), not for adjectives. Furthermore, we cannot rely on a well-established classification for adjectives. The classes themselves are subject to experimentation. We will test two different classifications, analyzing the empirical properties of the classes and the problems in their definition.

Another significant challenge is posed by polysemy, or the fact that one and the same adjective can have multiple senses. Different senses may fall into different classes, such that it is no longer possible to identify one single semantic class per adjective. Moreover, many adjectives exhibit similar sense alternations, in a phenomenon known as *regular* or *systematic polysemy* (Apresjan 1974; Copestake and Briscoe 1995). A special focus of the research presented, therefore, is on modeling regular polysemy. As an example of regular polysemy, take for instance the sense alternation for the adjective *econòmic* exemplified in Example (1). *Econòmic*, derived from *economia* ('economy'), can be translated as 'economic, of the economy', as in Example (1a), or as 'cheap', as in Example (1b). As we will see, each of these senses corresponds to a different semantic class in our classifications.

(1) a. recuperació econòmica
recovery economy$_{\text{SUFFIX}}$
'recovery of the economy'

b. pantalons econòmics
trousers economy$_{\text{SUFFIX}}$
'cheap trousers'

Other adjectives exhibit similar sense alternations; for example, *familiar* (derived from *família*, 'family') and *amorós* (derived from *amor*, 'love'), as shown in Example (2).

(2) a. reunió familiar / cara familiar
meeting family$_{\text{SUFFIX}}$ / face family$_{\text{SUFFIX}}$
'family meeting / familiar face'

b. problema amorós / noi amorós
problem love$_{\text{SUFFIX}}$ / boy love$_{\text{SUFFIX}}$
'love problem / lovely boy'

The first senses in Examples (1) and (2) have a transparent relation to the denotation of the deriving noun, as witnessed by the fact that they are translated as nouns in English (*economy*, *family*, *love*), whereas the other senses are translated as adjectives (*cheap*, *familiar*, *lovely*). For each of these adjectives, there is a relationship between the two senses, such that the sense alternations seem to correspond to a productive semantic process along the lines of Example (3) (Raskin and Nirenburg 1998, schema (43), page 173).

(3) PERTAINING TO [noun meaning] → CHARACTERISTIC OF [noun meaning]

Because of the systematic semantic relationship between the two senses of these adjectives, they constitute an instance of regular polysemy. In this article, therefore, we not only address the acquisition of semantic classes, but also the *acquisition of polysemy*: Our goal is to determine, for a given adjective, whether it is monosemous or polysemous, and to which class(es) it belongs. Note that we are not dealing with individual sense alternations, as related work on sense induction does (Schütze 1998; McCarthy et al. 2004; Brody and Lapata 2009), but with sense alternation *types*, that systematically hold across different lemmata. Thus, the present research is at the crossroad between sense induction and lexical acquisition.

Regularities in sense alternations are pervasive in human languages, and they are probably favored by the properties of human cognition (Murphy 2002). Regular polysemy has been studied in theoretical linguistics (Apresjan 1974; Pustejovsky 1995) and in symbolic approaches to computational semantics (Copestake and Briscoe 1995). It has received little attention in empirical computational semantics, however. This is surprising, given the amount of work devoted to sense-related tasks such as Word Sense Disambiguation (WSD). In WSD (see Navigli [2009] for an overview) sense ambiguities are almost exclusively modeled for each individual lemma, despite the ensuing sparsity problems (Ando [2006] is an exception). Properly modeling regular polysemy, therefore, promises to improve computational semantic tasks such as WSD and sense discrimination.

This article has the goal of finding a computational model that responds to the theoretical and empirical properties of regular polysemy. In this direction, we test two alternative approaches. We first model polysemy in terms of independent classes to be separately acquired (e.g., an adjective with two senses a_i and b_i belongs to a class AB defined independently of classes A and B), and show that this model is not adequate. A second approach, which posits that polysemous adjectives simultaneously belong to more than one class (e.g., an adjective with two senses a_i and b_i belongs to both class A and class B), is more successful. Our best classifier achieves 69.1% accuracy against a 51% baseline, which is satisfactory, considering that the estimated upper bound (human agreement) for this task is 68%. We discuss pros and cons of the two models described and ways to overcome their limitations.

In the following, we first review related work (Section 2) and linguistic aspects of adjective classification (Section 3), then present the two acquisition experiments (Sections 4 and 5), and finish with a general discussion (Section 6) and some conclusions and directions for future research (Section 7).

2. Related Work

As mentioned in the Introduction, there has been very little research in the semantic classification of adjectives. We know of only two articles on specifically this topic: Carvalho and Ranchhod (2003) used adjective classes similar to the ones explored here to disambiguate between nominal and adjectival readings in Portuguese. Adjective information, manually coded, served to establish constraints in a finite-state transducer part-of-speech tagger. Actually, POS tagging was also the initial motivation for the present research, as adjective–noun and adjective–verb (participle) ambiguities cause most difficulties to both humans and machines in languages such as English, German, and Catalan (Marcus, Santorini, and Marcinkiewicz 1993; Brants 2000; Boleda 2007). Bohnet, Klatt, and Wanner (2002) also has similar goals to the present research, as it is aimed at automatically classifying German adjectives. However, the classification

used is not purely semantic, polysemy is not taken into account, and the evidence and techniques used are more limited than the ones used here.

Other research on adjectives within computational linguistics is oriented toward different goals than ours. Yallop, Korhonen, and Briscoe (2005) tackle syntactic, not semantic classification, akin to the acquisition of subcategorization frames for verbs. Another relevant line of research pursues WSD. Justeson and Katz (1995) and Chao and Dyer (2000) showed that adjectives are a very useful cue for disambiguating the sense of the nouns they modify. Adjective classes could be further exploited in WSD in at least two respects: (1) to establish an inventory of adjective senses (if polysemous instances are correctly detected; this is where sense induction and our own work fits in), and (2) to exploit class-based properties for the disambiguation, similar to related work on verb classes (Resnik 1993; Prescher, Riezler, and Rooth 2000; Kohomban and Lee 2005).

The application where adjectives have received most attention, however, is Opinion Mining and Sentiment Analysis (Pang and Lee 2008), as adjectives are known to convey much of the evaluative and subjective information in language (Wiebe et al. 2004). The typical goal of this kind of study has been to identify subjective adjectives and their orientation (positive, neutral, negative). This type of research, from pioneering work by Hatzivassiloglou and colleages (Hatzivassiloglou and McKeown 1993, 1997; Hatzivassiloglou and Wiebe 2000) to current research (de Marneffe, Manning, and Potts 2010), has thus focused on *scalar adjectives*, that is, adjectives like *good* and *bad*, which can be translated into values that can be ordered along a scale. These adjectives typically enter into antonymy relations (the semantic relation between *good* and *bad*), and in fact antonymy is the main organizing criterion for adjectives in WordNet (Miller 1998), the most widely used semantic resource in NLP. However, when examining a large scale lexicon, it becomes immediately apparent that there are many other types of adjectives that do not easily fit in a scale-based or antonymy-based view of adjectives (Alonge et al. 2000). Some examples are *pulmonary*, *former*, and *foldable*. It is not clear, for instance, whether it makes sense to ask for an antonym of *pulmonary*, or to establish a "foldability" scale for *foldable*. These adjectives need a different treatment, and they are treated in terms of different semantic classes in this article.

The semantic properties of adjectives can also be exploited in advanced NLP tasks and applications such as Question Answering, Dialog Systems, Natural Language Generation, or Information Extraction. For instance, from a sentence like *This maimai is round and sweet*, we can quite safely infer that the (invented) object *maimai* is a physical object, probably edible. This type of process could be exploited in, for instance, Information Extraction and ontology population, although to our knowledge this possibility has received but little attention (Malouf 2000; Almuhareb and Poesio 2004).

As for polysemy, previous approaches to the automatic acquisition of semantic classes have mostly disregarded the problem, by biasing the experimental material to include monosemous words only, or by choosing an approach that ignores polysemy (Hindle 1990; Merlo and Stevenson 2001; Schulte im Walde 2006; Joanis, Stevenson, and James 2008). There are a few exceptions to this tradition, such as Pereira, Tishby, and Lee (1993), Rooth et al. (1999), and Korhonen, Krymolowski, and Marx (2003), who used soft clustering methods for multiple assignment to verb semantic classes (see Section 4.5).

There is very little related work in empirical computational semantics in modeling regular polysemy. A pioneering piece of research is Buitelaar (1998), which tried to account for regular polysemy with the CoreLex resource. CoreLex, building on the Generative Lexicon theory (Pustejovsky 1995), groups WordNet senses into 39 "basic

types" (broad ontological categories). In CoreLex, each word is associated to a polysemy class, that is, the set of all basic types its synsets belong to. Some of these polysemy classes constitute instances of regular polysemy, as recently explored in Utt and Padó (2011).

Lapata (2000, 2001) also addresses regular polysemy in the Generative Lexicon framework. This work attempts to establish all the possible meanings of adjective-noun combinations, and rank 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, for example, *easy text*, that is usually equivalent to *text that is easy to read*). Thus, the focus is on the meaning of adjective-noun combinations, not on that of adjectives alone as in the present research.

3. Basis for a Semantic Classification of Adjectives

Adjective classes in our definition are broad classes of lexical meaning. We will present lexical acquisition experiments in which, given the evidence found in corpora and other lexical resources, a semantic class can be assigned to a given adjective. For this purpose, two preconditions are required:

(a) a classification that establishes the number and characteristics of the target semantic classes;

(b) a stable relation between observable features and each semantic class.

There is no established semantic classification for adjectives in computational linguistics that we can use and, therefore, one subgoal of the research is to establish the classification in the first place, addressing (a), and exploiting the morphology–semantics and syntax–semantics interfaces for acquisition, addressing (b). We are thus facing a highly exploratory endeavor, and we do not regard the classifications we use as final. We test two different classifications: an initial classification, based on the literature, for the experiments reported in Section 4, and an alternative classification, for the experiments reported in Section 5. We next turn to presenting the two tested classifications.

3.1 Initial Classification

In the acquisition experiments reported in Section 4, we distinguish between **qualitative**, **intensional**, and **relational** adjectives, which have the following properties (Miller 1998; Raskin and Nirenburg 1998; Picallo 2002; Demonte 2011).

Qualitative adjectives. These are prototypical adjectives like *gran* ('big') or *dolç* ('sweet'), including scalar adjectives, which denote attributes or properties of objects. Adjectives in this class tend to be gradable and comparable (see Examples (4a–4b)). They are characterized by exhibiting the greatest variability with respect to their syntactic behavior: In Catalan, they can act as predicates in copular sentences and other constructions (Examples (4c–4d)), and they can typically act as both pre- and post-nominal modifiers (Examples (4e–4f)). When an adjective modifies a head noun in pre-nominal position,

the interpretation is usually nonrestrictive, as shown by the fact that they can modify proper nouns (Example (4e)).

(4) a. Taula molt gran / grandíssima
Table very big / $\text{big}_{\text{SUPERLATIVE}}$
'Very big table'

b. Aquesta taula és més gran que aquella
This table is more big than that
'This table is bigger than that one'

c. Aquesta taula és gran
This table is big
'This table is big'

d. Aquesta taula la veig massa gran
This table $\text{it}_{\text{OBJ-CL-FEM}}$ $\text{see}_{pres-1stp-sg}$ too big
'This table seems to me to be too big'

e. La gran Diana va seguir cantant
The great Diana PAST-AUX continue singing.
'Great Diana continued singing.'

f. Van portar una taula gran
PAST-AUX bring a table big
'They brought in a big table'

Intensional adjectives. These are adjectives like *presumpte* ('alleged') or *antic* ('former'), which according to formal semantics denote second-order properties (Montague 1974, and subsequent work). Most intensional adjectives modify nouns in pre-nominal position only (Example (5a)), and they cannot functionally act as predicates (Example (5b)). They are also typically not gradable (Example (5c)).

(5) a. El Joan és el presumpte assassí
The Joan is the alleged murderer
'Joan is the alleged murderer'

b. #El Joan és presumpte
The Joan is alleged
'#Joan is alleged'

c. #Més presumpte assassí / #presumptíssim assassí
More alleged murderer / $\text{alleged}_{\text{SUPERLATIVE}}$ murderer
'#More/very alleged murderer'

Intensional adjectives like *presumpte* may appear in any order with respect to qualitative adjectives, as in Example (6). The order, however, affects interpretation: Example (6a) entails that the referent of the noun phrase is young, whereas Example (6b) does not (McNally and Boleda 2004).

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

b. presumpte jove assassí
'alleged young murderer'

Relational adjectives. Adjectives such as *pulmonar, estacional, botànic* ('pulmonary, seasonal, botanical') denote a relationship to an object (in the mentioned examples, LUNG, SEASON, and PLANT objects). Most of them are denominal (e.g., *pulmonar* is derived from *pulmó*, 'lung') and can only modify nouns post-nominally (see Example (7a)). Also, contrary to qualitative adjectives, they are not gradable (Example (7b)) and act as predicates only under very restricted circumstances (Example (7c) vs. (7d)). If other adjectives or modifiers co-occur with relational adjectives, these occur *after* the adjective (Example (7e)). We will say relational adjectives are *adjacent* to the head noun.

(7) a. Tenia una malaltia pulmonar / #pulmonar malaltia
Had a disease pulmonary / pulmonary disease
'He/she had a pulmonary disease'

b. #Malaltia molt pulmonar / pulmonaríssima
Disease very pulmonary / pulmonary$_{\text{SUPERLATIVE}}$
#'Very pulmonary disease'

c. La decisió europea → ??Aquesta decisió és europea
The decision European → This decision is European
'The European decision → ??This decision is European'

d. La tuberculosi pot ser pulmonar
The tuberculose can be pulmonary
'Tuberculose can be pulmonary'

e. inflamació pulmonar greu / #inflamació greu pulmonar
inflamation pulmonary serious / inflamation serious pulmonary
'serious pulmonary inflammation'

Table 1 summarizes the properties just explained. Our goal is to use these properties to induce the semantic class of adjectives. For instance, if an adjective is denominal, appears almost exclusively in postnominal position, and is strictly adjacent to the head noun, we predict that it is relational. In the experiments reported in Sections 4 and 5, we

Table 1
Initial classification: Linguistic properties of qualitative, intensional, and relational adjectives.

	Qualitative	**Intensional**	**Relational**
	gran ('big')	*presumpte* ('alleged')	*pulmonar* 'pulmonary'
Property			
predicative	+	−	restricted
gradable/comparable	+	−	−
position with respect to head noun	both	pre-nom.	post-nom.
adjacent	−	−	+
denominal	−	−	+

extract data related to these and other properties of adjectives from linguistic resources, and use them as features in machine learning experiments.

3.2 Alternative Classification

In the acquisition experiments reported in Section 5, we distinguish between qualitative, relational, and event-related adjectives. The classification presented in Section 3.1 is thus altered in two ways: (1) The intensional class is dropped. (2) A new class, that of event-related adjectives, is added to the classification. The reasons for these changes will become clear in the discussion of the experiments in Section 4. Here, we describe the new class and provide a summary table of the alternative classification.

Event-related adjectives. Adjectives such as *exportador, promès, resultant* ('exporting, promised, resulting') denote a relationship to an event, in this case, EXPORT, PROMISE, and RESULT events, respectively. Most of them are deverbal. Like relational adjectives, they are typically nongradable (see Example (8a)) and prefer the postnominal position when modifying nouns (Example (8b)). Like qualitative adjectives, they typically can act as predicates (Example (8c)).

(8) a. És un país {exportador / #molt exportador} de petroli
Is a country {exporting / very exporting} of oil
'It is an oil exporting / #very exporting country'

b. #exportador país
'exporting country'

c. Aquest país és exportador
This country is exporting
'This is an exporting country'

Table 2 summarizes the properties of the alternative classification (for a more thorough discussion of previous research on the semantics of adjectives and more motivation for the classification, see Boleda [2007]). For comparison, we will briefly outline the treatment of adjectives in WordNet (Miller 1998; Alonge et al. 2000). As

Table 2
Alternative classification: Linguistic properties of qualitative, event-related, and relational adjectives.

	Qualitative	**Event-related**	**Relational**
	gran ('big')	*exportador* ('exporting')	*pulmonar* 'pulmonary'
Property			
predicative	+	+	restricted
gradable/comparable	+	typically not	–
position with respect to head noun	both	post-nom.	post-nom.
adjacent	–	–	+
derivational type	non-derived	deverbal	denominal

mentioned in Section 2, the main semantic relation around which adjectives are organized in WordNet is antonymy. Also as explained, however, not all adjectives have antonyms. This is solved in WordNet by the use of indirect antonyms (e.g., *swift* and *slow* are indirect antonyms, through the semantic similarity between *swift* and *fast*). Still, indirect antonymy only applies to a small subset of the adjectives in WordNet (slightly over 20% in WordNet 1.5). Therefore, some kinds of adjectives receive a differentiated treatment.

Specifically, two main kinds of adjectives are distinguished in WordNet: (1) Descriptive adjectives, akin to our qualitative adjectives, which are organized around antonymy (descriptive adjectives, however, include intensional adjectives). (2) Relational adjectives, as defined in this article, for which two different solutions are adopted. If a suitable antonym can be found for a given relational adjective (antonym in a broad sense; in Miller [1998, page 60], *physical* and *mental* are considered antonyms), it is treated in the same way as a descriptive adjective. Otherwise, it is linked through a PERTAIN-TO pointer to the related noun. In addition, a subclass of descriptive adjectives, having the form of past or present participles, is distinguished, and also receives a hybrid treatment. Those that can be accommodated to antonymy are treated as descriptive adjectives (*laughing–unhappy*, through the similarity between *laughing* and *happy*). Those which cannot are linked to the source verb through a PRINCIPAL-PART-OF pointer. Our event-related class includes not only past and present participles, but other types of deverbal adjectives. Thus, most of the classes used in this article are to some extent backed up by the organization of adjectives in WordNet.

3.3 The Role of Polysemy

As explained in the Introduction, some adjectives are *polysemous* such that each sense falls into a different class of the classifications just presented. Consider for instance the adjective *econòmic* in Example (1), repeated here as Example (9) for convenience. The two main senses of *econòmic* instantiate the relational (sense in Example (9a)) and the qualitative class (sense in Example (9b)), respectively.

(9) a. anàlisi econòmica
'economic analysis'

b. pantalons econòmics
'cheap trousers'

Crucially for our purposes, in each of the senses the adjective exhibits the properties of each of the associated classes. When used as a relational adjective, it is not gradable and cannot be used in a pre-nominal position (Example (10)). When used as a qualitative adjective, it is gradable and it can be used predicatively (see Example (11)). In the experiments that follow, we aim at capturing this hybrid behavior.

(10) a. #L'anàlisi molt econòmica de les dades
The-analysis very economic of the data
'#The very economic analysis of the data'

b. #Va dur a terme una econòmica anàlisi
'PAST-AUX bring to term an economic analysis
'#He/she carried out an economic analysis'

(11) Aquests pantalons són molt econòmics!
These trousers are very economic!
'These trousers are very cheap!'

Cases of regular polysemy between the intensional and qualitative classes also exist, as illustrated in Examples (12) and (13). *Antic* has two major senses, a qualitative one (equivalent to 'old, ancient') and an intensional one (equivalent to 'former'). Note again that, when used in the intensional sense, it exhibits properties of the intensional class: It appears pre-nominally (Example (13a)) and is not gradable (Example (13b)).

(12) a. edifici antic
building ancient
'ancient building'

b. edifici molt antic
building very ancient
'very ancient building'

(13) a. antic president
ancient president
'former president'

b. #molt antic president
very ancient president
'#very former president'

The new class in the alternative classification, that of event-related adjectives, also introduces regular polysemy, specifically, between event-related and qualitative adjectives, as illustrated in Examples (14) and (15). The participial adjective *sabut* ('known') has an event-related sense, corresponding to the verb *saber* ('know'), and a qualitative sense that can be translated as 'wise'. Likewise, the deverbal adjective *cridaner* derived from *cridar* ('to shout') alternates between an event-related sense and a qualitative sense.

(14) problema sabut / home sabut
problem known / man known
'known problem / wise man'

(15) noi cridaner / camisa cridanera
boy $\text{shout}_{\text{SUFFIX}}$ / shirt attention-gaining
'boy who shouts a lot / attention-gaining shirt'

Examples (14) and (15) represent cases of regular polysemy because, as can be drawn from the translations, there is a systematic shift from a transparent relation with the event to a quality that bears a more distant relation to the event. In the case of *sabut* the relation is clear (if a man knows a lot, he is wise); in the case of *cridaner*, a shirt qualifies for the adjective if it is for instance loud-colored or has an eccentric cut, such that it gains the attention of people, as shouting does.

In this article, we only consider types of polysemy that cut across the classification pursued. Other kinds of polysemy that have traditionally been tackled in the literature will not be considered. For instance, we will not be concerned with the polysemy illustrated in Example (16), which arguably has more to do with the semantics of the

modified noun than that of the adjective (Pustejovsky 1995). Both of the uses of *trist* ('sad') illustrated in Example (16) fall into the qualitative class, so, contrary to the work by Lapata (2000, 2001) cited previously, we do not treat the adjective as polysemous in the context of the present experiments.

(16) noi trist / pel·lícula trista
boy sad / film sad
'sad boy / sad film'

4. First Model: Polysemous Adjectives Constitute Independent Classes

Given the hybrid behavior of polysemous adjectives explained in Section 3, we can expect that they behave differently from adjectives in the basic classes. For instance, adjectives polysemous between a qualitative and a relational use should exhibit more evidence for gradability than pure relational adjectives, but less than pure qualitative adjectives. In this view, polysemous adjectives belong to a class, for instance, the qualitative-relational class, that is distinct from both the qualitative and the relational classes, typically exhibiting feature values that are in between those of the basic classes. In this section, we report on experiments testing precisely this model for regular polysemy. We will therefore distinguish between five types of adjectives: qualitative, intensional, relational, polysemous between a qualitative and an intensional reading (intensional-qualitative), and polysemous between a qualitative and a relational reading (qualitative-relational). There is one polysemous class missing (intensional-relational). No cases of polysemy between intensional and relational adjectives were observed in our data.

Recall from the previous sections that we cannot reuse an established classification, and that there is virtually no previous work on the automatic semantic classification of adjectives. The present experiments also aim at testing the overall enterprise of inducing semantic classes from distributional properties for adjectives. Given the exploratory nature of the experiment, we use clustering, an unsupervised technique, to uncover natural groupings of adjectives and test to what extent these correspond to the classes described in the literature.

4.1 Data and Gold Standard

The experiments reported in this section are based on an eight million word fragment of the CTILC corpus (*Corpus Informatitzat de la Llengua Catalana*; Rafel 1994), developed at the *Institut d'Estudis Catalans*. Each word is associated with its lemma, part of speech, and inflectional features, as well as syntactic function. Lemma and morphological information have been manually checked. We automatically added syntactic information with CatCG (Alsina et al. 2002). CatCG is a shallow parser that assigns one or more syntactic functions to each word. In the case of the adjective, CatCG distinguishes between (1) predicate of a copular sentence; (2) predicate in another construction; (3) pre-nominal modifier; (4) post-nominal modifier. As no full dependencies are indicated, the head noun can only be identified with heuristics.

In the experiments, we cluster all adjectives occurring more than ten times in the corpus (a total of 3,521 lemmata), and analyze the results using a subset of the data. This is a randomly chosen 101-lemma gold standard (available in the Appendix). Fifty

lemmata were chosen token-wise and 50 type-wise to balance high-frequency and low-frequency adjectives (one lemma was chosen with both methods, so the repetition was removed). Two lemmata were added in a post-hoc fashion, as explained subsequently.

The lemmata were annotated by four doctoral students in computational linguistics. The task of the judges was to assign each lemma to one of the five classes (qualitative, intensional, relational, qualitative-intensional, and qualitative-relational). The instructions for the judges included information about all linguistic characteristics discussed in Section 3, including syntactic and semantic characteristics.

The judges had a moderate degree of agreement, comparable to that obtained in other tasks on semantics or discourse, inter-annotator scores ranging between $\kappa = 0.54$ and 0.64 (see Artstein and Poesio [2008] for a discussion of agreement measures for computational linguistics). For comparison, Véronis (1998) reported a mean pair-wise weighted $\kappa = 0.43$ for a word sense tagging task in French; and Merlo and Stevenson (2001) obtained $\kappa = 0.53$–0.66 for the task of classifying English verbs as unergative, unaccusative, or object-drop. Poesio and Artstein (2005) report κ values of 0.63–0.66 (0.45–0.50 if a trivial category is dropped) for the tagging of anaphoric relations. Our judges reported difficulties in tagging particular kinds of adjectives, such as deverbal adjectives. This issue will be retaken in Section 4.5.

No intensional adjectives were identified in the data by the judges, and only one intensional-qualitative adjective was identified. Two intensional lemmata were manually added to be able to minimally track the class. This is clearly insufficient for a quantitative approach, however, so the intensional class is dropped in the alternative classification. It is striking that intensional adjectives, which have traditionally been the focus of formal semantic approaches to the semantics of adjectives, constitute a very small class (less than a dozen lemmata are mentioned in the reviewed literature).

4.2 Features

We use two sets of distributional features to model adjective behavior: on the one hand, theoretically motivated features (**theoretical features** for short); on the other hand, features that encode the part-of-speech distribution of a four-word window around the adjective (**POS features**). The former provide a theoretically informed model of adjectives, because they are cues to the properties of each class as described in the literature. The latter are meant to provide a theory-independent representation of adjectives, to test to what extent the structures obtained with theoretical and POS features are similar. Both sets of features take a narrow context into account (at most five words to each side of the adjective), because of the limited syntactic behavior of adjectives.

4.2.1 Theoretical Features. Theoretical features model the syntactic and semantic properties of the classes described in Section 3. The features used, together with their mean and standard deviation values (computed on all 3,521 adjectives), are summarized in Table 3. A feature value $v_{a,i}$ for an adjective lemma a and a feature i corresponds to the proportion of the occurrences in which i is observed for adjective a over all occurrences of a (see Equation (1); f stands for absolute frequency).

$$v_{a,i} = \frac{f(a,i)}{f(a)} \tag{1}$$

Table 3 is the translation of Table 1 into shallow cues that can be extracted from a corpus. The mean values in Table 3 are very low, which points to the sparseness of theoretically defined properties such as predicativity or gradability, at least in written texts (oral corpora would presumably yield different values). Also note that standard deviations are higher than mean values, which indicates a high variability in the feature values, something that will be exploited for classification.

From the discussion in Section 3, the following predictions with respect to the semantic features can be made.

(1) In comparison with the other classes, qualitative adjectives should have higher values for features *gradable, comparable, copular, predicative*, middle values for feature *prenominal*, and low values for feature *adjacent*.

(2) Relational adjectives should have an almost opposite distribution, with very low values for all features except for *adjacent*.

(3) Intensional adjectives should exhibit very low values for all features except for *pre-nominal*, for which a very high value is expected.

(4) With respect to polysemous adjectives, it can be foreseen that their feature values will be in between those of the basic classes. For instance, an adjective that is polysemous between a qualitative and a relational reading should exhibit a higher value for feature *gradable* than a monosemous relational adjective, but a lower value than a monosemous qualitative adjective.

Figure 1 shows that the predictions just outlined are met to a large extent, showing that the empirical (corpus) data support the theoretical predictions. This graph represents the value distribution of each feature in the form of boxplots. In the boxplots, the rectangles have three horizontal lines, representing the first quartile, the median, and the third quartile, respectively. The dotted line at each side of the rectangle stretches to the minimum and maximum values, at most 1.5 times the length of the rectangle. Values that are outside this range are represented as points and termed *outliers* (Verzani 2005). Note that the scale in Figure 1 does not range from 0 to 1; this is because the data are standardized, as will be explained subsequently.

Table 3
Theoretical features. The mean and SD values are computed on all clustered adjectives. Feature *copular* accounts for predicative constructions with the copula verbs *ser, estar* ('be'). Feature *predicative* accounts for other predicative constructions, such as Example (4d).

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
predicative	predicate syntactic tag	0.03	0.06
pre-nom	pre-nominal modifier syntactic tag	0.04	0.08
adjacent	first adjective in a series of two or more	0.03	0.05

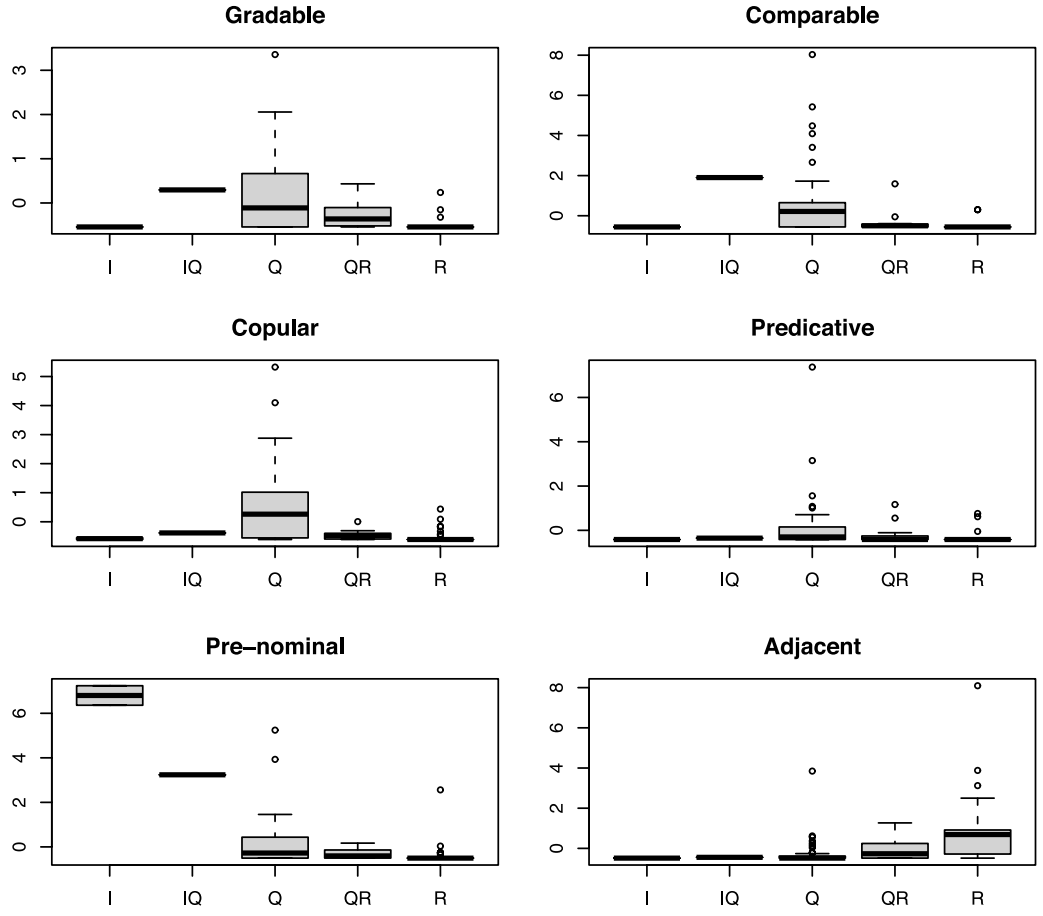


Figure 1
Theoretical features: Feature value distribution in the gold standard. Class labels: I = intensional; IQ = polysemous between intensional and qualitative; Q = qualitative; QR = polysemous between qualitative and relational; R = relational.

The differences in value distributions, although significant,[1] are not sharp, as most of the ranges in the boxes overlap. This affects mainly polysemous classes: Although they show the tendency predicted—exhibiting values that are in between those of the basic classes—they do not present clearly distinct values. The clustering results will be affected by this distribution, as will be discussed in Section 4.5.

4.2.2 POS Features. POS features encode the part-of-speech distribution of a four-word window around the adjective, providing a theory-independent representation of the linguistic behavior of adjectives. To avoid data sparseness, we encode possible POS for each position as a different feature. For instance, for an occurrence of *alta* ('tall') as in Example (17a), the representation would be as in Example (17b). In the example, the target adjective is in boldface, and the relevant word window is in italics. Negative numbers

1 Tested by 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. The test yields p-values lower than 0.05 (*predicative*), 0.01 (*comparable, pre-nominal, adjacent*), and 0.001 (*gradable, copular*), respectively.

Table 4
POS features. The mean and SD values are computed on all clustered adjectives.

Feature	Mean	SD	Feature	Mean	SD
−1 noun	0.52	0.25	−2 preposition	0.13	0.09
+1 punctuation	0.42	0.15	−1 adverb	0.10	0.11
−2 determiner	0.39	0.20	−1 verb	0.08	0.11
+2 determiner	0.24	0.13	−1 determiner	0.06	0.10
+1 preposition	0.21	0.15	+1 noun	0.06	0.10

indicate positions to the left, positive ones positions to the right. The representation in Example (17b) corresponds to the parts of speech of *és*, *més*, *que*, and *la*, respectively.

(17) a. la Bruna *és més* **alta** *que* *l'*Angelina
the Bruna is more tall than the-Angelina
'Bruna is taller than Angelina'

b. *-2 verb, -1 adverb, +1 conjunction, +2 determiner*

Feature values are defined as in theoretical features (see Equation (1)). The ten features with the overall highest mean value in our data (among a total of 36 features) are listed in Table 4. Note that the mean values are much higher for the POS features (Table 4) than for the theoretical features (Table 3), as theoretical features are much sparser.

4.3 Clustering Algorithm and Parameters

We use the *k*-means clustering algorithm (see Kaufman and Rousseeuw [1990] and Everitt, Landau, and Leese [2001] for comprehensive introductions to clustering).[2] 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) and the syntactic classification of Catalan verbs (Mayol, Boleda, and Badia 2005). Also, it performs hard clustering, which is adequate for our purposes (recall from Section 4.1 that we model polysemy in terms of separate classes). Additional experiments with other clustering methods yielded similar results: We tested two hierarchical and one flat algorithm, one of them agglomerative and the other two partitional, with several clustering criteria, always using the cosine distance measure.

K-means is a flat, partitional algorithm that aims at minimizing the overall distance from objects to their centroids (mean vectors of each cluster), which favors globular cluster structures. 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 reassigned to the cluster with the nearest centroid. The centroids are recomputed, and the process is iterated until no further changes take place, or a pre-specified number of times (20 in our case). Equation (2) shows the formula for the clustering criterion, where k is the total number of clusters and l are the lemmata in each cluster $c_1, \ldots, c_k$. To avoid the

2 More specifically, because we are using the cosine measure, the algorithm is spherical *k*-means (Dhillon and Modha 2001). All the experiments were performed with the CLUTO toolkit (Karypis 2002).

influence of the initial partition on the final structure, the whole experiment is repeated several times (25 in our case) with different random partitions, and the partition that better satisfies the clustering criterion is chosen.

$$minimize \sum_{i \in k} \sum_{l \in c_i} cos(l, centroid(c_i)) \quad (2)$$

We experimented with two representations of the feature values: raw and standardized proportions. In clustering, features with higher mean and standard deviation values tend to dominate over more sparse features. Standardization smooths the differences in the strengths of features. We standardize to z-scores, so that all features have mean 0 and standard deviation 1. As the most interpretable results were obtained with standardized values, we will restrict the discussion in the next section to the results obtained with standardized values.

4.4 Results

The discussion focuses on the cluster analyses with three and five clusters because our basis is three classes (intensional, qualitative, and relational) and we consider a total of five classes (basic classes plus polysemous classes: intensional-qualitative and qualitative-relational). A higher number of clusters introduces more noise (in the form of small clusters with no clear content).

The contingency tables of the clustering results with three clusters are depicted in Table 5. Part A of the table depicts the solution obtained with theoretical features, while Part B represents the solution obtained with POS features. Rows are gold standard classes and columns are clusters, labeled with the cluster number provided by the algorithm. The ordering of the cluster numbers corresponds to the quality of the cluster, measured in terms of the clustering criterion (see Equation (2)), 0 representing the cluster with the highest quality. In each cell C_{ij} of Table 5, the number of adjectives

Table 5
First model: Three-way solution contingency tables for theoretical and POS features. Rows are gold standard classes, columns are clusters. Row $Total_{GS}$ shows the number of Gold Standard lemmata and row $Total_{cl}$ the total number of lemmata contained in each cluster. Note that the column labeled *Total* represents the row sum for each part (as the number of items per class is identical).

	A: Theoretical			**B: POS**			
Cluster	*0*	*1*	*2*	*0*	*1*	*2*	*Total*
intensional (I)	0	0	2	0	2	0	*2*
intensional-qualitative (IQ)	0	0	1	0	1	0	*1*
qualitative (Q)	4	13	35	10	37	5	*52*
qualitative-relational (QR)	3	5	3	7	2	2	11
relational (R)	21	13	1	20	5	10	*35*
$Total_{GS}$	*28*	*31*	*42*	*37*	*47*	*17*	*101*
$Total_{cl}$	*834*	*1,287*	*1,400*	*1,234*	*1,754*	*533*	*3,521*

of class *i* that are assigned to cluster *j* by the algorithm is given. The largest value for each class is highlighted (see gray cells).

A striking feature of Table 5 is that results in the two parts (A and B) are very similar. The following can be observed:

(1) There is one cluster (cluster 0 in both solutions) that contains the majority of relational adjectives in the gold standard. This is the most compact cluster according to the clustering criterion.

(2) 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.

(3) The remaining cluster contains a mixture of qualitative and relational adjectives in both solutions.

(4) 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 (cluster 0).

The five-way results are depicted in Table 6. On the one hand, the table shows that the five-way structure found by the clustering algorithm is very similar to the three-way structure in Table 5. This means that the three clusters in A and B have basically been replicated by the three first clusters in C and D, respectively. On the other hand, the differences between the structures obtained using theoretical versus POS features are more obvious in the five-way solutions. From the set-up of the experiment, we had expected one cluster per class, plus QR and IQ adjectives isolated in a cluster of their own. This is clearly not borne out in Table 6. What we find instead is that (a) the mixed clusters persist and score high in the clustering criterion (see clusters 0 in solution C and 0–1 in solution D, with a mixture of Q, QR, and R adjectives), and (b) two additional small clusters are created (clusters 3 and 4 in both solutions) with no clear interpretation, suggesting that the three-way set-up matches better the structure uncovered by the clustering algorithm.

From the discussion of Tables 5 and 6 we conclude that the three-way clustering meets the target classification better than the five-way clustering, and that polysemous adjectives are not identified as a separate class. These results suggest that modeling

Table 6
First model: Five-way solution contingency tables. Information presented as in Table 5.

	C: Theoretical					**D: POS**					
Cluster	*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	1,754	275	258	*3,521*

polysemous adjectives in terms of additional, complex classes is not an adequate strategy (we return to this point subsequently).

Recall that we defined theoretical and POS features to compare the structures obtained using theoretically informed and theory-independent features. Further feature analysis, not reported here for space reasons, reveals a high correlation between the most descriptive features of solutions A and B.[3] This highlights the correspondence between the two feature representations with respect to the clustering results: The POS features elicited as most discriminative by the clustering algorithm are precisely those that correspond to the theoretical features. This correspondence explains the resemblance between the solutions obtained with the two types of representation and at the same time provides support for the present definition of the theoretical features.

Last but not least, note that we do not assign a score to each clustering solution. 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. We defer numerical evaluation until Section 5.

4.5 Discussion

4.5.1 Classification. The experiments presented provide feedback to the question, what is an appropriate broad semantic classification for adjectives? The clustering experiments provide empirical support for the qualitative and relational classes, as is particularly evident in the three-way solution (Table 5). These are classes that have traditionally been taken into account in descriptive grammar (Bally 1944; Picallo 2002) and computational resources such as WordNet (Miller 1998; Alonge et al. 2000), so we consider them to be quite stable and keep them in our classification.

Intensional and IQ adjectives, in contrast, are grouped together with qualitative adjectives in all solutions, because they do not exhibit distinctive enough distributional properties to differentiate them, a fact aggravated by the small size of the intensional class. From the point of view of NLP, it is reasonable to encode intensional adjectives by hand, given their limited number. For these reasons, we include the intensional class in the qualitative class in what follows (remember that, as mentioned in Section 3, WordNet also includes intensional adjectives in the qualitative—in their terms, *descriptive*—class).

"Hybrid" clusters, that is, clusters that contain adjectives from several semantic classes, play an interesting role in our cluster analyses. Such clusters seem to be coherent and stable, as they appear in all examined solutions (A, B, and also C and D in Tables 5 and 6) and have good scores in the clustering criterion. Significantly, however, most of the adjectives that are problematic for humans are assigned to hybrid clusters, where *problematic* means that they are not assigned to the same class by all four judges. Conversely, most adjectives in the hybrid clusters are problematic. Thus, hybrid clusters are useful to signal problems in the proposed classification. As an example, consider cluster 0 in Part C of Table 6: 17 out of the 24 (70.1%) gold standard adjectives in this hybrid cluster are problematic for humans. This cluster contrasts with the qualitative cluster (cluster 2 of Table 6), where only 10 out of its 42 (23.8%) lemmata are problematic.

Two kinds of adjectives crop up among problematic adjectives: so-called ethnic adjectives (*alemany* 'German', *menorquí* 'Menorcan', *sud-africà* 'South African', *xinès*

3 Descriptive features are defined here as those that are among the three features with highest or lowest mean values for at least three clusters in the five-way solution.

'Chinese'), and deverbal adjectives (*indicador* 'indicating', *parlant* 'speaking', *protector* 'protecting, protective', *salvador* 'savior'). Ethnic adjectives can act as predicates of copular sentences in a much more natural way than typical relational adjectives, and seem to be vague between a relational and a qualitative reading in their semantics (Raskin and Nirenburg 1998, page 173). This kind of adjective will mainly be treated as polysemous in the experiments reported in Section 5.

As for deverbal adjectives, they are clearly neither relational (they do not express a relationship to an object) nor intensional. They are also not typically qualitative, however, because they trigger a relationship to an event instead of denoting a simple property. For instance, *protector* triggers a relationship with a stable event of protecting in Example (18): A person named *Serra* belongs to the kind of associates who have as a primary role to *protect* the association.

(18) 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'

These considerations motivate the addition of a class of event-related adjectives in the overall classification. Event-related adjectives have not received much attention in the linguistic literature, except for one particular subtype, namely, adjectival uses of the participle (Bresnan 1982; Levin and Rappaport 1986; Bresnan 1995). As for computational resources, the English WordNet, as explained in Section 3, only distinguishes some participial adjectives. In the Italian WordNet, however, other event-related adjectives receive a specific treatment, through the encoding of the lexical relations CAUSES and LIABLE-TO, as exemplified in Example (19) (Alonge et al. 2000):

(19) a. *depuratorio* 'depurative, purifying' CAUSES *depurare* 'to depurate/purify'.
b. *giudicabile* 'triable' LIABLE-TO giudicare 'to judge'.

To sum up, the results of the experiments reported in this section motivate a three-way classification between qualitative, event-related, and relational adjectives. Note that, in the revised classification proposed in this section, classes are uniformly defined according to the *ontological type* of their denotation: Qualitative adjectives denote attributes or properties, relational adjectives denote relationships to objects, and event-related adjectives denote relationships to events. The classes correspond to the three major types of entities in an ontology (attributes, objects, events), more specifically, to the way adjectives participate from those entities. In this view, relational and event-related adjectives denote properties, just as qualitative adjectives do, but they are a specific type of property involving a relationship with either an object or an event. The classification is in fact similar to the one proposed in the Ontological Semantics framework (Raskin and Nirenburg 1998; Nirenburg and Raskin 2004).

Also note that the revised semantic classification bears a prominent relationship to morphology: In the default case, qualitative adjectives are not derived, event-related adjectives are deverbal, and relational adjectives are denominal. However, the correspondence between semantic classes and derivational type is not a one-to-one mapping. Although most event-related adjectives are deverbal, not only strictly deverbal adjectives evoke events: For instance, *tangible* 'tangible' evokes an event of touching, but there is no verb **tangir* in Catalan (*tangible* is built on the Latin verb *tangō*, 'touch'). Raskin and Nirenburg (1998, page 187) cite examples for English

such as *audible* or *ablaze*. Similarly, some object adjectives are not denominal (such as *botànic* 'botanical'). Conversely, some denominal or deverbal adjectives are qualitative: *vergonyós* 'shy' (from *vergonya* 'shyness'), *amable* (literally 'suitable to be loved'; has evolved to 'kind, friendly'). We will empirically check the correspondence between morphology and semantic class in Section 5.5.

4.5.2 Regular Polysemy. Our first series of experiments also provides feedback to the question, what is an adequate computational model for regular polysemy? Specifically, we have shown that the treatment of regular polysemy in terms of independent classes is not adequate. Remember that the motivation for the experiments presented in this section was the hypothesis that polysemous adjectives exhibit a linguistic behavior that participates from the basic classes involved in the regular polysemy, thus yielding feature values that are in between those of the basic classes (cf. Figure 1). Thus, we had expected that polysemous adjectives form a homogeneous group of lexical items, characterized precisely by the fact that they exhibit properties from each class to a certain degree. However, this expectation is not borne out in the results of the experiments. To this respect, it is striking that QR adjectives (polysemous between a qualitative and a relational reading) are spread throughout all the clusters in all solutions. They are not identified as a homogeneous group, nor as distinct from the rest. Crucially, as pointed out in Section 4.2, the differences between the feature values of polysemous adjectives and those of the basic classes are not strong enough to motivate a separate cluster.

We believe that the reason for these results is the fact that polysemous adjectives do *not* in fact have a homogeneous, differentiated profile: In a given corpus, most adjectives are used predominantly in one of their senses, corresponding to one of the basic classes, and thus the "hard" classification with three clusters fits better. For instance, the qualitative-relational adjective *irònic* ('ironic') is mainly used as a qualitative adjective in the corpus. Accordingly, it always appears in the qualitative clusters. Conversely, *militar* ('military') is mostly used as a relational adjective, and is consistently assigned to one of the relational clusters in all solutions. Thus, although polysemous adjectives on average do show a mixed behavior, each lexical item tends to pattern with one of the basic classes. An alternative conceptualization of regular polysemy and experimental design is called for, and this will be the topic of the next section.

5. Second Model: Polysemous Adjectives Simultaneously Belong to Different Classes

The experiments presented in the previous section pursued two goals: on the one hand, to test the initial classification proposal; on the other, to test a model of regular polysemy that treats polysemous adjectives in terms of separate classes. With respect to the first goal, the experiments in this section rely on the results of the previous experiments, and use the alternative classification described in Section 3.2. The alternative classification has in addition been supported by a clustering experiment not reported here for space reasons (see Boleda, Badia, and Batlle [2004] for details and discussion).

With respect to the second goal, we have shown that the first model is not successful at modeling regular polysemy. Furthermore, the analysis of feature values in the previous section suggests that the lack of success is not related to the specific technique used in the initial experiment, but to the properties of polysemous

adjectives: the fact that they are used predominantly in one of their senses, and the fact that the feature distributions of "polysemous classes" largely overlap with those of the basic classes.

In the present experiments, we develop an alternative approach to regular polysemy that is based on the perspective that polysemous adjectives belong to more than one semantic class, in the framework of **multi-label classification**. A typical example of a multi-label classification task is Text Categorization (Schapire and Singer 2000), where a document can be described via more than one label (e.g., *Health* and *Local*), so that it effectively belongs to more than one of the target classes. The motivation for this new approach is the fact that polysemous adjectives exhibit properties of all the classes involved (see Section 3.3). The hypothesis is that the evidence found for a polysemous adjective that is polysemous between, say, a relational and a qualitative use should be strong enough for the adjective to be assigned to both the relational and the qualitative classes. Note that by assigning the adjective to the two classes independently, we make an *implicit* classification of the adjective as polysemous. The success of the approach will depend on whether the different senses are sufficiently represented in the data, and it will be especially challenging to distinguish between noise and evidence for a given class.

5.1 Data and Gold Standard

The experiments reported in this section are based on a 16 million word fragment of the CTILC corpus (see Section 4.1). We additionally use an adjective database (Sanromà and Boleda 2010) with manually coded information about all adjectives occurring more than 50 times in the corpus (2,296 lemmata). The database codes the derivational type (deverbal, denominal, participial, non-derived) and suffix of each adjective.

A gold standard of 210 adjective lemmata (available in the Appendix) was selected from this database for the experiments. The lemmata were randomly sampled in a stratified fashion, balancing three factors of variability: frequency, morphological type, and suffix. Thus, the gold standard contains an equal number of adjectives from three frequency bands (low, medium, high), from the four derivational types, and from a series of suffixes within each type. This sampling method is aimed at achieving semantic variability.

Three experts assigned each of the 210 lemmata to one or two of the classes in the alternative classification, namely, event-related, qualitative, or relational. The decisions were reached by consensus and were based on expert knowledge together with the examination of the information in the database, corpus examples, and the judgments provided by 322 naive subjects in a large-scale annotation experiment.[4]

Table 7 shows the distribution of adjectives in the gold standard into classes according to the three experts. These are the data used in the experiments presented in this section. The proportion of polysemous adjectives is quite high, over 17%, with qualitative-relational being the most frequent type of polysemy. Also note that 51% of the adjectives are qualitative; this will be the baseline for the machine learning experiments presented subsequently.

4 For details on the annotation experiment, see Boleda, Schulte im Walde, and Badia (2008). The experiment yielded low inter-coder agreement scores (estimated κ 0.31–0.45, observed agreement 0.62-0.70). Note that the consensus classification is sub-optimal in the sense that its replicability cannot be estimated.

Table 7
Gold standard classification: Distribution and examples.

Class	Label	Example	#	%
qualitative	Q	*tenaç*, 'tenacious'	107	51.0
event	E	*informatiu*, 'informative'	37	17.6
relational	R	*cranià*, 'cranial'	30	14.3
qualitative-relational	QR	*familiar*, 'familiar'	23	11.0
qualitative-event	QE	*sabut*, 'known'	7	3.3
event-relational	ER	*comptable*, 'countable'	6	2.9
Total			*210*	*100*

Table 8
Feature sets. From left to right, each column depicts, for each feature set, an identifier, a description of the type of information used, the total number of features, and one example feature. Feature set *morph* contains two categorical features that are transformed into 25 if binarization is applied; the remaining feature sets are numerical.

Feature set	Description	#	Example
morph	morphological (derivational) properties	2 (25)	*suffix*
func	syntactic function of the adjective	4	*post-nom. modifier*
uni	uni-gram POS (1 word to left or to right)	24	*−1noun*
bi	bi-gram POS (1 word to left and 1 to right)	50	*−1noun+1adj*
theor	distributional cues of theoretical properties	18	*gradable*
Total		*98 (121)*	

5.2 Features

5.2.1 Feature Definition. We define five feature sets based on different types of linguistic information, to gain further insight into the properties of each class. In particular, we are interested in the properties of event-related adjectives, for which we do not have a description in the linguistic literature. Table 8 summarizes the properties of the feature sets used for the present experiments.

Feature set *morph* represents derivational properties of adjectives, as encoded in the adjective database. We include this type of information because of the relevance of morphology for the new classification (see Section 4.5). *Func* encodes the syntactic functions of the adjectives in the corpus, as explained in Section 4.1. *Uni* (for *unigram*) and *bi* (for *bigram*) encode the distribution of the adjective in the corpus in terms of the parts of speech of the surrounding words. Feature analysis of the first experiment showed that the word preceding and following the target were the most informative, so in the present experiment only a one-word window is taken into account. The unigram distribution (*uni*) encodes each part of speech separately, as was done in the first experiment, and the bigram distribution (*bi*) takes the left and right word jointly, to avoid feature correlation effects. In the latter feature set, only the 50 most frequent bigrams are considered, to avoid features that are too sparse.[5]

5 For a more detailed explanation of the information encoded in feature sets *uni* and *bi*, see Boleda (2007, section 5.2.2).

Table 9
New or revised features in feature set *theor*. Each row lists the property we aim to capture and the features through which the property is encoded. The information relies on the information in the corpus, which does not include full syntactic structure.

Property	**Features**
type of determiner	*NP headed by definite/indefinite/no determiner*
agreement properties	*gender and number of the NP*
syntactic function of head noun	*subject*, *object*, *complement to a preposition*
complement-bearing	*adjective followed by a preposition*
distance to the head	*linear distance (number of words)*

Finally, feature set *theor* (for *theoretical*) generalizes and adds to the theoretical properties used in the first experiment (Table 3 in Section 4.2). Upon inspection of the clustering solutions (not reported here for space reasons), some further potentially relevant distributional pieces of information cropped up that were included in the *theor* features of the present experiment. The new features, summarized in Table 9, cover several aspects of the noun phrases (NPs) in which adjectives occur: The type of determiner of the NP, agreement properties (as these can correlate with semantic properties), the syntactic function of the head noun, and the presence of a potential adjective complement. The latter are usually headed by prepositions (*El Joan està gelós d'en Pere*, 'Joan is jealous of Pere'). Finally, feature *distance to the head* is a reformulation of feature *adjacent* from Section 4.2. It encodes the mean distance of the adjective to the head, in number of words, as this is a more general definition that alleviates data sparseness.

As for feature values, they are computed as in the first experiment (see Equation (1)), with the following exceptions: (1) *morph* features are of categorical type, so their values are not numerical; (2) the two first features in Table 9, due to data sparseness considerations, are computed as proportions over the use of the adjective as a nominal modifier (see Equation (3), where a_{mod} is the number of occurrences of the adjective as modifier); (3) the values for feature *distance to the head*, also in Table 9, do not range from 0 to 1 as the other feature values, because they correspond to the mean distance to the head in number of words. The data set used for the present experiments is available at the ACL repository.[6]

$$v_{a,i} = \frac{f(a_{mod}, i)}{f(a_{mod})} \tag{3}$$

5.2.2 Feature Tuning. We test the effects of feature selection in the performance of the classifiers. The features are selected according to their performance within the machine learning algorithm used for classification. Accuracy for a given subset of features is estimated by cross-validation over the training data. Because the number of subsets increases exponentially with the number of features, this method is computationally very expensive, so we use a best-first search strategy. We also experiment with binarization of the two categorical features (*suffix*, *derivational type*).

6 http://aclweb.org/aclwiki/index.php?title=Database_of_Catalan_Adjectives_(Repository).

5.3 Method

The classification task is approached with a two-level architecture.

1. The decision on the class of the adjective is decomposed into three *binary decisions*: Is it qualitative or not? Is it event-related or not? Is it relational or not?
2. A complete classification is achieved by *merging* the results of the binary decisions. A consistency check is applied by which (a) if all decisions are negative, the adjective is assigned to the qualitative class (the most frequent one; this was the case for a mean of 4.6% of the class assignments); (b) if all decisions are positive, we randomly discard one (three-way polysemy is not foreseen in our classification; this was the case for a mean of 0.6% of the class assignments).

This is the standard architecture for multi-label classification tasks (Schapire and Singer 2000; Ghamrawi and McCallum 2005), and it has also been applied to NLP problems such as entity extraction and noun-phrase chunking (McDonald, Crammer, and Pereira 2005).

Note that in the present experiments we change both the classification and the approach (unsupervised vs. supervised) with respect to the first set of experiments presented in Section 4, which can be seen as a sub-optimal technical choice. After the first series of experiments that required a more exploratory analysis, however, we believe that we have now reached a more stable classification, which we can test by supervised methods. In addition, we need a one-to-one correspondence between gold standard classes and clusters for the approach to work, which we cannot guarantee when using an unsupervised approach that outputs a certain number of clusters with no mapping to the gold standard classes.

We test two types of classifiers. The first type are Decision Tree classifiers trained on different types of linguistic information coded as feature sets. Decision Trees are one of the most widely machine learning techniques (Quinlan 1993), and they have been used in related work (Merlo and Stevenson 2001). They have relatively few parameters to tune (a requirement with small data sets such as ours) and provide a transparent representation of the decisions made by the algorithm, which facilitates the inspection of results and the error analysis. We will refer to these Decision Tree classifiers as *simple classifiers*, in opposition to the ensemble classifiers, which are complex, as explained next.

The second type of classifier we use are ensemble classifiers, which have received much attention in the machine learning community (Dietterich 2000). When building an ensemble classifier, several class proposals for each item are obtained from multiple simple classifiers, and one of them is chosen on the basis of majority voting, weighted voting, or more sophisticated decision methods. It has been shown that in most cases, the accuracy of the ensemble classifier is higher than the best individual classifier (Freund and Schapire 1996; Dietterich 2000; Breiman 2001). The main reason for the general success of ensemble classifiers is that they are more robust towards the biases particular to individual classifiers: A bias shows up in the data in the form

of "strange" class assignments made by one single classifier, which are therefore overridden by the class assignments of the remaining classifiers.[7]

For the evaluation, 100 different estimates of accuracy are obtained for each feature set using 10-run, 10-fold cross-validation (*10x10 cv* for short). In this schema, 10-fold cross-validation is performed 10 times, that is, 10 different random partitions of the data (*runs*) are made, and 10-fold cross-validation is carried out for each partition. To avoid the inflated Type I error probability when reusing data (Dietterich 1998), the significance of the differences between accuracies is tested with the *corrected resampled t-test* as proposed by Nadeau and Bengio (2003).[8]

5.4 Results

5.4.1 Simple Classifiers. The accuracies for the simple classifiers are shown in Table 10. Part A of the table lists the results for each of the binary decisions (qualitative/non-qualitative, event/non-event, relational/non-relational). The accuracy for each decision is computed independently. For instance, a qualitative-event adjective is judged correct within the qualitative class iff the decision is *qualitative*; correct within the event class iff the decision is *event*; and correct within the relational class iff the decision is *non-relational*.

Part B reports the accuracies for the overall, merged class assignments, taking polysemy into account (qualitative vs. qualitative-event vs. qualitative-relational vs. event, etc.).[9] In Part B, we report two accuracy measures: full and partial. Full accuracy requires the class assignments to be identical (an assignment of qualitative for an adjective labeled as qualitative-relational in the gold standard will count as an error), whereas partial accuracy only requires some overlap in the classification of the machine learning algorithm and the gold standard for a given class assignment (a qualitative assignment for a qualitative-relational adjective will be counted as correct). The motivation for reporting partial accuracy is that a class assignment with some overlap with the gold standard is more useful than a class assignment with no overlap. The figures in the discussion that follow refer to full accuracy unless otherwise stated.

For the qualitative and relational classes, taking into account distributional information allows for an improvement over the default morphology–semantics mapping outlined in Section 4.5: Feature set *all*, containing all the features, achieves 75.5% accuracy for qualitative adjectives; feature set *theor*, with carefully defined features, achieves 86.4% for relational adjectives. In contrast, morphology seems to act as a ceiling for

7 The experiments discussed in this section were carried out with the Weka software package (Witten and Frank 2011), version 3.6. The Decision Tree algorithm used is J48, the latest open source version of C4.5 (Quinlan 1993), with default parameters (*binary splits* = False, *confidence factor for pruning* = 0.25, *minimum number of instances per leaf* = 2, *reduced-error pruning* = False, *subtree raising* = True, *unpruned* = False, *use Laplace* = False). AdaBoost has also been used with default parameters (*base classifier* = Decision Stump, *number of iterations* = 10, *random seed* = 1, *use resampling instead of reweighting* = False, *weight threshold* = 100). For Attribute Bagging, we used the Random Subspace algorithm, with J48 as base classifier (parameters as before), *bag size* = 1/3, and *random seed* = 1. We experimented with different values for the number of iterations (see Section 5.4.2).

8 Note that the corrected resampled t-test can only compare accuracies obtained under two conditions (algorithms or, as is our case, feature sets); ANOVA would be more adequate. In the field of machine learning, there is no established correction for ANOVA for the purposes of testing differences in accuracy (Bouckaert 2004). Therefore, we use multiple t-tests instead, which increases the overall error probability of the results for the significance tests.

9 Note that, for each adjective, only 10 different full classification proposals are obtained in each feature set, because each adjective is only used once per run for testing. Therefore, while the per-class accuracy for each feature set is assessed from 100 estimates (obtained via 10x10 cv), the accuracy of the different feature sets for full classification is assessed comparing 10 accuracies. This holds for Tables 10 and 11.

Table 10
Second model: Results with simple classifiers using different feature sets. The frequency baseline (first row) is marked in italics. The last row, headed by *all*, shows the accuracy obtained when using all features together for tree construction. The remaining rows follow the nomenclature in Table 8; a *FS* subscript indicates that automatic feature selection is used as explained in Section 4.2. For each feature set, we record the mean and the standard deviation (marked by ±) of the accuracies. Best and second best results are boldfaced. Significant improvements over the baseline are marked as follows: *p < 0.05; **p < 0.01; ***p < 0.001.

	A: Per-class accuracy			B: Overall accuracy	
	Qualitative	**Event**	**Relational**	**Full**	**Partial**
baseline	*65.2 ± 11.1*	*76.2 ± 9.9*	*71.9 ± 9.6*	*51.0 ± 0.0*	*65.2 ± 0.0*
morph	68.2 ± 11.1	**87.3**** ± 6.3	85.2*** ± 7.2	59.9*** ± 2.2	84.7*** ± 0.7
morph$_{FS}$	72.5* ± 7.9	**89.1**** ± 6.0	84.2*** ± 7.5	**60.6***** ± 1.3	**87.8***** ± 0.4
func	**75.1**** ± 9.0	76.1 ± 9.8	82.8*** ± 7.5	56.0*** ± 1.9	80.6*** ± 1.8
uni	64.2 ± 10.8	68.4 ± 12.0	82.1** ± 9.0	42.8 ± 2.7	74.8*** ± 2.6
uni$_{FS}$	66.0 ± 9.3	75.1 ± 10.6	82.2** ± 7.5	52.9 ± 1.9	77.0*** ± 2.0
bi	63.8 ± 9.9	66.2 ± 9.8	78.2* ± 8.2	46.1 ± 2.3	77.8*** ± 1.8
bi$_{FS}$	67.4 ± 10.6	72.3 ± 10.2	83.0*** ± 8.3	52.3 ± 1.7	76.7*** ± 1.0
theor	71.8 ± 10.0	74.1 ± 9.9	**86.4***** ± 7.6	54.8*** ± 1.7	81.8*** ± 1.8
all	**75.5**** ± 9.0	86.5** ± 6.4	**86.0***** ± 6.5	**62.5***** ± 2.5	**87.6***** ± 2.5

event-related adjectives: The best result, 89.1%, is obtained with morphological features using feature selection. As will be shown in Section 5.5, event-related adjectives do not exhibit a differentiated distributional profile from qualitative adjectives, which accounts for the failure of distributional features to capture this class. As could be expected, the best overall result is obtained with feature set *all*, that is, by taking all features into account: 62.5% full accuracy is a highly significant improvement over the baseline, 51.0%. The second best results are obtained with morphological features using feature selection (60.6%), due to the high performance of morphological information with event adjectives.

Also note that the POS feature sets, *uni* and *bi*, are not able to beat the baseline for full accuracy: Results are 42.8% and 46.1%, respectively, jumping to 52.9% and 52.3% when feature selection is used, still not enough to achieve a significant improvement over the baseline. Thus, for this task and this set-up, it is necessary to use well motivated features. In this respect, it is also remarkable that feature selection actually *decreased* performance for the motivated distributional feature sets (*func*, *sem*, *all*; results not shown in the table), and only slightly improved over morph (59.9% to 60.6% accuracy). Carefully defined features are of high quality and therefore do not benefit from automatic feature selection. Actually, Witten and Frank (2011, page 308) 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."

In the partial evaluation condition, however, all feature sets achieve a highly significant improvement over the baseline ($p < 0.001$). Therefore, the classifications obtained with any of the feature sets are more useful than the baseline, in the sense that they present more overlap with the gold standard.

5.4.2 Ensemble Classifiers. Error analysis on the results using simple classifiers (not reported for space reasons) revealed that the errors made by the different classifiers, using different feature sets, are qualitatively quite different. This motivated the use

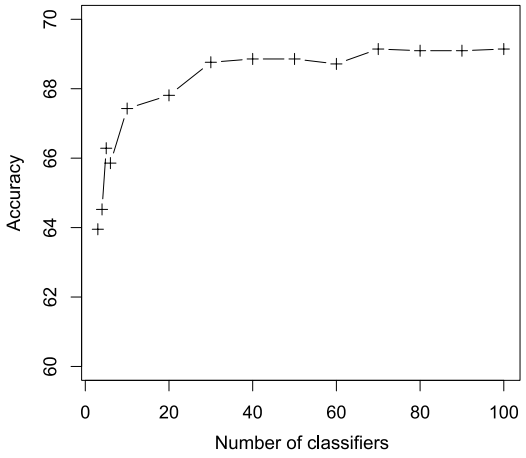


Figure 2
Accuracy of the Attribute Bagging classifier as a function of the number of random partitions i. Increasing i leads to a rapid increase of accuracy up to $i = 30$; after that, accuracy stabilizes and experiences only a slight increase.

of Attribute Bagging (Ho 1998; Bryll, Gutierrez-Osuna, and Quek 2003), an ensemble classifier (EC) in which the class assignments are obtained by majority voting over randomly sampled feature subsets.[10] Attribute Bagging has two main parameters: the bag size (number of features used for each classification; it was set to 1/3 given results reported in the literature, although varying this parameter did not affect the results much), and the number of iterations i (we tested 3, 4, 5, 6, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100; note that our total feature size is 121, see Table 8). Figure 2 shows that increasing i leads to a rapid increase of accuracy up to $i = 30$; with higher i, accuracy experiments only a slight increase.

Table 11 shows the results of Attribute Bagging, compared to the best simple classifier and human agreement (observed agreement, in percentage). The results obtained with AdaBoost (a standard EC; default parameters) are also included as a sanity check. The best results with Attribute Bagging, reported in the table, were obtained using both feature selection and binarization (binarization did not improve results for the remaining classifiers in Tables 10 and 11).

The Attribute Bagging EC with i = 5 achieves comparable accuracy to AdaBoost with default parameters. Full accuracy results with the Attribute Bagging classifier with i = 100 (69.1%) are significantly higher than those of the best simple classifier (62.5%; $p < 0.0001$) and the AdaBoost classifier ($p = 0.01$; recall however that we did not optimize AdaBoost's parameters). Ensemble classifiers are thus helpful for our task.

The best classifier in our experiments (Att. Bagg.$_{FS,bin}$,i = 100) obtains 69.1% full and 89.0% partial accuracy. This is comparable to the agreement between the expert annotation of the gold standard and naive subjects participating in a large-scale annotation experiment (p_o = 0.68, or 68%, and κ = 0.55 for full accuracy, p_o = 0.85, or 85%, and κ = 0.72 for overlapping accuracy; see Boleda, Schulte im Walde, and Badia [2008] for details on the comparison). If we view human agreement as an upper bound, we have reached

10 Grouping subsets according to linguistic considerations (i.e., building an EC over the feature subsets listed in Table 8) improved upon the best simple classifier, but not upon Attribute Bagging.

Table 11
Second model: Results of the ensemble classifiers, compared to the best simple classifier (first row) and to the human agreement on the gold standard (last row). *Att. Bagg.* stands for *Attribute Bagging*, and *i* corresponds to the number of iterations. Percentage human agreement is included in the last row. An *FS* subscript indicates feature selection, and *bin* binarization. Columns as in Table 10. Best and second best results are boldfaced. Significant improvements over the best simple classifier are marked as follows: *p < 0.05, **p < 0.01, ***p < 0.001.

	A: Per-class accuracy			B: Overall accuracy	
	Qualitative	**Event**	**Relational**	**Full**	**Partial**
best simple (*all*)	75.5 ± 9.0	**86.5** ± 6.4	86.0 ± 6.5	62.5 ± 2.5	87.6 ± 2.5
AdaBoost	**82.0*** ± 8.6	85.6 ± 7.1	88.0 ± 6.7	66.0* ± 1.9	**89.9*** ± 1.3
Att. Bagg.$_{FS,bin}$,*i*=5	77.0 ± 8.7	85.8 ± 7.1	**89.0** ± 6.5	**66.3*** ± 1.1	87.0 ± 1.5
Att. Bagg.$_{FS,bin}$,*i*=100	**81.0** ± 8.8	**86.1** ± 6.9	**90.1*** ± 5.3	**69.1***** ± 1.0	**89.0** ± 1.0
Human agreement	—	—	—	*68*	*85*

the maximum accuracy that could be obtained via machine learning for the present task. Further improvements will need to be preceded by an improvement in the agreement scores of human judges, that is, by a better definition of the classes and the classifying task.

Finally, Table 11 shows that the best results are obtained for the relational class (90.1%), followed by the event class (86.5%), and the qualitative class has the lowest scores (at most 82%). The qualitative class contains attribute-denoting adjectives, but in the present definition it is also populated with adjectives that simply do not fit into the other classes (such as intensional adjectives, as explained earlier). Also, whereas some adjectives in the class are prototypical qualitative adjectives such as *gros* 'big' or *llarg* 'long', others are unprototypical types of properties (*subaltern* 'subordinate', *subsidiari* 'subsidiary'). This factor brings heterogeneity into the class, which justifies the relatively poor performance of the classifier on this task. Significantly, also, ECs do not improve upon simple classifiers for the event class; again, morphological information acts as a ceiling and no combination of information serves to go beyond that ceiling, as will become clear in the error analysis explained next.

5.5 Error Analysis

Table 12 depicts the contingency table of the classifications by the experts (rows) and one randomly chosen run of the Attribute Bagging classifier with $i = 100$ (columns). The table shows that there are two major sources of errors: First, the confusion between the qualitative and event classes, which is responsible for 14 errors (see dark-gray shaded cells in the table; also note that the related Q–QE and E–QE misclassifications account for another 14 errors). To compare, note that the confusion between the qualitative and relational classes only accounts for six of the errors, and there are no cases of confusion between event and relational adjectives.

The second major source of errors is the overgeneration of polysemous adjectives (see medium-gray shaded cells): there are 26 adjectives tagged as monosemous by the experts and assigned a polysemous class by the system. To compare, the opposite case (i.e., tagging polysemous adjectives as monosemous) accounts for 13 errors only (see light-gray shaded cells). We next examine the two main types of errors in more detail.

Table 12
Contingency table comparing the gold standard (rows) against run 2 of the Attribute Bagging classifier with $i = 100$ (columns). Dark-gray cells highlight the confusion between the qualitative and event classes; medium-gray cells highlight the overgeneration of polysemous adjectives; light-gray cells highlight the opposite case, that is, the generation of monosemous adjectives that should have been tagged as polysemous.

	Q	E	R	QR	QE	ER	*Total*
Q	90	4	2	3	8	0	*107*
E	10	17	0	1	6	3	*37*
R	4	0	20	4	0	2	*30*
QR	5	0	4	13	0	1	*23*
QE	1	1	0	0	5	0	*7*
ER	0	0	2	1	0	3	*6*
Total	*110*	*22*	*28*	*22*	*19*	*9*	*210*

5.5.1 Distinguishing between Qualitative and Event Adjectives. Table 12 suggests that there are difficulties in distinguishing event-related from qualitative adjectives. Feature analysis confirms that the distinction between these two classes is only partially possible on morphological grounds, but not on distributional grounds. As for morphological information, Figure 3 shows that most event-related adjectives are deverbal or participle adjectives, although the reverse is not true: 14 deverbal and two participle adjectives are qualitative.

In fact, the class distribution varies with the suffix (see Table 13): Some types, such as *-or* and the participle, show a clear predominance of the event class (see dark-gray shaded cells); other types, such as *-ble*, *-iu*, or *-nt*, are more spread in their distribution (see light-gray shaded cells). Thus, the suffix seems to influence the resulting readings, with some active-like suffixes building a much more transparent relation to the event (*creador* 'creating', *exportador* 'exporting', *recomanat* 'recommended'), and some passive-like or stative suffixes being more prone to creating a stative meaning (*contingent* 'contingent', *formidable* 'formidable | terrific', *significatiu* 'significant'). The aspectual class of the deriving verb (Vendler 1957) also plays a role: For instance, although the meaning of *abundant* ('abundant') is related to that of the verb *abundar* ('abound'), it clearly has a more stative (property-like) meaning than many of the other event adjectives, due to the fact that the deriving verb is stative. Correspondingly,

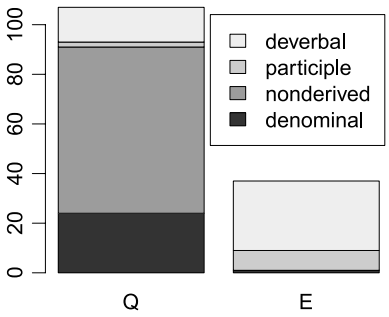


Figure 3
Derivational types in the qualitative (Q) and event-related (E) classes. The bars represent the classes, and the colors the derivational types, as shown in the legend.

Table 13
Contingency table of the most frequent deverbal suffixes (rows) and classification (columns). Dark-gray cells highlight suffixes that mostly create event-related adjectives, and light-gray cells indicate suffixes with a more spread class distribution.

	Q	E	R	QR	QE	ER	*Total*
-ble	3	6	0	0	1	1	*11*
-iu	3	1	1	2	0	4	*11*
-nt	4	6	0	0	0	1	*11*
-or	1	10	0	0	0	0	*11*
participle	2	8	0	0	5	0	*15*
Total	*16*	*36*	*3*	*2*	*7*	*6*	*70*

abundant is classified as qualitative by the Attribute Bagging algorithm. This variation in the morphology–semantics interface is also mirrored in the feature value distributions, as will be shown subsequently.

As for distributional information, Figure 4 depicts the feature value distribution for nine selected features across basic classes qualitative (Q), event-related (E), and relational (R), excluding polysemous adjectives. The figure clearly shows that, whereas relational adjectives tend to have a differentiated value distribution for many of the features, the values for the event-related class in general overlap with those of the qualitative class. In fact, of the 18 theoretically motivated features defined for the experiments, only two exhibit statistically significant differences in the distribution of the qualitative and event-related class according to a two-tailed t-test ($\alpha = 0.01$; no equality of variance assumed). These are *pre-nominal* and *complement-bearing* (graphs A and I in Figure 4; df = 108.9/54.6, t = 3.56/−3.09, p-value = 0.0005/0.003, respectively). These two features show that in general event-related adjectives appear less often than qualitative adjectives in pre-nominal position and tend to bear more complements. Both differences are presumably due to the fact that many event adjectives inherit the argument structure of the deriving verb, with arguments expressed via PPs, constituting heavier constituents that are placed after the head noun. This is but a slight tendency and it is not homogeneous through the class, however.

The remaining features do not show differences between event and qualitative adjectives, but rather properties of relational adjectives. In addition to the properties that were already known, the figure shows that relational adjectives appear more often in definite NPs acting as preposition complements (graphs E and H). Thus, the typical syntactic context for a relational adjective is *preposition + definite determiner + noun + relational adjective*). This type of adjective also appears slightly more often with feminine head nouns, which could be due to the fact that, in Catalan, many abstract nouns (*física* 'physics', *capacitat* 'ability') are feminine, for morphological reasons. These nouns are often modified by relational adjectives to select for subtypes of the class of objects denoted by the nouns (McNally and Boleda 2004).

Another difficulty in the distributional characterization of the event class is the fact that it is quite heterogeneous, due to the variation at the morphology–semantics interface discussed earlier. This can be traced in Figure 4 by the fact that for most of the features, the box of the event class is larger than the box of the other two classes, meaning that there is more variation within the event class than within the other two classes.

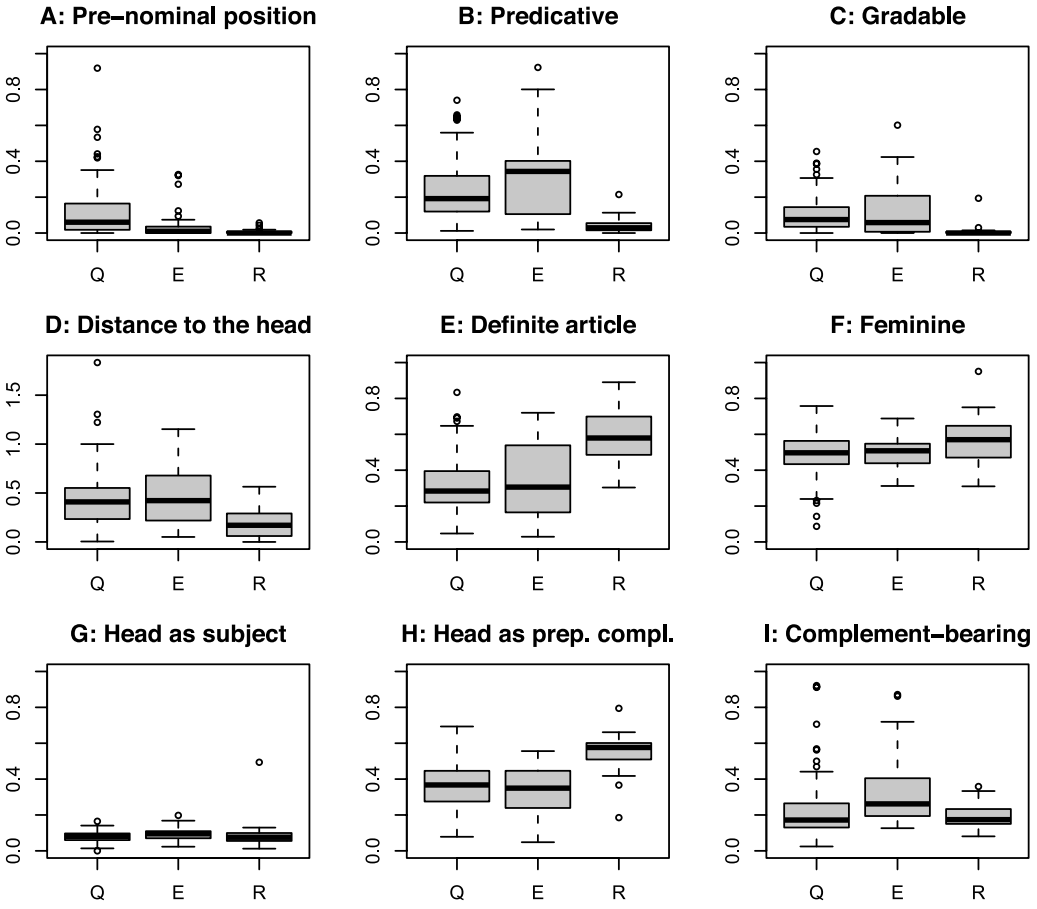


Figure 4
Feature value distribution across classes qualitative (Q), event-related (E), and relational (R) for nine selected features (see Section 5.2 for the definition of these features).

To sum up, morphological features can quite reliably spot event-related adjectives, but distributional information cannot. As a result, in the cases where morphology gives the wrong prediction, nothing can be done on the distributional side to remedy this. This results in the confusion of event-related and qualitative adjectives shown in Table 12.

5.5.2 Detecting Polysemous Adjectives. Recall from Table 12 that the system overgenerates polysemous adjectives: There are 26 monosemous adjectives assigned to a polysemous class. One of the reasons for this overgeneration is the procedure followed. The procedure treats the decision on each of the basic classes as if they were all independent. Thus, the probability for an adjective being polysemous amounts to the product of the probabilities of the adjective belonging to each of the basic classes, as expressed in Equation (4) for two arbitrary given classes, A and B.

$$p(AB) = p(A) * p(B) \tag{4}$$

Table 14 shows that the distribution of polysemous items predicted by Equation (4) is more similar to the distribution obtained with the best machine learning classifier

Table 14
Distribution of polysemous items and absolute numbers, according to the prediction (Equation (4); first column), in the machine learning (ML) results shown in Table 12 (second column), and in the gold standard (GS; third column).

	Predicted	ML	GS
QR	15	22	23
QE	19	19	7
ER	5	9	6

(ML) than to the distribution of polysemous items in the gold standard (GS) for the QE cases. The distribution is estimated from the frequency over the 210 adjectives in the gold standard, and shown as absolute numbers.

Both Equation (4) and the ML classifier assign 19 adjectives to the QE polysemy type, although the gold standard contains only 7 QE adjectives. The equation predicts fewer QR adjectives than observed in the data, but in this case the classifier produces a similar number of QR adjectives than attested (22 vs. 23). Finally, the classifier produces more ER adjectives than observed and also than predicted by Equation (4), but in this case the numbers are so small that no clear tendencies can be observed. Thus, the procedure followed can be said to cause the overgeneration of items for the QE polysemy type, but it does not account for the other two polysemous classes.

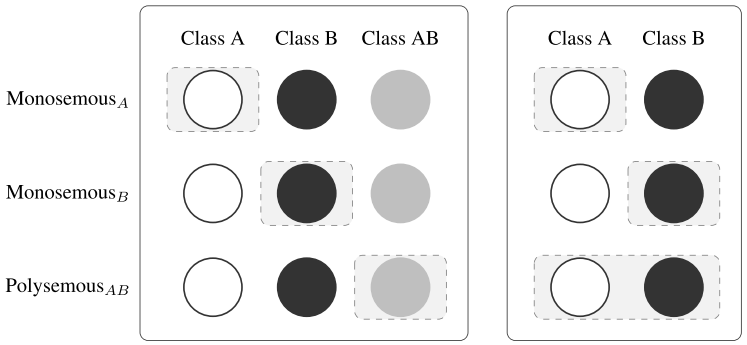
Further qualitative analysis on the overgenerated polysemous adjectives (corresponding to the middle-gray cells in Table 12; not reported because of space concerns) showed that different types of evidence motivate the inclusion of monosemous adjectives in two classes, causing them to be considered polysemous. This suggests that, because polysemous adjectives exhibit only partial or limited evidence of each class, the threshold for positive assignment to a class is lowered, resulting in the observed overgeneration. Recall that at the beginning of this section, when introducing the model, we warned that it would be specially challenging to distinguish between noise and evidence for a given class. We have indeed found this to be a challenge. The mentioned effect is amplified by the procedure followed, which assumes that the class assignments are independent, thus not adequately enough modeling the empirical distribution of polysemy.

6. Discussion: Towards a Model for Regular Polysemy

The acquisition experiments presented in Sections 4 and 5 correspond to two different underlying models of regular polysemy. Figure 5 represents the two models in a simplified scenario with just two basic classes (A and B). The first model (Figure 5(a); experiments in Section 4) treats polysemous words in terms of independent classes. The second model (Figure 5(b); experiments in Section 5) treats polysemous words alike to those of the basic classes: Polysemous assignments result from membership in two basic classes.

As can be seen in the figure, there are two main differences between the models. First, the number of classes considered: Whereas the second model only considers n classes, in the general case the first model will need to consider

$$n + \binom{n}{2}$$



(a) First model (b) Second model

Figure 5
The two models of regular polysemy tested in this article, assuming a simplified scenario with just two basic classes (A and B). The rows represent three different cases: one monosemous adjective of class A (*Monosemous*$_A$), one monosemous adjective of class B (*Monosemous*$_B$), and one polysemous adjective (*Polysemous*$_{AB}$). The columns represent the classes assumed in each model: Three classes (a), or only two (b). The correct class assignments for each case are shown as dashed rectangles.

classes (n monosemous classes plus $\binom{n}{2}$ polysemous classes, all the possible two-combinations of the monosemous classes). This formula assumes that only two-way regular polysemy is allowed, as in this article; polysemy across three or more classes would make the explosion of classes even worse. It is clear that the second model is easier to learn.

The second difference concerns the way class assignments to polysemous words are carried out. In the first model, polysemous words are assigned to one single, independent class, whereas in the second they are assigned to each of the two basic classes that give rise to the regular polysemy. Recall that the motivation for the first model was that—given that regularly polysemous adjectives show a particular hybrid behavior—we could expect that polysemous adjectives could be characterized as differentiated classes. This expectation has clearly not been borne out. A further problem with the first model it that it in principle allows for a polysemous class AB whose properties do not necessarily have anything to do with those of the basic classes A and B. The second model, in contrast, enforces that polysemous adjectives exhibit properties of each of the classes they participate in, which is both theoretically and empirically more adequate. For these reasons, we believe that the second model is more suitable to represent regular polysemy than the first model.

The second model is also not completely satisfactory, however. As discussed in the previous section, in the current implementation of the model the class assignments are assumed to be independent (though this need not be the case in other instantiations of the model). Also, in a way, it is at the opposite end of the scale with respect to the first model: Whereas in the first model polysemous adjectives do not need to have anything in common with the basic classes, in the second model a polysemous word is assumed to be just like any other word in each of the basic classes. For instance, a qualitative-relational adjective is assumed to function both as a full-fledged qualitative adjective and a full-fledged relational adjective. By their very nature, polysemous words will show only *some* evidence for each of the classes, as their occurrences (and thus their properties) will be distributed across the two classes. Therefore, they will be untypical members of at least one of the intervening classes.

An alternative instantiation of the second model could use soft clustering (Pereira, Tishby, and Lee 1993; Rooth et al. 1999; Korhonen, Krymolowski, and Marx 2003), which assigns a probability to each of the classes and is thus not bound to a hard yes/no decision, as our approach does. From a theoretical point of view (and for many practical purposes such as dictionary construction), however, a distinction between monosemous and polysemous words is desirable, which adds a further parameter to be optimized in a soft clustering setting. Overlapping clustering (Banerjee et al. 2005), which allows for membership in multiple clusters, avoids this difficulty. Both methods have the advantage that they do not assume independence of the decisions. The most serious problem for the experiments presented in this article, however, would presumably also be a problem for these settings: The fact that the skewed sense distribution of many words makes it difficult to distinguish evidence for a particular class from noise. In the soft clustering setting, for instance, it would be hard to distinguish whether 10% evidence for class A and 90% for class B corresponds to polysemy with a skewed distribution, to noise in the data, or simply to an untypical instance.

To sum up, the main problem for the models presented in this article is that neither model can capture the distributional connection between P(AB) and P(A), either because AB and A are seen as unrelated atoms in the first place (first model), or because AB is diluted into A and B (second model). A more refined statistical approach that can model this interdependency is needed for further progress. Such a model should take into account both the differences of polysemous adjectives with respect to the other adjectives in the basic classes (first model) *and* their similarities (second model), thus directly capturing their hybrid behavior.

7. Conclusion

This article has tackled the automatic induction of semantic classes for Catalan adjectives, with a special emphasis on regular polysemy. To our knowledge, this is the first time that such an endeavor has been carried out, as (1) related work on lexical acquisition has focused on verbs (and, to a lesser extent, nouns) and on major languages such as English and German; and (2) polysemy in general has been largely ignored in lexical acquisition, and regular polysemy has only been sparsely addressed in empirical computational semantics.

We have explored the relationship between observable cues and semantic properties for adjectives, and, specifically, the morphology–semantics and syntax–semantics interfaces. We have showed that there is a systematic relation between the type of denotation of an adjective and its morphological and distributional properties. Our experiments have furthermore related the linguistic properties of adjectives as described in the literature to the information that can be extracted from linguistic resources, such as corpora or lexical databases. The presented results and analyses provide empirical support for the qualitative and relational classes, defined in theoretical work, and bring event-related adjectives into focus, a type of adjective that has been largely neglected in the literature.

This article has focused on Catalan as a case study, but most of the properties discussed (predicativity, gradability, complementation patterns), as well as the types of polysemy explored, are relevant for a broader range of languages, specially Indo-European languages (Dixon and Aikhenvald 2004). The approach does not require deep-processing resources (full parsing, semantic tagging, semantic role labeling), which makes it useful for lesser-researched languages.

The experiments show that a major bottleneck for our purposes is the definition of the classification itself: The machine learning results obtained have reached an upper bound, as the best classifier has achieved 69.1% accuracy (against a 51.0% baseline), and the human agreement is 68%. Thus, improvements in the computational task will need to be preceded by improvements in the agreement scores, that is, by a better and clearer definition of the classification and the classification task. We have shown that this is by no means a trivial issue. In fact, low inter-coder agreement scores are a problem for machine learning approaches to semantic and discourse-related phenomena in general. This is in contrast to tasks such as POS tagging or syntactic parsing, where relatively high inter-coder agreement scores are achieved. This state of affairs is probably due to the fact that semantic and pragmatic phenomena are much less well understood than morphological or syntactic phenomena.

Our experiments have highlighted a number of problems with the current classification proposal. First, the distinction between event-related and qualitative adjectives. The event class cannot be distinguished from the qualitative class with the distributional information used in this article, and its members are not homogeneous. We have shown that factors such as the aspectual class of the deriving verb or the suffix of the deverbal adjective play a role in the semantic and syntactic behavior of these adjectives that should be further explored. Also, a crucial type of evidence remains to be explored, namely, the selectional preferences of adjectives. These may be a relevant clue to the differences between qualitative and event-related adjectives. The second main problem is the fact that the qualitative class contains adjectives that do not fit into the other classes, constituting a sort of "catch-all" class. A natural extension for the work presented in this article would be to define a finer-grained categorization including the problematic cases discussed earlier. For instance, adjectives deriving from stative verbs could be distinguished from those deriving from active verbs, and different types of qualitative adjectives could be treated as different classes.

As for regular polysemy, we have shown that polysemous adjectives exhibit a hybrid behavior, with properties from all the classes involved in each type of regular polysemy. We have empirically tested two models of the phenomenon aimed at exploiting this hybrid behavior. The first model treats polysemous words in terms of independent classes, and we have argued that it is not adequate, neither from a theoretical nor from an empirical perspective. The second model assumes that polysemous words belong to each of the basic classes participating in the regular polysemy. This model is more adequate than the first one, as it accounts for the properties of the basic classes found in polysemous words, but it fails to account for the differences between polysemous and monosemous words. To improve on the modeling of regular polysemy, we plan to move to token-based (word-in-context) models (Schütze 1998; Erk and Padó 2010), as opposed to type-based models as we have done in this article. This should in turn shed light into the problem of distinguishing between evidence for a particular class from noise, discussed previously.

Finally, at a methodological level, we have illustrated how the broad coverage, large-scale, radically empirical approaches developed in computational linguistics can be of use to uncover phenomena and facts that are relevant for the study of language, providing complementary evidence to the analytic tools traditionally used by linguists. Most prominently, we have shown that (1) by randomly sampling the set of words to be analyzed, new or neglected phenomena emerge; (2) the feature representation typically used by machine learning algorithms provides an empirical handle to the linguistic properties of words that can be explored in different ways (e.g., to test hypotheses about the morphology-syntax and semantics-syntax interfaces); (3) machine learning

experiments provide a framework for the systematic evaluation of different models of the phenomenon under study (in our case, both adjective classification and regular polysemy). Computational linguistic studies are also inherently limited in several aspects, such as the type of evidence that can be used or the ways in which it can be used. Despite these limitations, we believe that empirical computational linguistics approaches are a gold mine of new knowledge about language.

Appendix: Gold Standard Data

In the following, we include the lemmata that were manually classified for the first and second set of experiments, respectively (Sections 4 and 5). For details on the classes and the methodology, see the body of the article. The translation of the adjectives has been carried out with the help of the Spanish–English/English–Spanish Collins Dictionary (3rd edition) and Google Translator.[11] Different senses are separated with a vertical bar ('|'), different translations of the same sense with a comma (','). Whenever possible, we have included adjective equivalents; many of the relational adjectives, however, are equivalent to attributive uses of nouns. Such nominal translations have been marked with *(attr.)*.

Recall that the gold standard for the second experiment, together with its feature values, is available at the ACL repository (see URL in footnote 6).

Gold standard for the experiments with the first model (Section 4).

- **intensional (I)**: *mer* 'mere', *presumpte* 'alleged'.
- **qualitative (Q)**: *accidental* 'accidental', *accidentat* 'uneven, rough | injured', *alienant* 'alienating', *anticlerical* 'anticlerical', *avergonyit* 'ashamed', *bastard* 'bastard', *benigne* 'benign', *caracurt* 'short-faced', *coherent* 'coherent', *colpidor* 'striking', *contradictori* 'contradictory', *cosmopolita* 'cosmopolitan', *destructor* 'destructive', *diversificador* 'diversifying', *duratiu* 'durative', *escàpol* 'fleeing', *esfereïdor* 'terrifying', *evident* 'evident', *exempt* 'exempt', *expeditiu* 'expeditious', *fortuït* 'fortuitous', *gradual* 'gradual', *grandiós* 'grand', *gratuït* 'free | gratuitous', *honest* 'honest', *implacable* 'implacable', *infreqüent* 'infrequent', *innoble* 'ignoble', *inquiet* 'anxious | restless', *insalvable* 'insuperable', *inservible* 'useless', *invers* 'inverse', *irreductible* 'unyielding', *laberíntic* 'labyrinthine', *llaminer* 'sweet-toothed | appetising', *malalt* 'ill', *morat* 'purple', *negatiu* 'negative', *nombrós* 'numerous', *penós* 'distressing', *preeminent* 'pre-eminent', *preponderant* 'preponderant', *raonable* 'reasonable', *real* 'real', *representatiu* 'representative', *sobrenatural* 'supernatural', *subsidiari* 'subsidiary', *supraracional* 'supra-rational', *trivial* 'trivial', *uniforme* 'uniform', *usual* 'usual', *utòpic* 'Utopian', *vitalista* 'vitalist(ic)'.
- **relational (R)**: *adquisitiu* 'acquisitive', *alfabètic* 'alphabetical', *carbònic* 'carbonic', *cervical* 'neck (attr.), cervical', *climatològic* 'climatologic', *col·laborador* 'collaborating', *curatiu* 'curative', *diofàntic* 'diophantic', *formatiu* 'formative', *freudià* 'Freudian', *governatiu* 'governmental', *indicador* 'indicating', *onomàstic* 'name (attr.), onomastic', *parlant* 'talking',

11 http://translate.google.com.

penitenciari 'penitentiary, prison (attr.)', *periglacial* 'periglacial', *pesquer* 'fishing', *petri* 'stony', *preescolar* 'preschool (attr.)', *protector* 'protecting', *salvador* 'rescueing', *sociocultural* 'sociocultural', *sud-africà* 'South African', *tàctil* 'tactile', *terciari* 'tertiary', *terminològic* 'terminological', *topogràfic* 'topographic(al)', *toràcic* 'thoracic', *vaginal* 'vaginal', *valencianoparlant* 'Valencian-speaking', *ventral* 'ventral', *veterinari* 'veterinary', *vocàlic* 'vocalic, vowel (attr.)', *xinès* 'Chinese'.

- **intensional-qualitative (IQ)**: *antic* 'ancient | former'.
- **qualitative-relational (QR)**: *alemany* 'German', *celest* 'celestial | sky blue', *contaminant* 'pollutant', *cultural* 'cultural', *femení* 'female (attr.) | feminine', *irònic* 'irony (attr.) | ironic', *menorquí* 'Menorcan', *militar* 'war (attr.) | military', *sonor* 'sound (attr.) | sonorous', *triomfal* 'triumphal | triumphant', *viril* 'man (attr.) | virile, manly'.

Gold standard for the experiments with the second model (Section 5).

- **qualitative (Q)**: *absort* 'absorbed', *aleatori* 'random', *altiu* 'haughty', *ample* 'wide', *animal* 'animal', *anòmal* 'anomalous', *baix* 'low', *benigne* 'benign', *bord* 'infertile (plant) | stroppy (person)', *caduc* 'deciduous', *calb* 'bald', *capaç* 'able', *cardinal* 'cardinal', *caut* 'cautious', *cèlebre* 'famous', *concret* 'concrete', *conservador* 'conservative', *contingent* 'contingent', *cru* 'raw | crude', *curull* 'full', *decisiu* 'decisive', *deficient* 'deficient, defective', *deliciós* 'delicious', *desproporcionat* 'disproportionate', *dificultós* 'difficult', *esquerre* 'left', *excels* 'sublime', *exquisit* 'exquisite', *fluix* 'weak | loose', *foll* 'crazy', *formidable* 'formidable | terrific', *franc* 'frank', *fresc* 'fresh', *gros* 'big', *gruixut* 'thick', *humil* 'humble', *igual* 'equal, alike', *imperfecte* 'imperfect', *impropi* 'improper', *incomplet* 'incomplet', *inhumà* 'inhuman', *insuficient* 'insufficient', *integral* 'integral | wholegrain', *íntegre* 'entire', intel·ligent 'intelligent', *intern* 'intern', *líquid* 'liquid', *llarg* 'long', *llis* 'smooth', *mal* 'bad', *màxim* 'maximum', *menor* 'minor | smaller | younger', *mínim* 'minimum', *moll* 'wet', *morat* 'purple', *mutu* 'mutual', *notori* 'notorious', *ocult* 'hidden', *opac* 'opaque', *paradoxal* 'paradoxical', *peculiar* 'peculiar', *perillós* 'dangerous', *pertinent* 'pertinent', *pessimista* 'pessimistic', *plàcid* 'placid', *precoç* 'precocious', *predilecte* 'favorite', *primari* 'primary', *primitiu* 'primitive', *propens* 'prone', *pròsper* 'prosperous', *prudent* 'prudent', *punxegut* 'sharp-pointed', *quadrat* 'square', *reaccionari* 'reactionary', *recent* 'recent', *recíproc* 'reciprocal', *remarcable* 'remarkable', *responsable* 'responsible', *rígid* 'rigid', *roent* 'burning', *sant* 'saint', *semicircular* 'semicircular', *seriós* 'serious', *significatiu* 'significant', *silenciós* 'silent', *similar* 'similar', *simplista* 'simplistic', *subaltern* 'subordinate', *sublim* 'sublime', *subsidiari* 'subsidiary', *subterrani* 'underground', *superflu* 'superfluous', *tenaç* 'tenacious', *terrible* 'terrible', *típic* 'typical', *titular* 'titular, official', *tort* 'bent', *total* 'total', *tou* 'soft', *triangular* 'triangular', *vague* 'vague', *ver* 'true', *viciós* 'vicious', *vigorós* 'vigorous', *viril* 'virile', *vulgar* 'vulgar'.
- **event-related (E)**: *abundant* 'abundant', *abundós* 'plentiful', *acompanyat* 'accompanied', *admirable* 'admirable', *contradictori* 'contradictory', *convincent* 'convincing', *creador* 'creative', *divergent* 'divergent', *encarregat*

'in charge', *exigent* 'demanding', *exportador* 'exporting', *immutable* 'immutable', *imperceptible* 'imperceptible', *informatiu* 'informative', *irat* 'angry', *matiner* 'who gets up early', *motor* 'motor', *oblidat* 'forgotten', *orientat* 'oriented', *picat* 'pricked | minced | offended', *preferible* 'preferable', *productor* 'producing', *promès* 'promised', *protector* 'protecting, protective', *receptor* 'receiving', *recomanat* 'recommended', *regulador* 'regulating', *resultant* 'resulting', *revelador* 'revealing', *salvador* 'savior', *satisfactori* 'satisfactory', *sospitós* 'suspicious | suspect', *temible* 'fearsome', *treballador* 'working', *variable* 'variable', *victoriós* 'victorious', *vivent* 'living'.

- **relational**: *americà* 'American', *angular* 'angular', *atòmic* 'atomic', *barceloní* 'Barcelonian', *calcari* 'calcareous', *causal* 'causal', *ciutadà* 'city (attr.)', *conflictiu* 'conflict (attr.)', *corporatiu* 'corporate', *cranià* 'skull (attr.)', *diari* 'daily', *elèctric* 'electric(al)', *epistemològic* 'epistemological', *escènic* 'scenic', *estacional* 'seasonal', *fangós* 'muddy', *imperial* 'imperial', *lleidatà* 'Leridan', *manresà* 'Manresan', *marxià* 'Marx (attr.)', *melòdic* 'melodic', *mercantil* 'mercantile', *obrer* 'working-class, labour (attr.)', *ontològic* 'ontological', *pasqual* 'paschal', *peninsular* 'peninsular', *renaixentista* 'Renaissance (attr.)', *respiratori* 'respiratory', *terrestre* 'terrestrial', *viari* 'road (attr.)'.
- **event-qualitative (EQ)**: *animat* 'animate | lively', *cridaner* 'who usually shouts | loud-colored', *embolicat* 'wrapped up | messy', *encantat* 'charmed | happy', *obert* 'opened | open', *raonable* 'that can be reasoned on | reasonable, fair', *sabut* 'known | wise'.
- **event-relational (ER)**: *comptable* 'countable | account (attr.)', *cooperatiu* 'cooperative | cooperative (attr.)', *digestiu* 'digestive | digestion (attr.)', *docent* 'teaching | educational', *nutritiu* 'nutritive | nutritional', *vegetatiu* 'vegetative | vegetation (attr.)'.
- **qualitative-relational (QR)**: *alegre* 'cheerful', *amorós* 'lovely | love (attr.)', *anarquista* 'anarchistic | anarchist', *capitalista* 'capitalistic | capitalist', *catalanista* 'Catalanistic | Catalanist', *comunista* 'communistic | communist', *diürn* 'diurnal, day (attr.)', *eròtic* 'erotic | love (attr.)', *familiar* 'familiar | family (attr.)', *feminista* 'feminist | feminism (attr.)', *humà* 'humane | human', *infantil* 'childish | child (attr.)', *intuïtiu* 'intuitive | intuition (attr.)', *local* 'local | place (attr.)', *nocturn* 'nocturnal, night (attr.)', *poètic* 'poetic, idealized | poetry (attr.)', *professional* '(worker) who works well | professional, job (attr.)', *revolucionari* 'revolutionary | revolution (attr.)', *sensitiu* 'sensitive | sensation (attr.)', *socialista* 'socialistic | socialist', *turístic* 'touristy | tourist (attr.)', *unitari* 'unitary | union (attr.)', *utilitari* 'utilitarian | utility (attr.)'.

Acknowledgments

The authors wish to thank Àngel Gil, Laia Mayol, Martí Quixal, and Roser Sanromà for participating in the annotation of the gold standards; David Farwell, Louise McNally, Sebastian Padó, and Martí Quixal for comments and discussion on previous versions of this article; Josep Maria Boleda, Montse Cuadros, and Edgar Gonzàlez for technical help; and the anonymous reviewers for their constructive criticism, which has greatly helped improve the article. This work has been supported via Ph.D. grants to the first author by the Generalitat de Catalunya (2001FI 00582), the Fundación Caja Madrid, and the Universitat Pompeu Fabra; also by the Ministry of

Education and the Ministry of Science and Technology of Spain under contracts FFI2010-09464-E (REDISIM), FFI2010-15006 (OntoSem 2), TIN2009-14715-C04-04 (KNOW2), and JCI2007-57-1479; and by the European Union via the EU PASCAL2 Network of Excellence (FP7-ICT-216886). The second author was funded by the DFG Collaborative Research Center 732.

References

Almuhareb, Abdulrahman and Massimo Poesio. 2004. Attribute-based and value-based clustering: An evaluation. In *Proceedings of EMNLP 2004*, pages 158–165, Barcelona.

Alonge, Antonietta, Francesca Bertagna, Nicoletta Calzolari, Adriana Roventini, and Antonio Zampolli. 2000. Encoding information on adjectives in a lexical-semantic net for computational applications. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL '00, pages 42–49, San Francisco, CA.

Alsina, Àlex, Toni Badia, Gemma Boleda, Stefan Bott, Àngel Gil, Martí Quixal, and Oriol Valentín. 2002. CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation*, volume III, pages 1130–1135, Las Palmas.

Ando, Rie Kubota. 2006. Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 77–84, New York City, NY.

Apresjan, Juri D. 1974. Regular polysemy. *Linguistics*, 142:5–32.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bally, Charles. 1944. *Linguistique générale et linguistique française*. A. Francke, Berne.

Banerjee, Arindam, Chase Krumpelman, Joydeep Ghosh, Sugato Basu, and Raymond J. Mooney. 2005. Model-based overlapping clustering. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 532–537, New York, NY.

Bohnet, Berndt, Stefan Klatt, and Leo Wanner. 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*, pages 24–33, Las Palmas.

Boleda, Gemma. 2007. *Automatic Acquisition of Semantic Classes for Adjectives*. Ph.D. thesis, Pompeu Fabra University.

Boleda, Gemma, Toni Badia, and Eloi Batlle. 2004. Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1119–1125, Morristown, NJ.

Boleda, Gemma, Sabine Schulte im Walde, and Toni Badia. 2008. Analysis of agreement on adjective semantic classification. *Research on Language and Computation*, 6(3):247–271.

Bouckaert, Remco R. 2004. Estimating replicability of classifier learning experiments. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 15–23, New York, NY.

Brants, Thorsten. 2000. Inter-annotator agreement for a German newspaper corpus. In *Second International Conference on Language Resources and Evaluation*, LREC '02, Athens.

Breiman, Leo. 2001. Random forests. *Machine Learning*, 45:5–23.

Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA, pages 3–86.

Bresnan, Joan. 1995. Lexicality and argument structure. Invited talk at the *Paris Syntax and Semantics Conference*. 12 October.

Brody, Samuel and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA.

Bryll, Robert K., Ricardo Gutierrez-Osuna, and Francis K. H. Quek. 2003. Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6):1291–1302.

Buitelaar, Paul. 1998. CoreLex: An ontology of systematic polysemous classes. In *Proceedings of Formal Ontologies in Information Systems*, pages 221–235, Amsterdam.

Burnage, Gavin and Dominic Dunlop. 1992. Encoding the British National Corpus. In

English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, pages 79–95, Amsterdam.

Carvalho, Paula and Elisabete Ranchhod. 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 EACL-03*, pages 105–112, Budapest.

Chao, Gerald and Michael G. Dyer. 2000. Word sense disambiguation of adjectives using probabilistic networks. In *Proceedings of the 18th Conference on Computational Linguistics (COLING-00)*, pages 152–158, Morristown, NJ.

Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. 2010. "Was it good? It was provocative." Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 167–176, Stroudsburg, PA.

Demonte, Violeta. 2011. Adjectives. In Klaus von Heusinger, Claudia Maienborn, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2. Mouton de Gruyter, Berlin, pages 1314–1340.

Dhillon, Inderjit S. and Dharmendra S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175.

Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Dietterich, Thomas G. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:5–23.

Dixon, Robert M. W. and Alexandra Y. Aikhenvald, editors. 2004. *Adjective Classes*. Oxford University Press, Oxford.

Dorr, Bonnie J. and Douglas Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 322–333, Morristown, NJ.

Erk, Katrin and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 92–97, Stroudsburg, PA.

Everitt, Brian S., Sabine Landau, and Morven Leese. 2001. *Cluster Analysis*. Arnold, London, 4th edition.

Freund, Yoav and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, ICML '96, pages 148–156, San Francisco, CA.

Ghamrawi, Nadia and Andrew McCallum. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 195–200, New York, NY.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 172–182, Morristown, NJ.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics (EACL'97)*, pages 174–181, Morristown, NJ.

Hatzivassiloglou, Vasileios and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 299–305, Morristown, NJ.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, ACL '90, pages 268–275, Stroudsburg, PA.

Ho, Tin Kam. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832–844.

Joanis, Eric, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Justeson, John S. and Slava M. Katz. 1995. Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, 21(1):1–27.

Karypis, George. 2002. CLUTO—A Clustering Toolkit. Technical Report TR 02-017, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN.

Kaufman, Leonard and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, New York.

Kohomban, Upali S. and Wee Sun Lee. 2005. Learning semantic classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 34–41, Ann Arbor, MI.

Korhonen, Anna, Yuval Krymolowski, and Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics - Volume 1*, ACL '03, pages 64–71, Stroudsburg, PA.

Lapata, Maria. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA.

Lapata, Mirella. 2000. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Ph.D. thesis, University of Edinburgh.

Lapata, Mirella and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(2):45–73.

Levin, Beth and Malka Rappaport. 1986. The formation of adjectival passives. *Linguistic Inquiry*, 17:623–661.

Malouf, Robert. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, ACL '00, pages 85–92, Stroudsburg, PA.

Marcus, M., B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Mayol, Laia, Gemma Boleda, and Toni Badia. 2005. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39(4):295–312.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL '00, pages 256–263, San Francisco, CA.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL '04, pages 279–286, Stroudsburg, PA.

McDonald, Ryan, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 987–994, Stroudsburg, PA.

McNally, Louise and Gemma Boleda. 2004. Relational adjectives as properties of kinds. *Empirical Issues in Syntax and Semantics*, 5:179–196.

Merlo, Paola and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Miller, Katharine J. 1998. Modifiers in WordNet. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*. The MIT Press, Cambridge, MA, pages 47–67.

Montague, Richard. 1974. English as a formal language. In Richmond H. Thomason, editor, *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven, CT, chapter 6, pages 188–221.

Murphy, Gregory L. 2002. *The Big Book of Concepts*. The MIT Press, Cambridge, MA.

Nadeau, Claude and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52(3):239–281.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:10:1–10:69.

Nirenburg, Sergei and Victor Raskin. 2004. *Ontological Semantics*. The MIT Press, Cambridge, MA.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ACL '93, pages 183–190, Stroudsburg, PA.

Picallo, Carme. 2002. L'adjectiu i el sintagma adjectival. In Joan Solà, editor, *Gramàtica del català contemporani*. Empúries, Barcelona, pages 1643–1688.
Poesio, Massimo and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 76–83, Stroudsburg, PA.
Prescher, Detlef, Stefan Riezler, and Mats Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 649–655, Stroudsburg, PA.
Pustejovsky, James. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.
Quinlan, Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.
Rafel, Joaquim. 1994. Un corpus general de referència de la llengua catalana. *Caplletra*, 17:219–250.
Raskin, Victor and Sergei Nirenburg. 1998. An applied ontological semantic microtheory of adjective meaning for natural language processing. *Machine Translation*, 13(2-3):135–227.
Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.
Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA.
Sanromà, Roser and Gemma Boleda. 2010. The database of Catalan adjectives. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, LREC '10, Valletta.
Schapire, Robert E. and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168.
Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.
Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
Utt, Jason and Sebastian Padó. 2011. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 265–274, Stroudsburg, PA.
Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review*, 66:143–60.
Véronis, Jean. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and Advanced Papers of the Senseval Workshop*, pages 2–4, Herstmonceux Castle.
Verzani, John. 2005. *Using R for Introductory Statistics*. Chapman & Hall/CRC, Boca Raton, FL.
Wiebe, Janyce M., Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.
Witten, Ian H. and Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, Amsterdam, 3rd edition.
Yallop, Jeremy, Anna Korhonen, and Ted Briscoe. 2005. Automatic acquisition of adjectival subcategorization from corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 614–621, Stroudsburg, PA.