Research on Language and Computation **00:** 1–28, 2009.

An Analysis of Human Judgements on Semantic Classification of Catalan Adjectives**

GEMMA BOLEDA (gboleda@lsi.upc.edu)
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, 08034, Spain

SABINE SCHULTE IM WALDE (schulte@ims.uni-stuttgart.de)
Institute for Natural Language Processing,
University of Stuttgart,
70174 Stuttgart, Germany

TONI BADIA (toni.badia@upf.edu)
GLiCom,
Fundació Barcelona Media and Universitat Pompeu Fabra,
Barcelona 08003, Spain

February 11, 2009

Abstract. This article reports on a large-scale experiment for gathering human judgements with respect to a semantic classification of Catalan adjectives. The goal of our experiment was to classify 210 Catalan adjectives as basic, event-related, or object-related adjectives, allowing for multiple class assignments to account for polysemy. The experiment was directed at non-expert native speakers and administered via the Web, collecting data from 322 participants. We assess the degree of inter-annotator agreement through an innovative methodology based on observed agreement and kappa, and use weighted versions of these measures to account for partial agreement in polysemous assignments. Because the obtained scores (kappa 0.20-0.34) are too low to establish a reliably labelled dataset, we then perform a series of post-hoc analyses on the human judgements to investigate the sources of disagreement, by comparing the participants' classifications with a classification obtained from experts. Our analysis shows that polysemous items and event-related adjectives are more problematic than other types of adjectives. Furthermore, the analysis helps to distinguish disagreement caused by the task as opposed that caused by the experimental design, thus pointing to specific difficulties in both aspects of the research. The methodology developed for this analysis might therefore prove useful for the design of experiments for related tasks.

Key words: adjectives, Catalan, human judgements, inter-annotator agreement, semantic classes, web experiment

** The original publication is available at www.springerlink.com, doi 10.1007/s11168-008-9056-4.

1. Introduction

Human judgements play a key role in the development and the assessment of linguistic resources and methods in Computational Linguistics. For example, the annotation of a corpus requires the definition of guidelines, i.e., an inventory of categories as well as instructions on how to apply them, which are followed by the annotators when tagging the text. However, in many cases, neither an off-the-shelf inventory of categories nor a straightforward set of application criteria are available. Human judgements (e.g., gathered in a pilot annotation study), can be used to develop and fine-tune such guidelines. Furthermore, collecting human judgements in Computational Linguistics is typically not an end-task by itself, but an intermediate task to create a gold standard that is useful for training and evaluating NLP systems. For a gold standard to be reliable, though, independent judges have to arrive at similar decisions (Krippendorff, 2004). Thus, the production of reliable gold standard resources requires the development of solid methodologies for gathering human judgements and assessing the degree of agreement between them. Last but not least, systematically collected human judgements provide clues for research on linguistic issues that can not be easily obtained from an introspective analysis (because they provide many independent judgements) or from corpus data (because the target judgements concern aspects not readily provided by corpus data, such as semantic classes).

Experiments that gather human judgements on linguistic phenomena are, however, very difficult to design for two main reasons. First, the agreement between annotators decreases with the complexity of the task (Artstein and Poesio, pear). Second, in order to obtain judgements on a large scale, the experiments need to address non-expert participants in addition to expert participants. In fact, the use of naive subjects for linguistic tasks is not uncommon in Computational Linguistics (for instance, Fellbaum et al. (1998) compared naive and lexicographer subjects in the task of tagging a text with WordNet senses; Artstein and Poesio (2005) used 18 naive subjects for coreference tagging), but is deemed to cause difficulties for the non-expert judges if linguistic background is required.

This article reports on a large-scale experiment for gathering human judgements with respect to a semantic classification of Catalan adjectives. The specific goal of our experiment was to classify 210 Catalan adjectives as basic, event-related, or object-related adjectives, allowing for multiple class assignments to account for polysemy. The resulting classification was aimed at building a gold standard for lexical acquisition experiments with Machine Learning techniques. Furthermore, as the semantic classification of Catalan adjectives is not well established from a theoretical point of view, the experimental data were also expected to provide insight into adjective semantics.

In order to check the reliability of the human data through agreement measurement, we propose two methodological innovations to assess agreement in large-scale annotation experiments involving polysemy: (i) the computation of three different agreement scores, corresponding to the partial matches in polysemous assignments, and (ii) a robust method to compute confidence intervals for agreement data.

The resulting agreement scores obtained from our data are too low to establish a reliably labelled dataset. Thus, we perform a series of post-hoc analyses on the human judgements, (i) comparing the participants' classifications with a classification obtained from experts, and (ii) identifying types of adjectives that pose special difficulties to participants. Our analysis shows how the data provide insight into linguistic issues that are relevant for the semantic classification of adjectives. Furthermore, the analysis helps to distinguish disagreement caused by the classification scheme as opposed to the experimental design. We believe that such post-hoc analyses should be an integral part of experiments that collect human judgements. In that respect, our results might prove useful for the design of related experiments.

The article is structured as follows. Section 2 introduces the aspects of the target classification that are relevant to the experiment design, and Section 3 reviews the experimental method and data collection procedures. The agreement results and the post-hoc analyses are presented in Sections 4 and 5, respectively, and Section 6 finishes with some conclusions.

2. Classification

The definition and characterisation of our target semantic classification closely follows the proposal by Raskin and Nirenburg (1998) within the framework of Ontological Semantics (Nirenburg and Raskin, 2004).[1] In Ontological Semantics, an ontology of concepts modelling the world is explicitly defined and the semantics of words is provided by mapping the words onto elements of the ontology. The classification pursued in this article is based on the ontological sort of adjectival denotation: all adjectives denote properties, and these properties can be instantiated as simple attributes (*basic adjectives*), relationships to objects (*object-related adjectives*), or relationships to events (*event-related adjectives*).

Basic adjectives are the prototypical adjectives which denote attributes or properties that cannot be decomposed further (such as *bonic* 'beautiful', *gran* 'big'). In Ontological Semantics, these adjectives are mapped to concepts of type *attribute*. For instance, the semantics of the adjective *gran* specifies a mapping to the *size-attribute* element in the ontology. ***Event-related adjectives*** bear a reference to an event and are therefore mapped onto *event* concepts in the ontology. For instance, if something is *tangible* ('tangible'), then it can be touched. The semantics of *tangible* includes a

pointer to the event element *touch* in the ontology, together with a modality value to account for the meaning introduced by the *-ble* morpheme (Raskin and Nirenburg, 1998, p. 187ff.). Similarly, ***object-related adjectives*** are mapped onto object concepts in the ontology because they have an embedded object component in their meaning: *Deformació nasal* ('nasal deformity') can be paraphrased as *deformity that affects the nose*, so *nasal* evokes the object *nose*. This class of adjectives has been discussed in Romance linguistics at least since Bally (1944) and has recently received attention from semantic theory (Bosque and Picallo, 1996; McNally and Boleda, 2004).

Our interest in classifying adjectives is motivated by the fact that adjectives play an important role in sentential semantics: They are crucial in determining the reference of NPs, and in defining properties of entities. Establishing the semantic class of an adjective is a first step towards specifying its lexical semantic properties; further properties might be added in a subsequent step. As mentioned in the Introduction, so far, there has been little work on the semantic classification of adjectives (as opposed to, e.g., verbal semantic classification). Thus, we deliberately decided in favour of a small-scale, broad classification consisting of three classes, which can be refined and extended in subsequent work.

Our target classification as described above is semantic in nature. However, the semantic distinctions also correspond to distinctions at other levels of linguistic description, most notably, morphology and syntax. For instance, there is a clear relationship between morphological type and semantic class in Catalan: Basic adjectives are typically morphologically simple (non-derived), object-related adjectives tend to be denominal, and event adjectives are usually deverbal. This is the default mapping that one expects from the morphology-semantics interface. As for the syntax-semantics interface, basic adjectives in Catalan can be used as pre-nominal modifiers (mostly in non-restrictive uses) and also as predicates, while object adjectives typically cannot. The interfaces between the linguistic levels enable theoretical and computational work to exploit various cues to the semantic class of a particular adjective.

However, the correspondences between these linguistic properties and adjectival semantic classes are not one-to-one mappings. Taking the morphological level as an example, there are denominal adjectives which are basic (such as *vergonyós* 'shy', from *vergonya* 'shyness'). Conversely, some object adjectives are not synchronically denominal (such as *botànic* 'botanical') and some deverbal adjectives are not event-related, such as *amable* (*lit.* 'suitable to be loved'; has evolved to 'kind, friendly'). Furthermore, our classification (like any classification concerning lexical semantics) is affected by polysemy, i.e., some adjectives belong to more than one class. For instance, *familiar* has an object reading (related to the object 'family'), and a basic reading (corresponding to the English adjective 'familiar'). The two readings are

exemplified in (1). Similarly, the participial adjective *sabut* ('known') has an event sense corresponding to the verb *saber* ('know') and a basic sense equivalent to 'wise', as exemplified in (2).

(1) reunió familiar / cara familiar
meeting familiar / face familiar

'family meeting / familiar face'

(2) conseqüència sabuda / home sabut
consequence known / man wise

'known consequence / wise man'

Note, however, that not all cases of adjectival polysemy can be modelled in terms of semantic class alternation. For example, the two senses of *llarg* as in *discurs llarg / carrer llarg* ('long speech / long street'), and also the two senses of *trist* in *noi trist / pel·lícula trista* ('sad boy / sad film'), as discussed in Pustejovsky (1995), all correspond to the basic class. Within this article, we concentrate on polysemy that is between our classes, as exemplified in (1) and (2).

3. Experiment design

This section describes our web experiment to collect the human judgements on adjective classification, introducing the material (Section 3.1), the experiment design (Section 3.2), the participants (Section 3.3), and the collected data (Section 3.4).

3.1. MATERIAL

We selected 210 adjective lemmata from a manually developed database of Catalan adjectives (Sanromà, 2003). The database contained morphological information, namely, the derivational type of an adjective (whether it is denominal, deverbal, participial, or non-derived), and its suffix, in case it is derived. Information on each adjective's frequency in a balanced, 14.5 million word, Catalan corpus (Rafel, 1994) was also recorded, and only adjectives with at least 50 occurrences in the corpus were included in the database. The sample comprises approximately 10% of all adjectives in the database, and is representative of adjectives in Catalan, being balanced for three possible sources of variation: frequency, derivational type, and suffix. We next motivate and explain the sampling scheme.

Frequency: More frequent words exhibit a higher degree of polysemy (Zipf, 1949). To control for this factor, we divided the adjectives into three frequency bands (high, medium, low), based on an equal division of the

range of log-transformed frequencies, and randomly selected 70 adjectives from each band. Lapata et al. (1999) used the same procedure to choose material for plausibility ratings concerning adjective-noun combinations.

Morphology (derivational type and suffix): As explained in Section 2, there is a strong relationship between the morphological type and the semantic class of Catalan adjectives. To promote semantic variability, thus, it is reasonable to control for morphological variability. However, the derivational types (denominal, deverbal, participial, or non-derived) are not evenly distributed: For example, there are only 399 deverbal adjectives in the database, as opposed to 860 denominal adjectives. Moreover, the distribution of adjectives is particularly skewed with respect to the suffix within each of the denominal and deverbal groups. We therefore designed a stratified sampling approach to morphology, and took an (approximately) equal number of adjectives from each derivational type and from each suffix. The exception were suffixes with very few lemmata (less than 20), which were gathered in one common group.

The distribution of the adjectives in the experiment sample is shown in Table I, which lists the number of suffixes (second column) and the number of lemmata from each derivational type in each frequency band (columns 3-6). The table also demonstrates that there were equal or similar distributions among derivational types (non-derived, denominal, and deverbal; 70 adjectives each) and frequency bands (approximately 70 adjectives for each band).

Table I. Stratification of the adjective selection.

Morph. type	**# Suffixes**	**Low**	**Medium**	**High**	**Total**
non-derived	-	23	24	23	70
denominal	8	24	23	23	70
deverbal	6	25	27	18	70
total	14	72	74	64	210

The sample was randomly divided into 7 test sets with 30 adjectives each, and each participant of the experiment was randomly assigned one of the sets (see next section). The reason for this procedure was that we wanted the experiment to last about 30 minutes on average because longer experiments tend to discourage participation and decrease concentration.

3.2. DESIGN

Recall that the goal of our experiment was to classify the 210 Catalan adjectives in the sample as basic, event, or object, allowing for multiple class

assignments to account for polysemy. The most direct method to collect human judgements on adjective classes would have been to ask participants to assign class labels to the adjectives. However, we took into account (a) that the experiment addressed non-expert participants, and (b) that there was no pre-existing classification of Catalan adjectives and therefore the classification proposal introduced in Section 2 was to be assessed. Therefore, participants were asked to *define*, rather than *classify*, the adjectives according to pre-defined patterns. Each pattern corresponds to a semantic class and was realised by a paraphrase. We thus gathered judgements of native speakers with respect to paraphrased relationships between lexical items. Note that paraphrases are among the types of linguistic evidence most often used by semanticists (Chierchia and McConnell-Ginet, 2000).

Participants were asked to complete one or more patterns for each adjective by filling in a blank field corresponding to a noun, verb, or adjective (depending on the pattern). Completing a pattern (indicated as $\square$ in the examples that follow) implied selecting a definitional pattern and thus a particular kind of meaning or semantic class. The fact that participants had to provide information to fill in the blank instead of simply selecting the pattern ensured their full attention to the task, and also served to indicate which sense was perceived in each case. Each field was accompanied by an indication of the expected part of speech (adjective, noun or verb), so as to further constrain the task. Note that this design requires participants to be familiar with some linguistic notions, but these are very basic notions which are acquired in primary school in Spain.

We defined five patterns, and all patterns were available for the participants for each adjective to be classified. For basic adjectives, the definitional pattern was to be completed with a synonym or an antonym, since basic adjectives typically have lexical antonyms or near-antonyms (see Miller, 1998). The definitional pattern is given in (3a) and exemplified in (3b).

(3) a. Té un significat semblant a / contrari a $\square_{(adjectiu)}$
'Has a meaning similar to / opposite to $\square_{(adjective)}$'

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

For object-related adjectives, the definitional pattern expressed the relationship to an object lexicalised through a noun, as shown in (4).

(4) a. Relatiu a o relacionat amb (/el/la/els/les/l') $\square_{(nom)}$
'Related to (the) $\square_{(noun)}$'

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

For event-related adjectives, the definitional pattern expressed the relationship to an event lexicalised through a verb. Three definitional patterns were provided to account for the different meanings arising from different suffixation processes: an "active" meaning for suffixes such as *-iu* or *-or* (pattern in (5)), a "passive" meaning for the *-ble* suffix (pattern in (6)), and a resultative meaning for participial adjectives (pattern in (7)).

(5) a. que $\square_{(verb)}$
'that/which/who $\square_{(verb)}$'

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

(6) a. que pot ser $\square_{(verb)}$
'that can be $\square_{(verb)}$'

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

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

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

No instructions were provided as to how to use the patterns because reading too many instructions discourages participation. However, the general instructions provided some examples, and the participants were made to go through three trial adjectives (for which they were shown the expected answers) so as to clarify the task. Following standards in psycholinguistic research, no example sentences were provided for the adjectives during the experiment, so as not to bias the subjects' responses. Recall that participants could select more than one pattern in case of polysemy. This concept was not mentioned in the instructions, but an example was provided along with an explanation.

The experiment was performed via the Web. Web experiments are among the easiest ways to carry out large-scale experiments, as they allow a potentially larger quantity and variety of data to be gathered than traditional, laboratory-based experiments, at virtually no cost (Reips, 2002). In recent years, web experiments have been applied to gather psycholinguistic evidence

for computational linguistic tasks (Lapata et al., 1999; Corley and Scheepers, 2002; Melinger and Schulte im Walde, 2005).

Before launching the experiment, we performed a pilot study with 85 subjects, which altered the following aspects of the experiment design: (a) Initially, we set no constraint on the maximal number of definitional patterns to be selected. Our assumption was that at most two patterns would be enough to account for polysemy in our setting because much of the polysemy occurs within two classes. The pilot study confirmed this assumption and we therefore decided to explicitly ask participants to fill in only one or two of the patterns. This decision makes the task clearer and the analysis of the results easier, without significantly decreasing descriptive accuracy. (b) In the pilot study, the order of the definitional patterns was always the same (first the object pattern, then the three event patterns, then the basic pattern). Since we observed an overuse of the object pattern, in the final design the order of the patterns was randomised to avoid ordering effects. The final experiment was structured as follows:

- first page with introduction and classificatory questions (cf. Table II),
- second page with instructions and examples,
- three training adjectives, where participants were given the expected answer after they filled in the blank,
- actual experiment: 1 page per adjective (30 adjectives),
- final "thank-you" page, with a small explanation of the purpose of the experiment and the possibility for the participant to write a comment.

As mentioned in Section 3.1, for each participant, one of the 7 sample sets was randomly chosen, and the order of the 30 adjectives to be judged was also randomised.

3.3. PARTICIPANTS

603 subjects took part in the Web experiment. Participants were recruited via e-mail from several university departments and distribution lists, and received no payment.[2] To encourage participants to reveal their e-mail address, so that they would commit themselves to the experiment (Reips, 2002), we offered as prices 2 vouchers of 30 euros each.

Of the 603 participants, 101 (17%) only read instructions without classifying a single adjective. 131 (22%) filled in too little for results to be analysed (we set the threshold at 20 adjectives – 66% of the material – to be classified). The dropout rate, thus, seems to be quite high (39%), although we have not found reported dropout ratios for similar Web experiments for comparison. Finally, 15 (2%) participants filled in 3 patterns or more for at least 20 adjectives, and were excluded from the analysis. Table II describes the remaining 322 participants, from which the data analysed in the remainder of the article was collected.

Table II. Main characteristics of participants in Web experiment. *NR* stands for *not reported*.

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

3.4. DATA COLLECTION AND CLEANING

The data were collected in March 2006. The responses were checked for compliance with instructions by a semi-automatic procedure, and the following types of responses were discarded:

- Responses with three or more filled patterns.
- Responses composed of more than one word, with some exceptions such as compound nouns (*ésser humà* 'human being').
- Responses with a part of speech other than that indicated in the instructions.
- Non-existing words (see example (8); presumably, time constraints and performance pressure led to participants making words up).

(8) mutu → *mutuar
mutual → ? (non-existing deadjectival verb)

Spelling mistakes were corrected for normalisation. The total number of errors detected (358) corresponds to 3.2% of the data. For comparison, Corley and Scheepers (2002) excluded 3% of their experimental data in a Web-based syntactic priming experiment because the prime-to-target times were too long. Our noisy data has a similar proportion.

Almost two thirds of the errors were due to two types of errors which pointed to problems in the experimental design. First, one of the event patterns ('that/which/who $\square\!\!\square\!\!\square\!\!\square\!\!\square\!\!\square$(*verb*)') produced 131 multiple word errors, indicating that the pattern was not constrained enough. In addition, many dictionary entries for non-event adjectives begin with *que* ('that'). For instance, the definition of *abrupte* ('abrupt') in a standard Catalan dictionary (Institut d'Estudis Catalans, 1997) is *que presenta transicions sobtades o brusques* ('that presents sudden transitions'). Choosing a dictionary-like construction for the erroneous event pattern was thus a sub-optimal design

decision. Second, the basic pattern ('has a meaning similar to / opposite to $\square_{(adjective)}$') produced 92 errors where a wrong POS (mainly, a noun) was provided. There are presumably two main reasons for this: (a) The large proportion of part of speech ambiguity between adjective and noun in Catalan (Boleda, 2007), which produced responses corresponding to the noun homograph and not to the adjective (as in *obrer* → *patró* 'working-class$_{adjective}$' → 'boss'); (b) the notion of similarity of meaning (as glossed in the definitional pattern) is quite vague, and various types of semantic relationships other than synonymy or antonymy fit in, as in *alegre* → *tristesa* ('joyful' → 'sadness').

4. Measuring inter-annotator agreement

As stated in the introduction, creating a dataset on the basis of human judgements requires the collected data to be reliable. One of the main conditions for reliability is reproducibility, which in our case means that independently working subjects should arrive at a very similar classification (Krippendorff, 2004). This section is therefore concerned with analysing the extent to which the participants in our experiment agree in the classification they implicitly provide.

The assessment of inter-annotator agreement is a complex area, and statisticians do not agree on a single method or approach to address it in a variety of settings, or even within a single setting. Accordingly, this issue has also been a focus of ongoing discussions in Computational Linguistics (Carletta, 1996; Di Eugenio and Glass, 2004; Artstein and Poesio, pear). Due to space constraints, the discussion that follows is restricted to the aspects that are most relevant for our experimental setting.

4.1. METHODOLOGY FOR MEASURING AGREEMENT

4.1.1. *Overall proportion of agreement and kappa*

The most straightforward measure for agreement (and the most widely used measure) is *observed agreement*, or p_o, the proportion of cases where subjects agree in their judgement (Hripcsak and Heitjan, 2002). For two annotators and multiple categories C_i, p_o can be formalised as follows. Using a CxC contingency table, where C is the number of categories, and where the rows and columns correspond to the classifications provided by the two annotators, each cell n_{ij} represents the number of elements that Annotator 1 assigns to category C_i and Annotator 2 to category C_j. Equation 1 shows how p_o is computed. Because cases corresponding to agreement lie at the diagonal of the contingency table, and cases corresponding to disagreement are off-

diagonal, Equation 1 sums over the diagonal cells n_{ii} and normalises the sum by the total number of cases (N). This measure ranges between 0 and 1.

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

The formula yields an intuitive measure for inter-annotator agreement. However, it runs into problems when the categories are unevenly distributed (Hripcsak and Heitjan, 2002; Di Eugenio and Glass, 2004, among others): If most objects belong to one of the categories, the annotators achieve a high p_o just by chance. Also, annotators are likely to agree more by chance if the number of categories is small, regardless of their relative frequencies. These considerations have led scholars to propose indices that correct the observed agreement for chance, factoring out the agreement that would be expected if annotators provided their judgements just randomly. The general form of the corrected indices is provided by Equation 2, where p_o represents observed agreement (as in Equation 1), and p_e the agreement expected by chance. The denominator normalises the scale so that the scores range between -1 and 1; 1 indicates perfect agreement, 0 agreement by chance, and values below 0 some kind of systematic disagreement (Fleiss, 1981; Carletta, 1996; Artstein and Poesio, pear).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

The major difference among the corrected indices is the way the expected agreement (p_e) is modelled, that is, what the prior probabilities of each category are. We use Cohen's kappa (Cohen, 1960, see Equation 3), one of the most widely used indices in Computational Linguistics. In this case, the expected agreement is computed as the sum of the products of the marginal proportions. This computation assumes that "random assignment of categories to items is governed by prior distributions that are unique to each coder, and which reflect individual annotator bias" (Artstein and Poesio, pear).

$$p_e = \frac{1}{N^2} \sum_{i=1}^{C} n_{i.}\, n_{.i} \tag{3}$$

4.1.2. *Estimation of standard error*

No matter which agreement measures are used, their values are estimated from a sample only (i.e., the set of coders and the set of objects coded), and thus are subject to sampling error. It is therefore important to report the standard error (or confidence interval) in addition to the obtained agreement

scores, to give an estimate of the accuracy with which the sample values approach the population values (the "real" agreement values for our task). This issue is generally ignored in the Computational Linguistics literature, although it is discussed in other fields (Fleiss, 1981; Lui et al., 1999; Altaye et al., 2001; Krippendorff, 2004).

Typically, agreement scores are computed with a relatively large number of objects to be classified and a small number of subjects to classify them. The proposals for standard error computation in the literature mentioned in the previous paragraph correspond to this type of situation. Our situation, however, is the reverse: We have a large number of subjects for each object (32 to 59 annotators per adjective) and a small number of objects per subject (about 30). Giving consideration to this difference, one could compute a single agreement score for each of the over 7,000 annotator pairs arising from these data, and then compute a confidence interval based on these scores. However, this procedure would not be correct, because the data are not independent: each subject participates in more than one pair. We therefore propose an alternative procedure, i.e. a random assignment of subjects to pairs of subjects, such that each subject only participates in one pair. The agreement scores for pairs of subjects form a distribution with independent values, and the confidence interval can be estimated in the standard way using the t-distribution, assuming that the data are normally distributed. Equation 4 shows the general formula for confidence interval estimation, where $\bar{x}$ is the sample mean, s the sample standard deviation, N the sample size, α the significance level, and $t_{\alpha/2}$ the value from the t-distribution corresponding to the relevant significance level and degrees of freedom.

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{N}} \tag{4}$$

Our procedure corresponds to the usual practice (in medicine and other fields) of reporting mean kappa values when multiple subjects are involved, as mean pair-wise values are an approximation of multi-subject agreement. Although robust, the solution is not optimal in that it compares each subject with only one randomly chosen subject. Furthermore, it requires a large number of annotators per object, so it is only applicable to large-scale experiments.

4.1.3. *Weighted agreement for polysemy judgements*

One of the most challenging aspects in linguistic tasks, particularly with respect to lexical semantics, is the assessment of agreement when multiple categories are allowed, as is the case with polysemy judgements. Recall that we allowed subjects to select more than one definitional pattern, that is, to assign adjectives to more than one class in case of polysemy. In this setting,

partial agreement arises when some, but not all, class assignments coincide, which has to be taken into account when measuring agreement. The need to account for partial agreement also arises in other linguistic tasks, such as anaphoric relation annotation (Passonneau, 2004; Artstein and Poesio, pear).

To estimate the agreement values under different considerations of partial agreement, we use three definitions of agreement: *full agreement* requires all class assignments to coincide, *weighted agreement* gives some credit to partial matches, and *overlapping agreement* gives full credit to partial matches. Probably, full agreement is too strict and overlapping agreement to lax a definition of agreement; however, they serve as lower and upper bounds, respectively, for the actual agreement score to be estimated.

Full agreement relies on p_o and κ, as defined in Section 4.1.1 (Equations 1 to 3); weighted agreement uses *weighted kappa* (Cohen, 1968): wp_o, wp_e and $w\kappa$ are defined in Equations 5 to 7. The definitions are equivalent to their unweighted versions in Equations 1 to 3, but all cells in the CxC contingency table are considered instead of only the diagonal, and can potentially add some value to the final score, depending on the value of their associated weight w_{ij}.[3]

$$wp_o = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{C} w_{ij}\, n_{ij} \tag{5}$$

$$wp_e = \frac{1}{N^2} \sum_{i=1}^{C} \sum_{j=1}^{C} w_{ij} n_{i.}\, n_{.j} \tag{6}$$

$$w\kappa = \frac{wp_o - wp_e}{1 - wp_e} \tag{7}$$

The drawback of this measure is that, in general, the weighting scheme used is difficult to justify on independent grounds, and the obtained values vary substantially depending on the weighting scheme (Artstein and Poesio, pear). In our setting, it is possible to justify the weighting scheme by how judges make their decisions: assuming that subjects make three independent decisions (basic/non-basic, event/non-event, object/non-object)[4], we can assign equal weight (1/3) to each of the decisions, and thus arrive at the weighting scheme in Table III.[5] However, this approach implies assigning a weight of 1/3 to monosemous non-agreement (in cases where, e.g., one subject chooses the class *basic* and another one chooses the class *event*), because there is implicit agreement on not choosing the *object* class. We assign 0 in this situation by placing a further restriction on the weighting scheme, namely, that for weight w_{ij} to be > 0, there has to be at least one

Table III. Agreement weights for polysemous assignments.

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

positive agreement. This approach to weighting can be generalised to other tasks involving polysemy.

Weighted kappa also offers a natural way to accommodate the notion of overlapping agreement, namely, to assign a weight of 1 to all cells where there is some overlap between the categories involved. To compute overlapping agreement, thus, all non-zero cells in Table III would contain a 1.

4.2. AGREEMENT RESULTS

This section discusses the agreement scores obtained for our experiment. In all analyses to follow, we take the three eventive definitional patterns presented in examples (5-7) as indicative of a single class (the event class). Taking into account that a maximum of two patterns per adjective was allowed, there are six possible classifications for a given adjective:[6]

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

Recall from Section 4.1.2 that – for each test set – we calculated agreement scores for random pairs of subjects. The available number of subject pairs per test set ranges between 19 and 29.[7] The agreement scores of full, weighted, and overlapping agreement were obtained as follows. For each test set, the mean agreement scores were computed, and 95% confidence intervals were obtained using the t distribution. Table IV reports the observed agreement (p_o; wp_o for weighted p_o; op_o for overlapping p_o) and the kappa values (κ, $w\kappa$, and $o\kappa$), averaged over all test sets.

Table IV shows that in the most strict definition (row *full*), the p_o values for our task are between 0.37 and 0.51, and the κ values are between 0.20 and 0.34. These values represent a very low level of agreement. At the other end, the scores for op_o and $o\kappa$ (row *overlapping*) range between 0.73 and 0.83, and between 0.42 and 0.60, respectively. The two measures provide the lower and upper bounds for the agreement values, as discussed earlier. The values of the weighted observed agreement and kappa (row *weighted*) are

Table IV. Overall agreement values.

Agreement	Measure	Mean	Confidence Interval
full	p_o	0.44	0.37-0.51
	κ	0.27	0.20-0.34
weighted	wp_o	0.66	0.62-0.70
	$w\kappa$	0.38	0.31-0.45
overlapping	op_o	0.78	0.73-0.83
	$o\kappa$	0.51	0.42-0.60

between the p_o/κ and $op_o/o\kappa$ values, as expected: wp_o ranges from 0.62 to 0.70, and $w\kappa$ from 0.31 to 0.45. The weighting scheme in Table III therefore appears to account for partial agreement in a sensible manner.

Summarising the agreement results, the kappa value for our task is higher than 0.20 (lower extreme of the confidence interval for κ) and lower than 0.60 (upper extreme of the confidence interval for $o\kappa$). We consider the best estimate to correspond to $w\kappa$, so that the kappa of the Web experiment ranges from 0.31 to 0.45. This range is very low, too low in fact to consider the data to be reliable. Krippendorff (1980) demands as a very minimum a 0.67 value for his α measure (which yields slightly lower values than κ), and only considers values over 0.8 to be sufficient for reliability. According to Fleiss (1981), our scores represent poor to fair agreement, and Landis and Koch (1977) would consider them to be fair to moderate.

It is generally the case that in studies involving human judgements on semantics or discourse, high agreement values are very difficult to obtain. For example, the already mentioned study by Poesio and Artstein (2005) analysed an experiment in which 18 subjects tagged anaphoric relations. The authors reported κ values around 0.63-0.66, and noted that if a trivial category is dropped, κ drops to 0.45-0.50. Merlo and Stevenson (2001) discussed a classification of verbs into unergative, unaccusative, and object-drop. Three subjects with a high level of expertise tagged 59 verbs. Despite the expertise of the subjects, their kappa scores range between 0.53 and 0.66 (p_o 0.70 to 0.77). Véronis (1998) reported on experiments on tagging senses of French words. Six students of linguistics with no training in lexicography tagged 60 highly polysemous words (20 adjectives, 20 nouns and 20 verbs) with the set of senses listed in the *Petit Larousse* dictionary. The resulting pair-wise p_o was around 0.69 and weighted kappa around 0.43.

All these values are well below the 0.8 threshold for kappa, which can be interpreted as indicating that the field of computational semantics is not mature enough to yield reliable classifications. However, most of the values

reported are higher than our 0.31-0.45 values. While the figures are not entirely comparable (parameters such as the number and distribution of the classes and the evaluation procedures in the studies cited differ from the one presented here), they indicate that the agreement we obtained is lower than that obtained in related tasks. We next explore some explanations for the low agreement.

5. Exploring the sources of disagreement

The results described in the previous section warrant a study of the factors causing the low agreement. We carry out a two-fold analysis. First, we compare the participants' classifications with a classification obtained from experts (Section 5.1), identifying problems in the experimental design. Second, we analyse intra-item agreement to spot types of adjectives that pose special difficulties to participants (Section 5.2), thus providing insight into theoretical issues that are relevant for the semantic classification of adjectives.

5.1. EXPERT GOLD STANDARD AND PARTICIPANTS' CLASSIFICATIONS

The first part of our analysis uses an expert gold standard for the adjectives under consideration and compares it against the classification data from the participants, with the main goal of detecting biases or problems in the design of the experiment.

The expert classification was obtained from a committee of three experts in lexical semantics (two of the authors of the article and a researcher pursuing a PhD on Catalan adjectives) in three 2-hour sessions. They reviewed each of the 210 adjectives in the gold standard, and assigned them to one or two semantic classes. The experts based their decisions on their own intuitions, a Catalan dictionary (Institut d'Estudis Catalans, 1997), corpus examples, and the experimental data. Decisions were reached by consensus so as to avoid individual biases as far as possible.[8]

To allow for a direct comparison of the gold standard classification and the participants' judgements, we also created a consensus classification for the experimental data. Table V illustrates the participants' classifications for three example adjectives: For each adjective, the proportions of assignments to each of the semantic classes are provided. In the consensus classification, the adjective was assigned to the semantic class with the largest proportion of votes, given in the last column of the table.

As shown in Table V, 100% of the participants assigned *cranià* to the object class. For *conservador*, the judgements on class assignment varied, but half of the votes are nevertheless concentrated in the basic class, and a

Table V. Examples of the participants' classifications.

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

further third in the basic-event class. For *capaç* the judgements are spread across all classes, with only a slight majority (39%) for the basic-object class.

The agreement scores between the experts and the participants are shown in Table VI. They are still far from the 0.8 threshold, but they are much higher than the mean agreement between participants, with $\kappa = 0.55$ (in comparison to 0.27), and $w\kappa = 0.65$ (in comparison to 0.38). The reasons are presumably (a) that the participants' classification was obtained through a voting procedure (which ignored low-frequency classifications caused by individual biases), and (b) that the experts took the participants' classifications into account when building the gold standard classification, so the classification was influenced by the experiment data.

Table VI. Agreement experts/experiment.

p_o	κ	wp_o	$w\kappa$	op_o	$o\kappa$
0.68	0.55	0.79	0.65	0.85	0.72

The sources of disagreement between the two classifications can be traced in the contingency table in Table VII, which aligns the experts' and the participants' class assignments. The largest numbers are bold-faced.

Table VII. Contingency table: experts vs. participants.

		Participants						
		B	BE	BO	E	EO	O	*Total*
Experts	B	**79**	0	3	5	0	**20**	107
	BE	3	0	0	4	0	0	7
	BO	1	0	4	0	1	**17**	23
	E	2	1	1	**28**	1	4	37
	EO	0	0	0	2	2	2	6
	O	0	0	0	0	0	**30**	30
	Total	85	1	8	39	4	73	210

For all three monosemous classes (B, E, O), we find large numbers in the diagonal of the matrix, which indicates that there is a consensus on what the classes mean. However, large numbers are also found in two off-diagonal cells: for adjectives which experts tagged as basic and participants tagged as object (i.e., the B-O cell), and for adjectives which experts tagged as polysemous between basic and object and participants tagged as object only (BO-O). Furthermore, we identified two general phenomena: (i) many cases of disagreement appear between polysemous class assignments by the experts and monosemous class assignment by the participants (the underlined cases in the table); (ii) many cases of disagreement involved the event class. In what follows, we provide an interpretation of the various types of disagreement.

B-O disagreement. This type of disagreement involves adjectives such as *calb* ('bald'), *intel·ligent* ('intelligent'), *recíproc* ('reciprocal'), and *sant* ('holy'), where a deadjectival noun corresponding to the attribute denoted by the adjective exists: *calbície* ('baldness'), *intel·ligència* ('intelligence'), *reciprocitat* ('reciprocity'), and *santedat* ('holiness'), respectively. These nouns denote attributes, and the "related to" pattern cannot be properly applied to the adjectives to describe their meaning. The adjective *calb*, for instance, is related to the meaning of *calbície*, but it does not **mean** "related to baldness", which is what the use of the object pattern was meant to imply. This kind of disagreement thus indicates a problem with the definition of the pattern. However, the problem does not transfer to all basic adjectives. For instance, while 18 out of 58 participants provided the deadjectival nouns *amplada* for *ample* ('wide'), *amplària* and *amplitud* ('wideness'), thus assigning the adjective to the object class, the antonym *estret* ('narrow') motivated an overwhelming majority of participants (49 out of 58) to use the basic pattern. The overall behaviour of the participants suggests that attribute-denoting nouns are particularly salient for the above-mentioned adjectives, and that a suitable synonym or antonym (indicative of the basic class) is not as salient as the derived noun. The filtering procedure in Section 3.4 did not discard cases like *calb-calbície* because it is subjective to decide whether a noun refers to an attribute or to an object. It is clear, however, that in many cases the usage of this pattern did not correspond to its intended usage, and the experiment design should have avoided this confusion.

BO-O disagreement. This type of disagreement involves two cases of adjectival polysemy. First, adjectives such as *anarquista* ('anarchist(ic)') or *comunista* ('communist') are vague between an attribute reading (mostly when referring to humans) and a relation to an object (the abstract object corresponding to the underlying ideology). The experts therefore considered the basic-object to best represent this ambiguity. Most participants, though, only identified the relationship to the ideology. Second, adjectives falling in this type of disagreement correspond to true polysemy, and are cases of object-related adjectives that have also acquired a basic reading, as

explained in Section 2, example (1). Because such adjectives, e.g., *amorós* ('affectionate|of love'), *familiar* ('familiar|of family'), and *humà* ('human'), are denominal, participants tended to provide only the object reading and gloss over the basic reading, thus failing to identify polysemous adjectives. The design of the task should be improved to elicit polysemy.

Disagreement caused by polysemy. These considerations lead us to a more general phenomenon. We had wanted the participants to provide multiple class assignments in cases of polysemy. However, in general, the participants provided multiple responses in difficult or vague cases instead. In the cases in which participants consistently provided multiple classifications (which were very few), this did not indicate polysemy. An example of this is the adjective *capaç* (see Table V): The class assignments are spread over all classes, with the strongest class (BO) accounting for only 39% of the assignments. However, the most frequent responses (*incapaç* 'unable' for the basic pattern and *capacitat* 'ability' for the object pattern) do not indicate different senses. In fact, although some participants provided several classes per adjective in many cases (depending on personal taste or understanding of the task), as a collective they almost exclusively assigned monosemous classes, which indicates wide disagreement in the use of polysemous assignments. This behaviour is illustrated in Table VII above, in which the cases where experts provided a polysemous class and participants provided a monosemous class are underlined. These cases constitute the third main source of disagreement. Out of the 7 cases tagged as basic-event by experts, the participants assigned three to basic and four to event. Similarly, of the 23 BO expert cases, one was disambiguated as basic, and 17 were assigned to object only. Also, of the 6 adjectives classified as EO by the experts, two were assigned to event, two to object, and only 2 to EO by the participants.

Disagreement caused by event class. Out of the 67 cases where experts and participants disagree with respect to the semantic class of the adjectives, 28 (42%) involve the event class (classes BE, E, and EO). Of the remaining 41 cases, 37 correspond to the B-O and BO-O cases explained above. We have argued that B-O and BO-O disagreements are due to experimental design problems (which caused confusion about the object pattern) and to the inconsistent use of multiple responses to encode polysemy judgements. However, the classes basic and object seem to be well defined apart from this misunderstanding (see B-B and O-O cells in Table VII). In contrast, disagreements involving the event class are spread all over Table VII. This corresponds to random disagreement that indicates confusion with respect to the definition of the event class, as will be shown in the next section.

5.2. USING ENTROPY TO MEASURE THE DIFFICULTY OF AN ADJECTIVE

The previous section analysed the divergences between experts and participants, with the goal of shedding light on the sources of confusion that may explain the high disagreement between participants. This section pursues the same goal with different means: we analyse intra-item agreement, which allows us to identify groups of adjectives with a higher or lower degree of agreement, thus offering new kinds of analyses. The starting point for the analyses to follow is Table V in Section 5.1. Intuitively, if all the judgements are concentrated in one class (as in the case of *cranià*), there is strong consensus regarding the judgements; if they are evenly spread (as with *capaç*), there is no consensus.

We formalise this intuition using the Information Theory measure of entropy introduced by Shannon (1948). Entropy measures the average uncertainty in a random variable. If X is a discrete random variable with probability distribution $p(x) = Pr(X = x)$, x being each of the possible outcomes of the variable, its entropy is computed as in Equation 8.[9] If the outcome of the variable is totally predictable, the uncertainty (and thus the entropy) is 0; as the unpredictability increases, entropy also increases, with an upper bound determined by the number of possible outcomes of the random variable. In our case, the random variable is the class of the adjective, and predictability amounts to coincidence among subjects.

$$H(X) = -\sum p(x)\ \log_2 p(x) \tag{8}$$

Table VIII repeats the class proportions from Table V, and also shows the respective entropy values. The entropy values illustrate that the measure corresponds to our intuitions: for *cranià*, with total coincidence, entropy is 0; for *conservador*, with half of the probability mass in one class (B) and one third in another class (BE), entropy increases to 1.17. And finally, for *capaç*, with uneven proportions, it increases to 1.52. Summarising, a higher entropy value indicates a greater difficulty or confusion with respect to a given adjective. The upper bound for entropy in our data is 2.58, which applies to the case when all classes have an equal probability, $1/6^{10}$. However, the largest entropy value attested was 1.74, for the adjective *orientat* ('oriented'). In the following discussion, the entropy values are used to assess some of the sources of disagreement that were discussed in Section 5.1.

Polysemous adjectives: Adjectives classified as polysemous by the experts should correspond to less compact judgements than monosemous adjectives, and therefore have a higher entropy, because, as shown in Section 5.1, some participants choose just one of the two relevant monosemous classes, and some choose the polysemous class. Since each of the possible monosemous classes and also the polysemous class are considered to be different, separate classes, this distribution corresponds to higher entropy values for polysemous adjectives than for monosemous adjectives.

Table VIII. Entropy values from participants' classification.

Lemma	Trans.	B	BE	BO	E	EO	O	Entropy
cranià	cranial	0	0	0	0	0	1	0
conservador	conservative	0.5	0.33	0.02	0.11	0.04	0	1.17
capaç	able	0.06	0.11	0.39	0.17	0.03	0.25	1.52

Disagreement between experts and participants: More generally, cases where participants and experts disagree are expected to be more controversial than cases where they agree, due to the combination of two factors: (i) adjectives that are difficult or do not fit into the classification should exhibit a more uneven distribution of judgements across classes; (ii) a tendency can be expected towards more arbitrary decisions for these cases, which is likely to cause mismatches between participants and experts. Therefore, adjectives on which experts and participants disagree are expected to exhibit higher entropy values. These are mainly the BO-O and B-O cases discussed in the previous section.

The boxplots in Figure 1 show that our above predictions concerning polysemy are met.[11] Polysemous adjectives have higher entropy values (mean = 1.2, standard deviation = 0.29) than monosemous adjectives (M = 1.05, SD = 0.38). The difference is significant ($t(62.3)$ = -2.6, p = 0.01, two-tailed).[12] Adjectives prone to disagreement also exhibit higher entropy values (M = 1.25, SD = 0.30) than the rest (M = 0.99, SD = 0.38). The difference is again significant ($t(160.2)$ = -5.28, p $< 10^{-}6$, two-tailed). Note that the differences in entropy values are higher for disagreements between experts and participants than for polysemous cases, which indicates that disagreement predicts difficulty to a larger extent than polysemy.

Semantic class: We argued in Section 5.1 that the various cases of disagreement concerning the event class indicate a confusion with respect to the class definition, while the disagreements concerning the B-O, BO-O distinctions where mainly due to problems in the experimental design. Again, entropy provides us with a means to test this explanation: Adjectives classified as event adjectives by the experts should have higher entropy values, since the participants are more unsure about the class assignment, resulting in an uneven distribution of judgements across classes. The left-hand graph in Figure 2 supports this explanation and shows that event-related adjectives (classes BE, E, EO) are indeed more controversial than the rest. In contrast, object-related adjectives (class O) are the least problematic cases, which supports the analysis in Section 5.1. One-way ANOVA confirms that mean entropy values differ depending on the class ($F(5, 29.3) = 23.1, p < 10^{-}8$).[13]

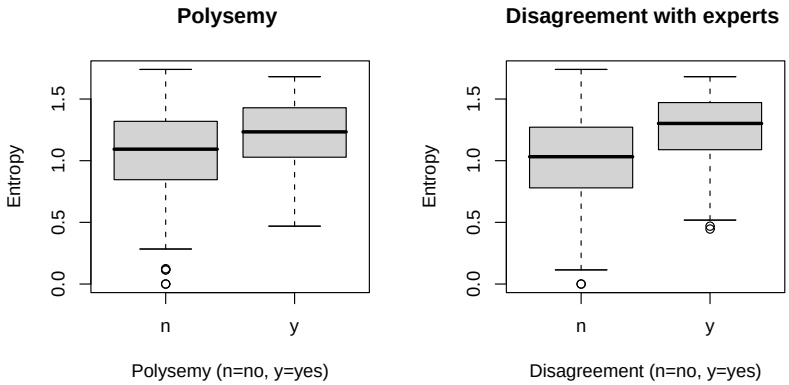

Figure 1. Explaining differences in entropy values I.

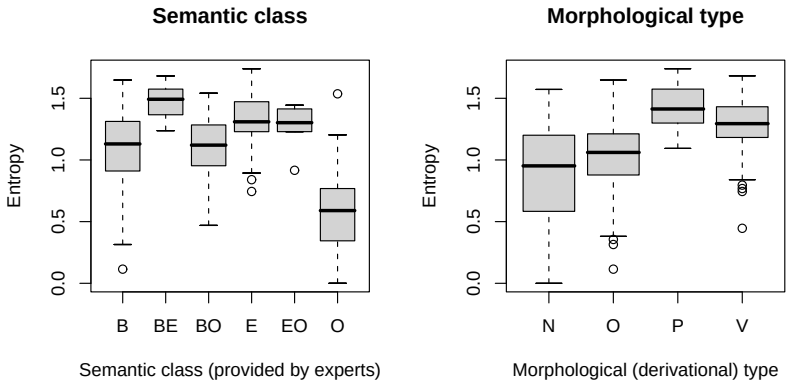

Figure 2. Explaining differences in entropy values II. Legend: N-denominal, O-non-derived, P-participial, V-deverbal.

Our hypothesis is that event adjectives are problematic because they are less homogeneous from a semantic point of view than the other two classes, due to two main factors. First, the semantic effects of morphological variation for event adjectives are more diverse than for object adjectives: Even though we found only 8 different suffixes for deverbal adjectives but 22 suffixes for denominal adjectives in our manually annotated database, object adjectives show a much more compact semantics than event adjectives, as shown by the fact that we defined 3 patterns to account for the semantics of event adjectives, but only one for object (and basic) adjectives. Second, the semantics contributed by the source verb is highly variable, mainly due to the *Aktionsart* or aspectual class of the verb (Vendler, 1957). Stative verbs produce more "basic-like" event adjectives. For instance, *abundant* was clas-

sified as event by the experts due to its relationship with the verb *abundar*. The participants, on the other hand, classified it as basic because it has an antonym *escàs* ('sparse'), corresponding to the fact that it seems to denote an attribute, like a basic adjective. Adjectives derived from process-denoting verbs (e.g., *protector*) have a more distinct semantics.

Because the semantic class and the derivational type of an adjective are related, we expect the differences in difficulty between event-related adjectives and all other adjectives to map to the morphological level. The right-hand graph in Figure 2 shows that participial and deverbal adjectives (closely corresponding to event adjectives, as explained in Section 2) have higher entropy values than the rest, and are indeed more controversial. The results of an ANOVA test again confirm this analysis ($F(3, 68.4) = 27.1, p < 10^{-}10$).

6. Conclusion

This article has described a large-scale experiment that collected and analysed human judgements, with the main goal of semantically classifying a set of Catalan adjectives. The collection of the judgements was carried out as a Web experiment with 322 non-expert subjects, who were asked to define adjectives through the use of pre-defined definitional patterns, each corresponding to one semantic class. The elicitation method thus used paraphrase relationships, which is a methodological innovation in collecting data for Computational Linguistics purposes.

The two main parts of this article provided two kinds of analyses on the experimental data. In the first analysis, we investigated the inter-annotator agreement concerning our classification task. We have proposed three methodological innovations, namely (i) a robust method to estimate confidence intervals, (ii) a principled weighting scheme to account for polysemy judgements, and (iii) three different definitions of agreement (full, weighted, and overlapping; equivalent to three different weighting schemes) to estimate the agreement scores under different considerations of partial agreement. With this methodology, however, the inter-annotator agreement on the classification task with its current experimental design has proved too low to establish a reliably labelled dataset: The most accurate estimate of the agreement score for our task, we have argued, is kappa $0.31 - 0.45$ (weighted agreement).

The second kind of analysis focused on the sources of disagreement. We have used two methods for this analysis: (i) a comparison of the participants' data with a classification performed by experts; and (ii) an analysis of intra-item agreement, in which entropy was used to identify the properties of difficult adjectives. This analysis has shown that the low level of agreement for our experiment has to do with the design of the experiment as well as with difficulties in the classification. As for the former aspect, the main problem is that, although the experiment was addressed to non-expert subjects, it asked

for metalinguistic judgements (definitions). The task should be redefined so that subjects provide intuitive judgements; how to best define such a task remains an open question. As for the difficulties in the classification, the analysis suggests that the event class is the least clearly defined of the three classes, a result that was further supported by Machine Learning experiments subsequently performed on the data (Boleda et al., 2007). Thus, the analysis also provides insight into a theoretical question, namely, the definition and characterisation of a semantic classification for adjectives.

One of the most difficult aspects of our task is the elicitation of polysemy judgements. This remains an unresolved challenge, a result consistent with the difficulties encountered in related research (Véronis, 1998; Fellbaum et al., 1998). We expected subjects to assign adjectives to several classes in the case of polysemy. However, the analysis of the agreement data has shown that they instead use several classes for either vague or difficult cases. Future experimental designs should improve this aspect.

To sum up, we believe that experiments and analyses of the sort explained in this article are a very useful source of insight into the design of experiments with human subjects for the elicitation of linguistic judgements, and should eventually lead to more robust resources and methods in Computational Linguistics. Furthermore, they can also provide feedback on empirical and theoretical linguistic questions in ways complementary to introspective methods and corpus analysis.

Acknowledgements

The authors wish to thank Ron Artstein, Alexander Fraser and two anonymous reviewers for their comments on a previous version of this article. Also thanks to Stefan Evert for discussion and many ideas regarding the assessment of inter-annotator agreement, and to Alissa Melinger for methodological advice on the design of the experiment. Special thanks are due to Roser Sanromà for providing us with the adjective database and for participating in the expert committee, and to all participants in the Web experiment. The research presented in this article has been partially funded by Fundación Caja Madrid, Universitat Pompeu Fabra, and Fundació Barcelona Media.

Notes

[1]For a review of research on adjectives from a formal semantics point of view, see Hamann (1991). For more details and justification of the adopted classification, see Boleda (2007).

[2]In order to adhere to ethical standards, we asked for permission to advertise the experiment from the relevant authorities. The experiment was also included in *Lan-*

guage Experiments (`http://www.language-experiments.org/`), a portal for psychological experiments on language, and an advertisement was placed in the first author's homepage.

[3]Fleiss (1981), among others, notes that the standard kappa is a particular case of the weighted kappa, where $w_{ij} = 0$ for all $i \neq j$.

[4]In fact, the decisions are not completely independent because the number of classifications was restricted to a maximum of two.

[5]An analogous weighting scheme was proposed by Passonneau (2004) for disjoint (0), intersecting (1/3), proper subsets (2/3), and identical (1) sets of coreference chains.

[6]Polysemous "classes" are not real classes, but a convenient way to represent the cases where an adjective belongs to more than one class.

[7]The total number of pairs was 158, corresponding to 316 subjects. For 6 out of the 7 test sets, an odd number of subjects was obtained, and one subject was randomly discarded.

[8]Note, however, that the expert gold standard thus built is not necessarily more reliable than the data from the Web experiment. For reliability, reproducibility is a necessary condition. The methodology used for the expert gold standard does not allow assessment of reproducibility, as decisions were not reached independently but by consensus. However, it does provide a good indication of the kind of classification that experts in the field (as opposed to non-expert native speakers) would build for the given set of adjectives.

[9]We use base 2 because entropy is usually measured in bits.

[10]$p(x) = 1/6$; $H(class) = -6(1/6)\log_2(1/6) = -\log_2(1/6) = 2.58$.

[11]The boxplots represent the value distribution of a continuous variable. The rectangles have three horizontal lines, representing the first quartile, the median, and the third quartile. 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 outliers and represented as points (Verzani, 2005).

[12]Equality of variance is not assumed.

[13]Homogeneity of variance is not assumed for the ANOVAs performed in this article.

References

Altaye, M., A. Donner, and M. Eliasziw: 2001, 'A general goodness-of-fit approach for inference procedures concerning the kappa statistic'. *Statistics in Medicine* **20**(16), 2479–2488.

Artstein, R. and M. Poesio: 2005, 'Bias decreases in proportion to the number of annotators'. In: G. Jaeger, P. Monachesi, G. Penn, J. Rogers, and S. Wintner (eds.): *Proceedings of FG-MoL 2005*. Edinburgh, pp. 141–150.

Artstein, R. and M. Poesio: to appear, 'Inter-coder agreement for computational linguistics'. Accepted for publication in *Computational Linguistics*.

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

Boleda, G.: 2007, 'Automatic acquisition of semantic classes for adjectives'. Ph.D. thesis, Pompeu Fabra University.

Boleda, G., S. Schulte im Walde, and T. Badia: 2007, 'Modelling adjective polysemy as multi-label classification'. In: *Proceedings of the Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning*.

Bosque, I. and C. Picallo: 1996, 'Postnominal adjectives in Spanish DPs'. *Journal of Linguistics* **32**, 349–386.

Carletta, J.: 1996, 'Assessing agreement on classification tasks: The kappa statistic'. *Computational Linguistics* **22**(2), 249–254.

Chierchia, G. and S. McConnell-Ginet: 2000, *Meaning and Grammar: An Introduction to Semantics*. Cambridge, MA: MIT Press, 2nd edition.

Cohen, J.: 1960, 'A coefficient of agreement for nominal scales'. *Educational and Psychological Measurement* **20**, 37–46.

Cohen, J.: 1968, 'Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit'. *Psychological Bulletin* **70**, 213–220.

Corley, M. and C. Scheepers: 2002, 'Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study'. *Psychonomic Bulletin and Review* **9**(1), 126–131.

Di Eugenio, B. and M. Glass: 2004, 'The kappa statistic: A second look'. *Computational Linguistics* **30**(1), 95–101.

Fellbaum, C., J. Grabowski, and S. Landes: 1998, 'Performance and confidence in a semantic annotation task'. In: C. Fellbaum (ed.): *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press, Chapt. 9, pp. 217–237.

Fleiss, J. L.: 1981, *Statistical Methods for Rates and Proportions*, Wiley series in probability and mathematical statistics. New York: John Wiley & Sons, second edition.

Hamann, C.: 1991, 'Adjectivsemantik/Adjectival semantics'. In: A. von Stechow and D. Wunderlich (eds.): *Semantik/Semantics. Ein internationales Handbuch der Zeitgenössischen Forschung. An International Handbook of Contemporary Research*. Berlin/New York: de Gruyter, pp. 657–673.

Hripcsak, G. and D. F. Heitjan: 2002, 'Measuring agreement in medical informatics reliability studies'. *Journal of Biomedical Informatics* **35**(2), 99–110.

Institut d'Estudis Catalans: 1997, *Diccionari de la llengua catalana*. Barcelona: Edicions 62.

Krippendorff, K.: 1980, *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Krippendorff, K.: 2004, *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage, second edition.

Landis, J. R. and G. C. Koch: 1977, 'The measurement of observer agreement for categorical data'. *Biometrics* **33**(1), 159–174.

Lapata, M., S. McDonald, and F. Keller: 1999, 'Determinants of adjective-noun plausibility'. In: *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, pp. 30–36.

Lui, K.-J., W. G. Cumberland, J. A. Mayer, and L. Eckhardt: 1999, 'Interval estimation for the intraclass correlation in dirichlet-multinomial data'. *Psychometrika* **64**(3), 355–369.

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

Melinger, A. and S. Schulte im Walde: 2005, 'Evaluating the relationships instantiated by semantic associates of verbs'. In: *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa, Italy.

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

Miller, K. J.: 1998, 'Modifiers in WordNet'. In: C. Fellbaum (ed.): *WordNet: an Electronic Lexical Database*. London: MIT, pp. 47–67.

Nirenburg, S. and V. Raskin: 2004, *Ontological Semantics*. Cambridge, MA: MIT Press.

Passonneau, R. J.: 2004, 'Computing reliability for coreference annotation'. In: *Proceedings of the Language Resources and Evaluation Conference*. Lisbon, Portugal.

Poesio, M. and R. Artstein: 2005, 'The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account'. In: *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, pp. 76–83.

Pustejovsky, J.: 1995, *The Generative Lexicon.* Cambridge, MA: MIT Press.
Rafel, J.: 1994, 'Un corpus general de referència de la llengua catalana'. *Caplletra* **17**, 219–250.
Raskin, V. and S. Nirenburg: 1998, 'An applied ontological semantic microtheory of adjective meaning for Natural Language Processing'. *Machine Translation* **13**(2-3), 135–227.
Reips, U. D.: 2002, 'Standards for internet-based experimenting'. *Experimental Psychology* **49**(4), 243–256.
Sanromà, R.: 2003, 'Aspectes morfològics i sintàctics dels adjectius en català'. Master's thesis, Universitat Pompeu Fabra.
Shannon, C. E.: 1948, 'A mathematical theory of communication'. *Bell System Technical Journal* **27**, 379–432.
Vendler, Z.: 1957, 'Verbs and times'. *The Philosophical Review* **66**, 143–60.
Verzani, J.: 2005, *Using R for Introductory Statistics.* Boca Raton: Chapman & Hall/CRC.
Véronis, J.: 1998, 'A study of polysemy judgements and inter-annotator agreement'. In: *Programme and advanced papers of the Senseval workshop.* Herstmonceux Castle, England, pp. 2–4.
Zipf, G. K.: 1949, *Human Behaviour and the Principle of Least-Effort.* Cambridge: Addison-Wesley.