Eliciting Subjectivity and Polarity Judgements on Word Senses

Fangzhong Su
School of Computing
University of Leeds
`fzsu@comp.leeds.ac.uk`

Katja Markert
School of Computing
University of Leeds
`markert@comp.leeds.ac.uk`

Abstract

There has been extensive work on eliciting human judgements on the sentiment of words and the resulting annotated word lists have frequently been used for opinion mining applications in Natural Language Processing (NLP). However, this word-based approach does not take different senses of a word into account, which might differ in whether and what kind of sentiment they evoke. In this paper, we therefore introduce a human annotation scheme for judging both the subjectivity and polarity of *word senses*. We show that the scheme is overall reliable, making this a well-defined task for automatic processing. We also discuss three issues that surfaced during annotation: the role of annotation bias, hierarchical annotation (or underspecification) and bias in the sense inventory used.

1 Introduction

Work in psychology, linguistics and computational linguistics has explored the affective connotations of words via eliciting human judgements (see Section 2 for an in-depth review). Two important parameters in determining affective meaning that have emerged are subjectivity and polarity. *Subjectivity identification* focuses on determining whether a language unit (such as a word, sentence or document) is *subjective*, i.e. whether it expresses a *private state, opinion or attitude*, or is factual. *Polarity identification* focuses on whether a language unit has a positive or negative connotation.

Word lists that result from such studies would, for example tag *good* or *positive* as a positive word, *bad* as negative and *table* as neither. Such word lists have frequently been used in natural language processing applications, such as the automatic identification of a review as favourable or unfavourable (Das and Chen, 2001). However, the word-based annotation conducted so far is at least partially unreliable. Thus Andreevskaia and Bergler (2006) find only a 78.7% agreement on subjectivity/polarity tags between two widely used word lists. One problem they identify is that word-based annotation does not take different senses of a word into account. Thus, many words are *subjectivity-ambiguous* or *polarity-ambiguous*, i.e. have both subjective and objective or both positive and negative senses, such as the words *positive* and *catch* with corresponding example senses given below.[1]

(1) positive, electropositive—having a positive electric charge;"protons are positive" (*objective*)

(2) plus, positive—involving advantage or good; "a plus (or positive) factor" (*subjective*)

(3) catch—a hidden drawback; "it sounds good but what's the catch?" (*negative*)

(4) catch, match—a person regarded as a good matrimonial prospect (*positive*)

Inspired by Andreeivskaia and Bergler (2006) and Wiebe and Mihalcea (2006), we therefore explore the subjectivity and polarity annotation of *word senses* instead of *words*. We hypothesize that annotation at the sense level might eliminate one possible source of disagreement for subjectivity/polarity annotation and will therefore hopefully lead to higher agreement than at the word level.

---

[1]All examples in this paper are from WordNet 2.0.

An additional advantage for practical purposes is that subjectivity labels for senses add an additional layer of annotation to electronic lexica and can therefore increase their usability. As an example, Wiebe and Mihalcea (2006) prove that subjectivity information for WordNet senses can improve word sense disambiguation tasks for subjectivity-ambiguous words (such as *positive*). In addition, Andreevskaia and Bergler (2006) show that the performance of automatic annotation of subjectivity at the *word* level can be hurt by the presence of subjectivity-ambiguous words in the training sets they use. A potential disadvantage for annotation at the sense level is that it is dependent on a lexical resource for sense distinctions and that an annotation scheme might have to take idiosyncracies of specific resources into account or, ideally, abstract away from them.

In this paper, we investigate the reliability of manual subjectivity labeling of word senses. Specifically, we mark up subjectivity/attitude (subjective, objective, and both) of word senses as well as polarity/connotation (positive, negative and no polarity). To the best of our knowledge, this is the first annotation scheme for judging both subjectivity and polarity of word senses. We test its reliability on the WordNet sense inventory. Overall, the experimental results show high agreement, confirming our hypothesis that agreement at sense level might be higher than at the word level. The annotated sense inventory will be made publically available to other researchers at http://www.comp.leeds.ac.uk/markert/data.

The remainder of this paper is organized as follows. Section 2 discusses previous related work. Section 3 describes our human annotation scheme for word sense subjectivity and polarity in detail. Section 4 presents the experimental results and evaluation. We also discuss the problems of bias in the annotation scheme, the impact of hierarchical organization or underspecification on agreement as well as problems with bias in WordNet sense descriptions. Section 5 compares our annotation to the annotation of a different scheme, followed by conclusions and future work in Section 6.

2   Related Work

Osgood et al. (1957) proposed semantic differential to measure the connotative meaning of concepts. They conducted a factor analysis of large collections of semantic differential scales and

pointed out three referring attitudes that people use to evaluate words and phrases—evaluation (good-bad), potency (strong-weak), and activity (active-passive). Also, they showed that these three dimensions of affective meaning are cross-cultural universals from a study on dozens of cultures (Osgood et al., 1975). This work has spawned a considerable amount of linguistic and psychological work in affect analysis on the *word* level. In psychology both the **Affective Norms for English Words (ANEW)** project as well as the **Magellan** project focus on collecting human judgements on affective meanings of words, roughly following Osgood's scheme. In the ANEW project they collected numerical ratings of pleasure (equivalent to our term polarity), arousal, and dominance for 1000 English terms (Bradley and Lang, 2006) and in Magellan they collected cross-cultural affective meanings (including polarity) in a wide variety of countries such as the USA, China, Japan, and Germany (Heise, 2001). Both projects concentrate on collecting a large number of ratings on a large variety of words: there is no principled evaluation of agreement.

The more linguistically oriented projects of the **General Inquirer (GI)** lexicon[2] and the **Appraisal** framework [3] also provide word lists annotated for affective meanings but judgements seem to be currently provided by one researcher only. Especially the General Enquirer which contains 11788 words marked for polarity (1915 positive, 2291 negative and 7582 no-polarity words) seems to use a relatively ad hoc definition of polarity. Thus, for example *amelioration* is marked as no-polarity whereas *improvement* is marked as positive.

The projects mentioned above center on subjectivity analysis on words and therefore are not good at dealing with subjectivity or polarity-ambiguous words as explained in the Introduction. Work that like us concentrates on *word senses* includes approaches where the subjectivity labels are automatically assigned such as **WordNet-Affect** (Strapparava and Valitutti, 2004), which is a subset of WordNet senses with semi-automatically assigned affective labels (such as emotion, mood or behaviour). In a first step, they manually collect an affective word list and a list of synsets which contain at least one word in this word list. Fine-

---

[2]Available at http://www.wjh.harvard.edu/ inquirer/

[3]Available at http://www.grammatics.com/appraisal/

grained affect labels are assigned to these synsets by the resource developers. Then they automatically expand the lists by employing WordNet relations which they consider to reliably preserve the involved labels (such as similar-to, antonym, derived-from, pertains-to, and attribute). Our work differs from theirs in three respects. First, they focus on their semi-automatic procedure, whereas we are interested in *human judgements*. Second, they use a finer-grained set of affect labels. Third, they do not provide agreement results for their annotation. Similarly, **SentiWordNet**[4] is a resource with *automatically determined polarity* of word senses in WordNet (Esuli and Sebastiani, 2006), produced via bootstrapping from a small manually determined seed set. Each synset has three scores assigned, representing the positive, negative and neutral score respectively. No human annotation study is conducted.

There are only two human annotation studies on subjectivity of word senses as far as we are aware. Firstly, the **Micro-WNOp** corpus is a list of about 1000 WordNet synsets annotated by Cerini et al. (2007) for polarity. The raters manually assigned a triplet of numerical scores to each sense which represent the strength of positivity, negativity, and neutrality respectively. Their work differs from us in two main aspects. First, they focus on polarity instead of subjectivity annotation (see Section 3 for a discussion of the two concepts). Second, they do not use absolute categories but give a rating between 0 and 1 to each synset—thus a synset could have a non-zero rating on both negativity and positivity. They also do not report on agreement results. Secondly, **Wiebe and Mihalcea (2006)** mark up WordNet senses as subjective, objective or both with good agreement. However, we expand their annotation scheme with polarity annotation. In addition, we hope to annotate a larger set of word senses.

3 Human Judgements on Word Sense Subjectivity and Polarity

We follow Wiebe and Mihalcea (2006) in that we see subjective expressions as private states "that are not open to objective observation or verification" and in that annotators distinguish between subjective (*S*), objective (*O*) and both subjective/objective (*B*) senses.

Polarity refers to positive or negative connotations associated with a word or sense. In contrast to other researchers (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005), we do not see polarity as a category that is dependent on prior subjectivity assignment and therefore applicable to subjective senses only. Whereas there is a dependency in that most subjective senses have a relatively clear polarity, polarity can be attached to objective words/senses as well. For example, *tuberculosis* is not subjective — it does not describe a private state, is objectively verifiable and would not cause a sentence containing it to carry an opinion, but it does carry negative associations for the vast majority of people. We allow for the polarity categories positive (*P*), negative (*N*), varying (*V*) or no-polarity (*NoPol*).

Overall we combine these annotations into 7 categories—*S:N*, *S:P*, *S:V*, *B*, *O:N*, *O:P*, and *O:NoPol*, which are explained in detail in the subsequent sections. Figure 1 gives an overview of the hierarchies over all categories.

As can be seen in Figure 1, our annotation scheme allows for hierarchical annotation, i.e. it is possible to only annotate for subjectivity or polarity. This can be necessary to achieve higher agreement by merging categories or to concentrate in specific applications on only one aspect.

3.1 Subjectivity

3.1.1 Subjective Senses

Subjective senses include several categories, which can be expressed by nouns, verbs, adjectives or adverbs. Firstly, we include emotions. Secondly, we include judgements, assessments and evaluations of behaviour as well as aesthetic assessments of individuals, natural objects and artefacts. Thirdly, mental states such as doubts, beliefs and speculations are also covered by our definition. This grouping follows relatively closely the definition of attitudinal positioning in the Appraisal scheme (which has, however, only been used on words, not on word senses before).

These types of subjectivity can be expressed via direct references to an emotion or mental state (see Example 5 or 8 below) as well as by expressive subjective elements (Wiebe and Mihalcea, 2006). Expressive subjective elements contain judgemental references to objects or events. Thus, *pontificate* in Example 6 below is a reference to a speech event that always judges it negatively; *beautiful* as
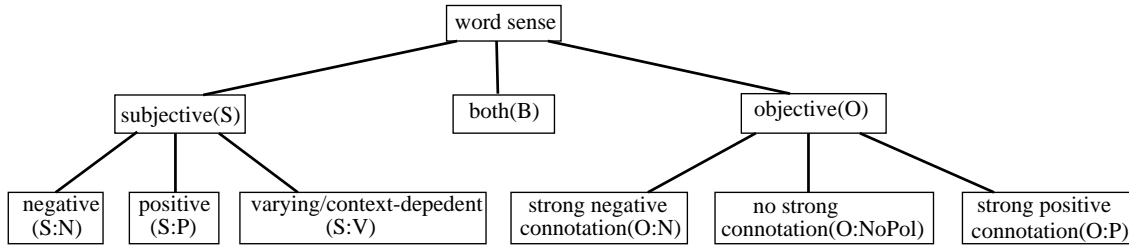
Figure 1: Overview of the hierarchies over all categories

in Example 7 below is a positive judgement.

(5) angry—feeling or showing anger; "angry at the weather; "angry customers; an angry silence" (*emotion*)

(6) pontificate—talk in a dogmatic and pompous manner; "The new professor always pontificates" (*assessment of behaviour*)

(7) beautiful—aesthetically pleasing (*aesthetic assessment*)

(8) doubt, uncertainty, incertitude, dubiety, doubtfulness, dubiousness—the state of being unsure of something (*mental state*)

3.1.2 Objective Senses

Objective senses refer to persons, objects, actions, events or states without an inherent emotion or judgement or an expression of a mental state. Examples are references to individuals via named entities (see Example 9) or non-judgemental references to artefacts, persons, animals, plants, states or events (see Example 10 and 11). If a sentence contains an opinion, it is not normally due to the presence of this word sense and the sense often expresses objectively verifiable states or events. Thus, Example 12 is objective as we can verify whether there is a war going on. In addition, a sentence containing this sense of *war* does not necessarily express an opinion.

(9) Einstein, Albert Einstein – physicist born in Germany who formulated the special theory of relativity and the general theory of relativity; Einstein also proposed that light consists of discrete quantized bundles of energy (later called photons) (1879-1955) (*named entity*)

(10) lawyer, attorney – a professional person authorized to practice law; conducts lawsuits or gives legal advice (*non-judgemental reference to person*)

(11) alarm clock, alarm – a clock that wakes sleeper at preset time (*non-judgemental reference to object*)

(12) war, warfare – the waging of armed conflict against an enemy; "thousands of people were killed in the war" (*non-judgemental reference to event*)

3.1.3 Both

In rare cases, a sense can be both subjective and objective (denoted by *B*). The following are the two most frequent cases. First, a WordNet sense might conflate a private state meaning and an objective meaning of a word in the gloss description. Thus, in Example 13 we have the objective literal use of the word *tarnish* mentioned such as *tarnish the silver*, which does not express a private state. However, it also includes a metaphorical use of *tarnish* as in *tarnish a reputation*, which implicitly expresses a negative attitude.

(13) tarnish, stain, maculate, sully, defile—make dirty or spotty, as by exposure to air; also used metaphorically; "The silver was tarnished by the long exposure to the air"; "Her reputation was sullied after the affair with a married man"

The second case includes the inclusion of near-synonyms (Edmonds, 1999) which differs on sentiment in the same synset list. Thus in Example 14, the term *alcoholic* is objective as it is not necessarily judgemental, whereas the other words in the synset such as *soaker* or *souse* are normally insults and therefore subjective.

(14) alcoholic, alky, dipsomaniac, boozer, lush, soaker, souse—a person who drinks alcohol to excess habitually

3.2 Polarity

3.2.1 Polarity of Subjective Senses

The polarity of a subjective sense can be positive (Category *S:P*), negative (*S:N*), or varying, dependent on context or individual preference (*S:V*). The definitions of these three categories are as follows.

- *S:P* is assigned to private states that express a positive attitude, emotion or judgement (see Example 7).

- *S:N* is assigned to private states that express a negative attitude, emotion or judgement (see Example 5, 6 and 8).

- *S:V* is used for senses where the polarity is varying by context or user. For example, it is

likely that you give an opinion about somebody if you call him *aloof*; however, only context can determine whether this is positive or negative (see Example 15).

(15) aloof, distant, upstage—remote in manner; "stood apart with aloof dignity"; "a distant smile"; "he was upstage with strangers" (*S:V*)

3.2.2 Polarity of Objective Senses

There are many senses that are objective but have strong negative or positive connotations. For example, *war* describes in many texts an objective state ("He fought in the last war") but still has strong negative connotations. In many (but not all) cases the negative or positive associations are mentioned in the WordNet gloss. Therefore, we can determine three polarity categories for objective senses:

- *O:NoPol* Objective with no strong, generally shared connotations (see Example 9, 10, 11 and 16).

- *O:P* Objective senses with strong positive connotations. These refer to senses that do not describe or express a mental state, emotion or judgement but whose presence in a text would give it a strong feel-good flavour (see Example 17).

- *O:N* Objective senses with strong negative connotations. These are senses that do not describe or express an emotion or judgement but whose presence in a text would give it a negative flavour (see Example 12). Another example is (18): you can verify objectively whether a liquor was diluted, but it is normally associated negatively.

(16) above—appearing earlier in the same text; "flaws in the above interpretation" (*O:NoPol*)

(17) remedy, curative, cure – a medicine or therapy that cures disease or relieve pain (*O:P*)

(18) adulterate, stretch, dilute, debase—corrupt, debase, or make impure by adding a foreign or inferior substance; often by replacing valuable ingredients with inferior ones; "adulterate liquor" (*O:N*)

We only allow positive and negative annotations for objective senses if we expect *strong* connotations that are *shared* among most people (in Western culture). Thus, for example *war*, *diseases* and *crimes* can relatively safely be predicted to have shared negative connotations. In contrast, a sense like the one of *alarm clock* in Example 11 might

have negative connotations for late risers but it would be annotated as *O:NoPol* in our scheme. We are interested in strong shared connotations as the presence of such "loaded" terms can partially indicate bias in a text. In addition, such objective senses are likely to give rise to figurative subjective senses (see Example 18).

4 Experiments and Evaluation

This section describes the experimental setup for our annotation experiments, presents reliability results and discusses the benefits of the use of a hierarchical annotation scheme as well as the problems of bias in the annotation scheme, annotator preferences and bias in the sense inventory.

4.1 Dataset and Annotation Procedure

The dataset used in our annotation scheme is the Micro-WNOp corpus[5], which contains all senses of 298 words in WordNet 2.0. We used it as it is representative of WordNet with respect to its part-of-speech distribution and includes synsets of relatively frequent words, including a wide variety of subjective senses. It contains 1105 synsets in total, divided into three groups *common* (110 synset), *group1* (496 synsets) and *group2* (499 synsets). We used *common* as the training set for the annotators and tested annotation reliability on *group1*.

Annotation was performed by two annotators. Both are fluent English speakers; one is a computational linguist whereas the other is not in linguistics. All annotation was carried out independently and without discussion during the annotation process. The annotators were furnished with guideline annotations with examples for each category. Annotators saw the full synset, including all synonyms, glosses and examples.

4.2 Agreement Study

Training. The two annotators first annotated the *common group* for training. Observed agreement on the training data is 83.6%, with a kappa (Cohen, 1960) of 0.76. Although this looks overall quite good, several categories are hard to identify, for example *B* and *S:V*, as can be seen in the confusion matrix below (Table 1) with Annotator 1 in columns and Annotator 2 in the rows.

Testing. Problem cases were discussed between the annotators and a larger study on *group 1* as test

---

Table 1: Confusion matrix for the training data

	B	S:N	S:P	S:V	O:NoPol	O:N	O:P	total
B	1	0	0	0	2	0	0	3
S:N	0	13	0	0	0	2	0	15
S:P	0	0	8	1	1	0	0	10
S:V	1	1	0	13	6	0	0	21
O:NoPol	1	0	0	0	50	0	0	51
O:N	0	0	0	0	2	4	0	6
O:P	0	0	1	0	0	0	3	4
total	3	14	9	14	61	6	3	110

data was carried out. Table 2 shows the confusion matrix for all 7 categories.

Table 2: Confusion matrix on the test set

	B	S:N	S:P	S:V	O:NoPol	O:N	O:P	total
B	7	2	0	2	0	0	0	11
S:N	0	41	1	0	0	0	0	42
S:P	0	0	65	4	0	0	2	71
S:V	0	0	7	17	3	0	0	27
O:NoPol	9	1	2	6	253	5	8	284
O:N	0	14	0	2	0	25	0	41
O:P	1	0	5	0	1	0	13	20
total	17	58	80	31	257	30	23	496

The observed agreement is 84.9% and the kappa is 0.77. This is good agreement for a relatively subjective task. However, there is no improvement over agreement in training although an additional clarification phase of the training material took place between training and testing.

We also computed single category kappa in order to estimate which categories proved the most difficult. Single category-kappa concentrates on one target category and conflates all other categories into one *non-target* category and measures agreement between the two resulting categories. The results showed that *S:N* (0.80), *S:P* (0.84) and *O:NoPol* (0.86) were highly reliable with less convincing results for *B* (0.49), *S:V* (0.56), *O:N* (0.68), and *O:P* (0.59). *B* is easily missed during annotation (see Example 19), *S:V* is easily confused with several other categories (Example 20), whereas *O:N* is easily confused with *O:NoPol* and *S:N* (Example 21); and *O:P* is easily confused with *O:NoPol* and *S:P* (Example 22).

(19) antic, joke, prank, trick, caper, put-on—a ludicrous or grotesque act done for fun and amusement (*B vs O:NoPol*)

(20) humble—marked by meekness or modesty; not arrogant or prideful; "a humble apology" (*S:V vs S:P*)

(21) hot—recently stolen or smuggled; "hot merchandise"; "a hot car" (*O:N vs O:NoPol*)

(22) profit, gain—the advantageous quality of being beneficial (*S:P vs O:P*)

Our annotation scheme also needs testing on an even larger data set as a few categories such as *B* and *O:P* occur relatively rarely.

4.3 The Effect of Hierarchical Annotation

As mentioned above, our annotation scheme allows us to consider the subjectivity or polarity distinction individually, leaving the full categorization underspecified.

Subjectivity Distinction Only. For subjectivity distinctions we collapse *S:V*, *S:P* and *S:N* into a single label *S* (subjective) and *O:NoPol*, *O:N* and *O:P* into a single label *O* (objective). *B* remains unchanged. The resulting confusion matrix on the test set is in Table 3.

Table 3: Confusion matrix for *Subjectivity*

	B	S	O	total
B	7	4	0	11
S	0	135	5	140
O	10	30	305	345
total	17	169	310	496

Observed agreement is 90.1% and kappa is 0.79. Single category kappa is 0.49 for *B*, 0.82 for *S* and 0.80 for *O*. As *B* is a very rare category (less than 5% of items), this is overall an acceptable level of distinction with excellent reliability for the two main categories.

Polarity Distinction Only. We collapse *O:N* and *S:N* into a single category *N* (negative) and *O:P* and *S:P* into *P* (positive), leaving the other categories intact. This results in 5 categories *B*, *S:V/V*, *NoPol*, *N* and *P*. The resulting confusion matrix is in Table 4.

Table 4: Confusion matrix for *Polarity*

	B	N	P	V	NoPol	total
B	7	2	0	2	0	11
N	0	80	1	2	0	83
P	1	0	85	4	1	91
V	0	0	7	17	3	27
NoPol	9	6	10	6	253	284
total	17	88	103	31	257	496

Observed agreement is 89.1% and kappa is 0.83. Single category kappa is as follows: *B* (0.49), *N* (0.92), *P* (0.85), *V* (0.56), and *NoPol* (0.86). This means all categories but *B* and *V* (together about 10% of items) are reliably identifiable.

Overall we show that both polarity and subjectivity identification of word senses can be reliably annotated and are well-defined tasks for automatic classification. Specifically the per-

centage agreement of about 90% for word sense polarity/subjectivity identification is substantially higher than the one of 78% reported in Andreeivskaia and Bergler (2006). Agreement for polarity-only is significantly higher than for the full annotation scheme, showing the value of hierarchical annotation. We believe hierarchical annotation is also appropriate for this task, as subjectivity and polarity are linked but still separate concepts. Thus, a researcher might want to mainly focus on explicitly expressed opinions as exemplified by subjectivity, whereas another can also focus on opinion bias in a text as expressed by loaded words of positive or negative polarity.

4.4 Bias in Annotation Performance, Sense Inventory and Annotation Guidelines

Why do annotators assign different labels to some senses? Three main aspects are responsible for non-spurious disagreement.

Firstly, individual perspective or bias played a role. For example, Annotator 2 was more inclined to give positive or negative polarity labels than Annotator 1 as can be seen in Table 4, where Annotator 2 assigned 103 positive and 88 negative labels,whereas Annotator 1 assigned only 91 positive and 83 negative labels.

Secondly, the WordNet sense inventory conflates near-synonyms which just differ in sentiment properties (see Section 3.1.3 and Example 14). Although the labels *B* and *S:V* were specifically created in the annotation scheme to address this problem, these cases still proved confusing to annotators and do not readily lead to consistent annotation.

Thirdly, WordNet sometimes includes a connotation bias either in its glosses or in its hierarchical organization. Here we use the word connotation bias for the inclusion of connotations that seem highly controversial. Thus, in Example 23, the WordNet gloss for *Iran* evokes negative connotations by mentioning allegations of terrorism.[6] In Example 24 *skinhead* is a hyponym of bully, giving strong negative connotations for *all* skinheads. Although the annotation scheme explicitly encourages annotators to disregard especially such controversial connotations as in Example 23 such examples can still confuse annotators and show that word sense annotation is to a certain degree depen-

dent on the sense inventory used.

(23) Iran, Islamic Republic of Iran, Persia—a theocratic islamic republic in the Middle East in western Asia; Iran was the core of the ancient empire that was known as Persia until 1935; rich in oil; involved in state-sponsored terrorism

(24) skinhead ⟵ bully, tough, hooligan, ruffian, roughneck, rowdy, yob, yobo, yobbo

Some of our *good* reliability performance might be due to one particular instance of bias in the annotation guidelines. We strongly advised annotators to only annotate positive or negative polarity for objective senses when strong, shared connotations are expected,[7] thereby "de-individualising" the task of polarity annotation. This introduces a bias towards the category *NoPol* for objective senses. We also did not allow varying polarity for objective senses, instructing annotators that such polarity would be unclear and should be annotated as *NoPol* as not being a strong shared connotation. It can of course be questioned whether the introduction of such a bias is good or not. It helps agreement but might reduce the usefulness of the annotation as individual connotations are not annotated for objective senses. However, to consider more individual connotations needs an annotation effort with a much larger number of annotators to arrive at a profile of polarity connotations over a larger population. We leave this for future work. Our current framework is comprehensive for subjectivity as well as polarity for subjective senses.

4.5 Gold Standard

After discussion between the two annotators, a gold standard annotation was agreed upon. Our data set consists of this agreed set as well as the remainder of the Micro-WNOp corpus (*group2*) annotated by one of the annotators alone after agreement was established.

How many words are subjectivity-ambiguous or polarity-ambiguous, i.e. how much information do we gain by annotating senses over annotating words? As the number of senses increases with word frequency, we expect rare words to be less likely to be subjectivity-ambiguous than frequent words. The Micro-WNOp corpus contains relatively frequent words so we will get an overestimation of subjectivity-ambiguous word types from this corpus, though not necessarily of word tokens. Of all 298 words, 97 (32.5%) are subjectivity-ambiguous, a substantial number. Fewer words are

---

[6]Note that this was part of WordNet 2.0 and has been removed in WordNet 2.1.

[7]See Section 3.2.2 for justification.

polarity-ambiguous: only 10 words have at least one positive and one negatively annotated sense with a further 44 words having at least one subjective sense with varying polarity (*S:V*). This suggests that subjective and objective uses of the same word are more frequent than reverses in emotional orientation.

5 Comparison to Original Polarity Annotation (Cerini et al.)

We can compare the reliability of our own annotation scheme with the original (polarity) annotation in the Micro-WNOp corpus. Cerini et al. (2007) do not present agreement figures but as their corpus is publically available we can easily compute reliability. Recall that each synset has a triplet of numerical scores between 0 and 1 each: positivity, negativity and neutrality, which is not explicitly annotated but derived as $1 - (positivity + negativity)$. Subjectivity in our sense (existence of a private state) is *not* annotated.

The ratings of three annotators are available for *Group 1* and of two annotators for *Group 2*. We measured the Pearson correlation coefficient between each annotator pair for both groups for both negativity and positivity scoring. As correlation can be high without necessarily high agreement on absolute values, we also computed a variant of kappa useful for numerical ratings, namely alpha (Artstein and Poesio, 2005), which gives weight to degrees of disagreement. Thus, a disagreement between two scores would be weighted as the absolute value of $score1 - score2$. The results are listed in Table 5.

Table 5: Reliability of original annotation on Micro-WNOp

dataset	raters	score type	correlation	alpha
Group 1	1 and 2	negative	83.7	64.9
Group 1	1 and 3	negative	86.4	71.8
Group 1	2 and 3	negative	82.5	56.9
Group 1	1 and 2	positive	80.5	60.9
Group 1	1 and 3	positive	87.8	74.9
Group 1	2 and 3	positive	78.2	57.5
Group 2	1 and 2	negative	95.9	90.7
Group 2	1 and 2	positive	92.2	84.9

Correlation between the annotators is high. However, Rater 2 (in Group1) still behaves differently from the other two raters, giving consistently higher or lower scores overall, leading to low alpha. Thus, we can conclude that Group 2 is much more reliably annotated than Group 1 and that es-pecially Rater 2 in Group 1 is an outlier in this (small) set of raters. This also shows that work with several annotators is valuable and should be conducted for our scheme as well.

6 Conclusion and Future Work

We elicit human judgements on the subjectivity and polarity of word senses. To the best of our knowledge, this is the first such annotation scheme for both categories. We detail the definitions for each category and measure the reliability of the annotation. The experimental results show that when using all 7 categories, only 3 categories (*S:N*, *S:P*, and *O:NoPol*) are reliable while the reliability of the other 4 categories is not high. We also show that this is improved by the virtue of hierarchical annotation and that the general tasks of subjectivity and polarity annotation on word senses are therefore well-defined. Moreover, we also discuss the effect of different kinds of bias on our approach.

In future we will refine the guidelines for the more difficult categories, including more detailed advice on how to deal with sense inventory bias. We will also perform larger-scale annotation exercises with more annotators as the latter is necessary to deal with more individualised polarity connotations. In addition, we will use the data to test learning methods for the automatic detection of subjectivity and polarity properties of word senses.

References

Andreevskaia, Alina and Sabine Bergler. 2006. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *Proceedings of EACL'06.*

Artstein, Ron and Massimo Poesio. 2005. $Kappa^3$=alpha(or beta). *Technical Report CSM-437, University of Essex.*

Bradley, Margaret and Peter Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings *Technical report C-1, the Center for Research in Psychophysiology, University of Florida. .*

Cerini, Sabrina, Valentina Compagnoni, Alice Demontis, Maicol Formentelli, and Caterina Gandini. 2007. Micro-WNOp: A Gold Standard for the Evaluation of Automatically Compiled Lexical Resources for Opinion Mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics.*

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, Vol.20, No.1*.

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. *Proceedings of APFA'01*.

Edmonds, Philip. 1999. Semantic Representations Of Near-Synonyms For Automatic Lexical Choice. *PhD thesis, University of Toronto*.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC'06*.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of ACL'97*.

Heise, David. 2001. Project Magellan: Collecting Cross-culture Affective Meanings via the Internet. Electronic Journal of Sociology.

Osgood, Charles, William May, and Murray Miron. 1975. Cross-cultural Universals of Affective Meaning. *University of Illinois Press.*

Osgood, Charles, George Suci, and Percy Tannenbaum. 1957. The Measurment of Meaning. *University of Illinois Press.*

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet-Affect: an Affective Extension of WordNet. *Proceedings of LREC'04*.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. *Proceedings of ACL'05*.

Wiebe, Janyce and Rada Micalcea. 2006. Word Sense and Subjectivity. *Proceedings of ACL'06*.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*.