Annotation Scheme and Gold Standard for Dutch Subjective Adjectives

Isa Maks, Piek Vossen

Faculteit der Letteren, Vrije Universiteit Amsterdam,

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

email: {e.maks, p.vossen}@let.vu.nl

Abstract

Many techniques are developed to derive automatically lexical resources for opinion mining. In this paper we present a gold standard for Dutch adjectives developed for the evaluation of these techniques. In the first part of the paper we introduce our annotation guidelines. They are based upon guidelines recently developed for English which annotate subjectivity and polarity at word sense level. In addition to subjectivity and polarity we propose a third annotation category: that of the attitude holder. The identity of the attitude holder is partly implied by the word itself and may provide useful information for opinion mining systems. In the second part of paper we present the criteria adopted for the selection of items which should be included in this gold standard. Our design is aimed at an equal representation of all dimensions of the lexicon , like frequency and polysemy, in order to create a gold standard which can be used not only for benchmarking purposes but also may help to improve in a systematic way, the methods which derive the word lists. Finally we present the results of the annotation task including annotator agreement rates and disagreement analysis.

1. Introduction

In recent years much attention has been paid to the automatic detection of opinions, sentiments, beliefs and emotions (subjectivity) in text. Most of the techniques use some kind of word list annotated with polarity and subjectivity features. Initially, simple word lists were compiled and automatically annotated for negative or positive polarity only (Kamps et al. (2004), Hatzivassiloglou (1997)). In these lists *friendly* would be tagged as positive, *sad* as negative and *chemical* as neutral. One of the main problems with these lists is that the annotation is at the word level, ignoring the possibility that a word may have both objective and subjective senses, or may have both positive and negative senses.

Such subjectivity-ambiguous and polarity-ambiguous words may cause major errors in the applications they are used in (Andreevskaia et al. 2006). To overcome this problem, annotation at word-sense level is introduced (Andreevskaia (2006), Wiebe (2006)) and annotation schemes sense level are developed for English. The most recent annotation scheme (Su and Markert (2008)) combines labels for subjectivity and polarity and applies them at the word-sense level. These schemata, however, lack information that may be annotated at word sense level as well, i.e. information related to the identity of the attitude holder.

In this paper we present a gold standard for Dutch subjective adjectives following new annotation guidelines which permit the annotation of subjectivy, polarity and attitude holder. The remainder of the paper is organised as follows. In the next section we discuss the annotation guidelines in detail. In section 3 we discuss the selection of the gold standard data. Section 4 presents the experimental results of the annotation task as well as inter-annotator agreement rates and disagreement analysis. In section 5 the final gold standard data are presented and analysed with respect to different lexicon layers.

2. Annotating Polarity, Subjectivity and Annotation Holder

The annotation schema we propose is a futher elaboration of previous schemata developed for word sense subjectivity tagging in English and in particular of the schema developed by Su and Markert (2008). Important aspects – as illustrated by examples (1a-d) - of these existing annotation schemata are (1) tagging at word sense level (2) distinction between objective and subjective words and (3) both objective and subjective words may have positive and negative polarity.

ex. (1a) (Subjective:Negative) angry—feeling or showing anger; “angry at the weather; “angry customers; an angry silence”

ex. (1b) (Subjective:Positive) beautiful—aesthetically pleasing

ex.(1c) (Objective:NoPolarity) alarm clock, alarm – a clock that wakes the sleeper at a preset time

ex.(1d) (Objective: negative) war, warfare – the waging of armed conflict against an enemy; ”thousands of people were killed in the war”

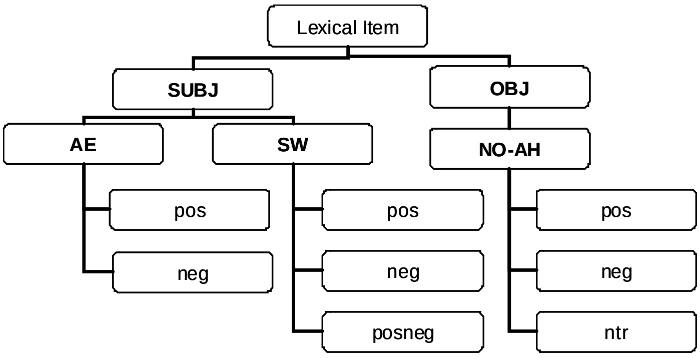
In our view, however, this schema lacks information about about whose attitude, opinion or point of view is expressed. The schema focusses on the degree of subjectivity of a word , i.e. whether it expresses an opinion or attitude , or is factual. And it focusses on polarity, i.e. whether a sense unit has a positive or negative connotation. In opinion mining and sentiment mining these two aspect are highy important but equally important is the identification of the attitude holder. Kim and Hovy (2006) define attiudes and opinions as consisting of three elements: a topic, a valence or polarity, and a holder. The detection of the topic is beyond the scope of the present study; the detection of the valency relies upon the use of word sense lists annotated for subjectivity and polarity; in our view, the detection of the holder of the attitude may rely as well on the word sense lists if they are annotated with attitude holder information. This can be illustrated by the following examples:

ex. (2) Bush is angry over Obama's leeking of private conversation ..... [attitude/judgment of Bush on Obama]
ex. (3) Bush is bad for the economy …. [attude/judgement of Speaker/Writer on Bush]

These two sentences both give opinions with a negative valency which is expressed by the negative polarity of *angry* and *bad*. The difference is whose opinion is expressed , i.e. who is the holder of the opinion or attitude : in the first it is Bush's opinion about Obama and in the second case it is the speaker or writer of the proposition whose opinion / attitude is expressed. The identification of the attitude holder is closely related to characteristics of the subjectivity clues *angry* and *bad* respectively which imply the possible identity of the attitude holder. In example (2) it is the person the adjective is attributed to, in this case represented by the logical subject of the sentence, whose attitude is expressed. In example (2) it is the implicit speaker or writer whose attitude is expressed. These different perspectives, and the possible inferences implied by them, are part of the semantics of the word itself. In combination with the syntactic structure of the sentence it can be used to identify the attitude holder.
Therefore, we propose to extend the annotation schema with an extra layer for attitude holder (cf. figure 1) which make it possible to assign the attitude to individuals. This layer leads to the further distinction of the category subjective in 2 subcategories : Speaker/Writer (SW) and Agent/Experiencer (AE). The SW-group includes words which imply an attitude holder who is the speaker/ writer of the text (cf. ex. 3) or of the embedded proposition (cf. ex 4). In example (4) it is McCain, the speaker/writer of the embedded opinion, who thinks Palin will make an excellent president. The AE-group implies attitude holders that are explicity mentioned in the text, as the logical subject of the utterance or as the noun that is modified by the adjective. The group includes two types: the Experiencer who inactively undergoes an emotion (cf. ex. 2 and 5) ; and the Agent who actively takes a stance or attitude(cf. ex. 6). In both cases the attitude is expressed by an adjective which is ascribed to this Agent or Experiencer .

ex. (4) McCain thinks Palin will make an excellent vice president ...
ex. (5) . . excited fans welcome Obama.. [attitude/judgement of fans]
ex. (6) McCain was critical of Iraq war from the beginning …. [attitude/judgement of McCain on Iraq war]

Summarizing we come to the following annotation schema :



SUBJ-subjective	pos-positive
OBJ-objective	neg-negative
AE-agent/experiencer	posneg-positive/negative
SW-speaker/writer	ntr-neutral

[*[AE-pos] blij (content, happy) ex.: W.B. is blij met de keuze voor .. W.B. is happy with the choice for .*
[AE-neg] McCain was al vanaf het begin kritisch over de oorlog te gen Irak *McCain was critical of the Iraq war from the beginning ….*
[SW-pos] mooi (*beautiful*) ex.: ze rijden daar in mooie auto's rond *They drive around in beautiful cars*
[SW-neg] slecht (*bad)* ex.: Bush is slecht voor de economie *Bush is bad for the economy*
[SW-posneg] voorzichtig (*cautious, conservative*) ex.: voorzichtige schatting *conservative estimate*
[OBJ-pos] waterbestendig (*water resistant)* ex.: waterdicht horloges *water resistant watches*
[OBJ-neg] kreupel (*crippled)* ex.: een kreupel paard *a crippled horse*
[OBJ-ntr] burgerlijk (*civil, civilian*) ex.: burgerlijk huwelijk *civil marriage*

Figure 1: annotation schema subjectivity, polarity and attitude holder

- SW refers to those word senses which imply the attitude, point of view or perspective of the speaker or writer.
- AE refers to those word senses where the attitude is atttributed to somebody explicitly mentioned in the text. This may be an Experiencer, who experiences an emotion or it may be a person , a cognitive Agent who is a more conscious stance-taker.
- OBJ refers to those word senses which are often neutral, and which do not imply a specific attitude holder (No-AH). They may evoke strong positive or negative associations which are probably shared by both speaker/writer and reader/listener.
- Pos refers to word senses which express a positive attitude towards somebody or something; in the case Obj-pos they may evoke a positive association.
- Neg refers to word senses which express a negative attitude towards somebody or something; in the case Obj-neg they may evoke a negative association.
- Posneg refers to those word senses which express an attitude or feeling that might be felt differently in different situations, not depending on linguistic context.
- Ntr refers to neutral word senses which do not express or evoke any positive or negative attitude. This category prototypically includes classifying adjectives like *dental (laboratory), social (services)*, etc.. It also includes words whose polarity very much depends on the context they appear in. They have ‘contextual’ instead of ‘prior’ polarity (cf. Wilson et al., 2005), like many frequent adjectives (e.g. *long, short, big, high*).

3. Design of the Gold Standard

An important aspect of the design of the gold standard concerns which type of word sense items should be included: lexical units or synsets. Recent annotation schemata for English are applied at Wordnet synsets (cf. Su et al. (2008), Wiebe et al. (2003), Cerini et al.(2007)) assuming that all synonyms of a synset share the characteristics with regard to subjectivity and polarity. As many synsets are internally ambiguous with regard to polysemy and subjectivity (cf. Maks and Vossen (2010) and section 6.2 below), our annotations will be performed at the lexical unit level. We make use of the Dutch Cornetto Database which combines two lexical resources with different semantic organisations: the Dutch Wordnet with its synset organisation and the Dutch Reference Lexicon with its form-meaning composites or lexical units. As nearly all automatic lexical acquisition techniques make use of Wordnet, it is important to note that within the Cornetto Database, each synonym in a synset is linked to the corresponding lexical unit of the Dutch Reference Lexicon.

Another issue is which items of this collection of lexical units should be included. We aim at a selection which is representative for the whole lexicon and relevant for subjectivity identification and annotation and set the following requirements:

(1) inclusion of relevant data. This requirement is not as trivial as it seems. A random selection of words would result in a data set useless as gold standard since it would be largely composed of neutral, non sentiment bearing words. As for Dutch no balanced subjectivity word lists exist, we compose our gold standard from scratch. We did a preliminary test and found that approximately 40% of the items were objective-neutral. Although the inclusion of objective or neutral senses is important, we think that a portion of 40% is too large and would lead to a gold standard non representative of all different characteristics related to subjective items. Therefore, we eliminated part of the neutral senses by randomly selecting one out of two, thus reducing their portion to approximately 20%.

(2) representativeness of the lexicon with regard to word frequency, polysemy and large synset membership. As the gold standard serves as a benchmark for evaluation and comparison of the automatically annotated sentiment word lists and helps to improve the methods which derive these word lists, these three lexicon dimenions are important:

- Frequency: most recently used goldstandards for English, like General Inquirer (Stone, 1966) and Micro-Wnop (Cerini, 2007) seem to be unbalanced with regard to frequency which may lead to a non representative high degree of subjectivity-ambiguous words (Su et al. , 2008). Moreover, as the number of infrequent or rare words is very high in subjective text (Wiebe et al., 2004), we think it is important to include high frequency as well as low frequency words in the gold standard.
- Polysemy is mentioned by by Andreevskaia et al. (2006) as one of the major origins of errors of their lexicon tagging sentiment system. Especially the identification of occurrences of sentiment-laden meanings of a given polysemous word from the occurrences of its neutral meanings, hampers seriously the performance of the system.
- Large synset membership: in our own experience (Maks and Vossen (2010)) sentiment-bearing words tend to cluster in large synsets which may be internally inconsistent with respect to subjectivity and polarity. To be able to test for this phenomenon as well, we included also some members of large synsets in the test set.

We aim at an equal distribution of test items across these three dimensions (frequency, polysemy and large synset membership) and three values (high, medium and low). For each dimension and each value we selected approximately 70 items. Thus, each item focusses on a single phenomenon which distinguishes it from all other test items and the performance of systems can be evaluated for isolated phenomena. Table (1) shows the results of this selection procedure: the 609 test items are distributed in such a way that each dimension-value combination includes at least 129 word senses.

	frequency	**Polysemy**	**synset-size**
high	179	129	202
mid	164	239	256
low	266	241	151
	609	609	609

Table 1: Distribution of gold standard items

4. Annotation Results

This section presents the results of the annotation task (section 4.1), discusses the problematic cases (section 4.2) and compares the results with other studies (section 4.3).

Annotation is performed by 2 annotators (A1 and A2) who prepared the guidelines, did a first annotation task for training and discussed the problems before the gold standard annotation task was carried out. During and after the annotation task there was no further interaction between the annotators. Polarity and attitude holder were annotated as combined categories and will be evaluated together and as separate categories. Separate evaluations allow for comparison with other studies which may have more limited annotations than we have.

4.1 Inter annotator rates

- Polarity

Agreement for all four polarity categories– where attitude holder categories AH, SW and OBJ are neglected - is 86.3% with a Cohen kappa (κ) of 0.80.

	neg	pos	ntr	posneg	Totals	Kappa
neg	216	3	6	6	231	0.90
pos	3	197	13	3	216	0.81
Ntr	6	17	101	4	128	0.76
Posneg	5	15	2	12	34	0.38
Totals	230	232	122	25	609	

Table 2: Confusion Matrix - polarity

Single category kappa computation (cf. table 2, column 7) – with one category of interest and all other categories combined into one non-relevant category - shows that all categories but ‘posneg’ are reliable identifiable. Table 2 shows that the low scores for ‘posneg’ are due to an easy confusion of ‘posneg’ with all other categories and in particular the category ‘positive’.

- Attitude Holder

Agreement for all attitude holder categories– where polarity categories are neglected is 87% with a kappa of 0.73. As can be seen from table 3 (column 6), all three categories are reliably identifiable. Interesting is that the new distinction within the subjective items between AE and SW does not cause considerable lower performance. Category AE is rather small but easy to identify (cf. table 3).

	AE	SW	OBJ	Totals	Kappa
AE	38	8	0	46	0.81
SW	6	359	20	385	0.72
OBJ	2	44	132	177	0.73
Totals	46	411	152	609	

Table 3: Confusion Matrix – attitude

- Both Polarity and Attitude Holder

Agreement for the full annotation scheme with polarity and attitude as combined categories, is 79 % with a kappa of 0.73. Table 4 (last column) shows that SW-pn, Obj-n and Obj-p are not reliable identifiable. As can be seen from table 4 as well, these categories have relatively few members. All large categories , like SW-pos, SW-neg and OBJ-ntr are reliable identifiable.

	AE-neg	AE-pos	SW-neg	SW-pos	SW-pn	OBJ-ntr	OBJ-neg	OBJ-pos	Totals	Kappa
AE-neg	29	0	2	0	2	0	0	0	33	0.87
AE-pos	0	9	0	4	0	0	0	0	13	0.63
SW-neg	3	0	161	3	2	3	5	0	177	0.85
SW-pos	0	3	1	159	3	4	0	4	174	0.77
SW-pn	0	0	5	13	12	2	0	2	34	0.39
OBJ-ntr	0	1	4	11	4	101	2	5	128	0.75
OBJ-neg	1	0	9	0	2	3	6	0	21	0.34
OBJ-pos	0	0	2	12	0	9	0	6	29	0.23
	33	13	184	202	25	122	13	17	609	

Table 4 : Confusion matrix for attitude and polarity

4.2 Disagreement Analysis

In this section we will discuss the systematic confusions which concern the problematic categories with a low kappa-value: Posneg (κ=0.38), OBJ-pos (κ=0.23), OBJ-neg (κ=0.34) and SW-posneg(κ= 0.37).

- Posneg vs. Positive (cf. table 2). Also in other annotation studies, a confusion between positive, neutral , and – if relevant – positive/negative is noted. These three categories seem to have similar characteristics for example with regard to the intensity of the expressed attitude and with regard to the linguistically unmarkedness of its forms (e.g. *honest, reliable*), whereas the category 'negative' includes many items that express more intense attitudes and that are morphologically recognizable by a negativity marker (e.g. *unreliable, dishonest*). In fact, Kim and Hovy (2004) reported that inter-annotator agreement of their annotations increases from 76% to 89% when positive and neutral categories are merged.
- OBJ-neg vs. SW-neg (cf. table 5). The results show that OBJ-neg (0.34) is easily confused with SW-neg (cf. table 11). Examples of cases where annotators disagree are *ongelukkig (disabled), oud (old, having lived for a relatively long time), doofstom (mute)* and *kaalhoofdig (bald-headed).* These word senses refer to observable and truth-conditional concepts for which they may be considered as objective. However, they will be often used to express an appreciation of somebody's looks or capabilities for which they may be considered as subjective. According to the guidelines they should have been annotated as OBJ-neg. Another subgroup of problematic cases is formed by are words like *droog (dry, free from liquid), langzaam (slow, not moving quickly),* and *onscherp (vague, shadowy).* These words refer to measurable and observable concepts for which they are annotated as objective; however, these concepts are also relative (a slow train is faster than a fast bike) and sometimes subjective (i.e. people may have different opinions of what is slow or vague) for which they are annotated as subjective. Again, according to the guidelines they should be annotated as OBJ-neg or even OBJ-neutral.
- OBJ-pos vs. SW-pos (cf. table 5). Likewise, there is confusion between OBJ-pos and SW-pos. These cases seem to fall into the same group as *slow* and *vague* as they refer to measurable but also relative and sometimes subjective concepts. Examples are *luchtig (light, airy), zuiver(pure)* and *fijn (detailed).*
- SW-posneg vs. SW-pos and all other categories (cf. table 7). Category SW-posneg is easily confused with all other categories. This may be explained by the more general confusion between positive and positive/negative. However, there is one particular group of interest, that of the intensifiers with word senses like *heftig (intense), razend (extreme), zwaar (heavy-a heavy cold).* Sometimes, by some annotators these words are viewed as having contextual polarity (their polarity depends on the word they modify) and then annotated 'positive/negative'; and by others they are annotated 'positive' meaning something like 'highly intensive'. We think that none of these annotations is satisfactory as they are both difficult to interpret and that it might be better to isolate the intensifiers from the other adjectives and treat them differently.
- AH vs. SW. Although confusion between the categories AH and SW concerns only a few cases, it is interesting as it is related to the extra features of our annotation schema. An example is the word *belust (bent on)* like in a sentence as *hij is belust op geld (he is bent on money). Belust* is annotated by one annotator as AH-pos and by the other as SW-neg; these contradictory annotations are due to the fact that the word expresses two attitudes: a positive attitude from the subject 'he' (AE) towards money and a negative attitude of the Speaker/Writer towards the behaviour of the 'he' who is longing for 'bad' things (like money). Both annotations are correct which may lead to the conclusion that the model should allow for double annotations. We decided not to introduce double annotations as some experiments pointed out that this is too complicated to perform in a consistent and reliable manner. It is however interesting to note that the further distinction of 'subjectivity' in SW and AH , helps to disambiguate really ambiguous annotations like subjective-positive vs. subjective-negative for the same word as, in the new model, the polarity is related to the attitude holder.

What all these cases have in common is that they are very hard to solve. They are not caused by individual biases of the annotators but all annotators confuse them equally and it is the question if better guidelines would be of any help.

4.3 Comparison to other annotation works

We will compare our results with the following works:
Jijkoun and Hoffmann (2007), created a goldstandard for Dutch subjectivity words. The data set includes 1916 adjectives which are annotated for 3 polarity categories (positive/negative/neutral) by 2 annotators. Most important difference with our work is that the annotation is done at word form level instead of at word sense level and that they do not annotate subjectivity but only polarity.

For English, the General Inquirer lexicon (GI) (Stone, 1966) is often used as a gold standard. GI includes hand-labeled words with – among others - labels like "Positiv" or "Negativ". Labeling is performed at word sense level but the degree of polysemy is extremely low. Andreevskaia and Bergler (2006) calculated inter annotator's agreement between GI's polarity annotation and that of a data set compiled and hand-annotated by Hatzivassiloglou et al. (1997).

Closest to our work is the gold standard for English compiled by Su et al. (2008). They re-annotated the

MICRO-Wnop corpus (Cerini et al., 2007) with labels for both polarity and subjectivity. An important difference with our study is that their anotations have been applied at synset level and that their statistics do not distinguish between nouns, verbs and adjectives. Another difference is that their subjectivity annotations do not refer to attitude holdership.

	Polarity	attitude holder	Both
current study	86% K=0.80	87% κ=0.73	79% κ=0.73
Jijkoun et al. (2009)	79% K =0.66	-	-
Andreevskaia et al. (2006)	79%	-	-
Su et al. (2009)	89% K =0.83	90% κ=0.79	85% κ=0.77

Table 6: comparison inter annotator results with other studies

As can be seen from table 6, inter-annotator agreement on polarity of our work is in agreement with the agreement measured by other studies ranging between 79 % (κ=0.66) and 89% (κ=0.83). Jijkoun et al.'s results for Dutch and Andreevskaia et al.'s results for English, which are lower than ours, show that manual polarity annotation at sense level performs considerably better than polarity annotation at word level. With regard to both attitude holder annotation and full annotation, our scores are lower than Su et al.'s scores for English. This may be due to the fact that we introduced the extra layer of attitude holder with 3 categories and therefore have a more complex annotation schema. Another reason may be that 51% of the English data set is annotated as Objective-neutral which is a category rather easy to identify, while we reduced the amount of neutral senses from 40% to 20% resulting in a more mixed but also more complex data set.

5. Gold standard

5.1 Final Gold Standard

A third annotator (A3) annotated all 609 items to adjudicate the disagreements between A1 and A2. Inter-annotator agreement dropped remarkably: overall agreement by 3 annotators is achieved for 59% of the items only. As we do not have inter-annotator statistics for 3 annotators from other studies, it is diffult to interpret these results. Usually, disagreements are discussed and resolved between the two annotators.

	A1-A2-A3	A1-A2	A1-A3	A2-A3
polartiy	70.6%	86.4% κ 0.80	77% K 0.67	75% κ 0.65
attitude holder	71.5%	87% κ 0.73	79.4% K 0.61	76.8% κ0.55
Both	59.2%	79.3% κ0.73	67.3% K 0.59	65% κ0.56

Table 7: Overall Inter-annotator agreement – 3 annotators

The effect of training sessions and guideline preparation – in which A3 was not involved - seems to be rather high as A1 and A2 have considerable higher mutual agreement (79.3%) than A1 and A3 (67.3%), and A2 and A3 (65%) respectively. Many disagreement cases are caused by A3's bias towards the categories 'ntr' and 'posneg' at the cost of the more salient categories 'neg' and 'pos'. Better agreement might probably be achieved by more training of the third annotator and by discussing the disagreements. However, we decided not to do so and to use the third jugdment without further correction, for (1) the adjudication of disagreements between A1 and A2; and for (2) the identification of that part of the gold standard items which can be considered as the core of the sentiment category. With regard to this last purpose, we follow Andreevskaia and Bergler (2006) who consider the category of sentiment as a fuzzy category where some words are very central, prototypical members while other are less central words that may be interpreted differently by different people. According to them, the variability should be considered as inherent to the meaning of the word. Better training may solve disagreements caused by individual bias or perspective but also tends to eliminate disagreements caused by these inherent aspects. We hypothesize that items with high agreement coincide with the core members of the sentiment category and we regard this distinction between core and non core members as an important extra dimension of the gold standard.

Tables 8 presents the statistics of the final gold standard with one judgement for each item. Agreement is high (H) for those items where 3 annotators agree, medium (M) where only 2 agree. In those cases where all annotators disagree the agreement score is set to low (L) and the final annotation label is set to 'posneg' - in the case of polarity values - and to the combination of the agreed categories - in the case of the full annotation scheme (cf. table 8, columns H, M and L).

Polarity and Attitude Holder				
	Total	H	M	L
AE-neg	32	23	9	0
AE-pos	14	8	6	0
SW-neg	176	131	42	3
SW-pos	177	110	65	2
SW-pn	36	5	19	12
OBJ-ntr	129	82	47	0
OBJ-neg	15	2	6	7
OBJ-pos	24	0	13	11
OBJ-pn	6	0	0	6
Total	609 (100%)	361 (59%)	207 (34%)	41 (7%)
Polarity				
	Total	H	M	L
Neg	223	193	30	0
Pos	215	150	65	0
Ntr	129	82	47	0
Posneg	42	5	19	18
Total	609 (100%)	430 (71%)	161 (26%)	18 (3%)
Attitude Holder				
	Total	H	M	L
AE	46	32	14	0
SW	385	302	83	0
OBJ	178	102	76	0
	609 (100%)	436 (72%)	173 (28%)	0 (0%)

Table 8: Statistics Gold Standard

With regard to polarity, the final gold standard consists of equal parts for positive (215 (35%)) and negative (223(37%)) items, a smaller portion (129(21%)) of neutral items and a relatively small part (42(7%)) of mixed positive/negative items. With regard to attitude holder, the distribution of the items across the different categories is less balanced: SW is a large category presented with 63% (385 items) while the AE category has 46 (8%) members only. (cf. table 8, column 'Total').

5.2 Gold standard across different lexicon layers

- Frequency, polysemy and large-synset-membership

We can now explore the correlations between the human annotations and the different lexicon dimensions, like frequency, polysemy and large-synset-membership. Table 9 shows the distribution of the items with reliable annotations, i.e. those annotations which are agreed upon by all three annotators, across these dimensions. For example, from table 9 (row 3) can be seen that – with regard to polarity annotation – 65% of the highly polysemous items prove to be reliably identifiable ; 69% of the less polysemous and 75% of the low polysemous (i.e. monosemous) items are reliable identifiable. This implies that highly polysemous items are harder to identify with regard to polarity annotation than less polysemous items . This is in line with what we expect: the usually more or less related meanings of polysemous words may be hard to separate from each other and lead to ambiguous annotations. The same pattern holds for the combination of polarity annotation and frequency. We see a negative correlation between polarity annotation and frequency (cf. row 4): the more frequent the items are, the harder they are to annotate in a reliable manner. As the number of senses increases with frequency, we indeed expect similar results for frequency and polysemy. With regard to large synset membership there is a positive correlation (cf. row 5): items that are members of large synsets are more easy to annotate than items that are not. This can be explained by the observation (Maks and Vossen (2010)) that subjective items tend to cluster in large synsets. These 'really' subjective items are characterised by a lack of denotation – for which they are easily grouped together - and a high degree of connotational polarity – for which they are so easy to annotate.

Polarity			
	high	Mid	Low
Polysemy	65%	69%	75%
Frequency	59%	72%	78%
LargeSynset	79%	68%	65%
Attitude Holder Annotation			
	high	Mid	Low
Polysemy	71%	68%	76%
Frequency	71%	67%	74%
LargeSynset	79%	69%	67%

Table 9: correlation annotation and lexicon dimensions

There is an equal positive correlation between attitude holder annotation and large synset membership (cf. row 10). However, we cannot see a coherent pattern for the combination of attitude holder annotation with polysemy or frequency (rows 8 and 9) . Items which are in the mid categories of polysemy and frequency seem harder to reliably identify than items in the low and high categories. We may conclude that – at least on the basis of these results- there is no clear relation between, on the one side, the type of attitude holdership and/or degree of subjectivity of a word and , on the other side, its being frequent or polysemous.

- Ambiguous Words

Ambigous words are words that are polysemous whose senses are differently annotated. To know the effects annotating word senses over annotating word forms, we counted the amount of words that are ambiguous for attitude holdership and/or polarity. The data set includes 344 items (lexical units) which belong to 125 polysemous words. 76 (61%) of these words have senses with different annotations: 61 (49%) is polarity-ambiguous and 61(49%) is ambiguous with regard to annotation holder. These numbers confirm that a considerable number of disagreement can be avoided by annotation at word sense level.

- Ambiguous Synsets

Ambiguous synsets are synsets with several members which are differently annotated. The 609 lexical units of our goldstandard are distributed across 576 different synsets of which 373 have more than one member (synset size Low and High) . We did not select all synonyms of these 'plurimember' synsets in the gold standard but we took care to include some of them. As a result, 68 synsets have more than one member actually included in the gold standard. This makes it possible to see to what extent, synsets are internally ambiguous with respect to polarity or attitude holder. Of the 68 synsets, 21 (31%) have members with different annotations: 15 (22%) are polarity ambiguous and 11 (16%) are ambiguous with respect to the attitude holder. These numbers seem to support our claim (Maks and Vossen, 2010) that fine-grained polarity and subjectivity (as modelled in our annotation schema) cannot be annotated at synset level but must be annotated at the level of the individual lexical units unless the synsets are restructured in such a way that they are not ambiguous any more.

6. Conclusions

We argued that existing annotation schemata for polarity and subjectivity must be extended with an extra layer for the atttitude holder and we showed that these more complex annotation schema can be applied in a reliable manner leading to large groups of reliable annotations. Although the new categories on the level of attitude holder, i.e. Speaker/writer and Agent/Experiencer, prove to be rather small, they are annotated in a reliable manner. We also showed the advantage of making a gold standard that is representative of the lexicon as a whole as opposed to using sentiment word lists compiled indepedently form the lexicon. As we took great care that all dimensions of the lexicon that are relevant to subjectivity and polarity identification, like frequency, polysemy and large synset membership, are equally represented in this gold standard, we could find correlations between the human annotations and these three dimensions. We assume that these correlations will also be relevant for automatic annotations.

In future we will apply the guidelines on other word categories like nouns , verbs and adverbs. Moreover we will use the gold standard to test methods and techniques to build a sentiment lexicon for Dutch.

7. Acknowledgments

This research has been carried out within the project From Text To Political Positions (http://www2.let.vu.nl/oz/cltl/t2pp/). It is funded by the VUA Interfaculty Research Institute CAMeRA.

8. References

Andreevskaia, Alina and Sabine Bergler (2006). Sentiment Tagging of Adjectives at the Meaning Level. In *LNAI 4013:. Advances in Artificial Intelligence. 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI-2006*, Springer-Verlag, Heidelberg and Berlin, Germany.

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). *Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming)*, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

Esuli, Andrea and Fabrizio Sebastiani. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC-2006*, Genova, Italy.

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

Hatzivassiloglou, V., McKeown, K.B. (1997) Predicting the semantic orientation of adjectives. In *Proceedings of ACL-97*, Madrid, Spain.

Kim, S. and E. Hovy (2004) Determining the sentiment of opinions. In *Proceedings of COLING*, Geneva, Swtizerland.

Kim, S.M. and E.H. Hovy. (2006) Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY.

Kamps, J., R. J. Mokken, M. Marx, and M. de Rijke (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings LREC-2004*, Paris.

Maks, I. and P. Vossen (2010) Modeling Attitude, Polarity and Subjectivity in Wordnet. In *Proceedings of Fifth Global Wordnet Conference*, Mumbai, India.

Martin, J. R. and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave, London, UK. (http://grammatics.com/appraisal/).

Miller, G. (1998) Introduction. In: WordNet: An Electronic Lexical Database (C. Felbaun (ed.). The MIT Press, Cambridge, Mass.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Su, F.and K. Markert Eliciting Subjectivity and Polarity Judgements on Word Senses. In *Proceedings of Coling-2008*, Manchester, UK.

Wilson, T. , Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP-2005).

Wiebe, J., Theresa Wilson , Rebecca Bruce , Matthew Bell , and Melanie Martin (2004). Learning subjective language. Computational Linguistics 30 (3).

Wiebe, Janyce and Rada Micalcea.(2006) . Word Sense and Subjectivity. In *Proceedings of ACL'06*, Sydney, Australia.

Wiebe J., Theresa Wilson , Rebecca Bruce , Matthew Bell , and Melanie Martin (2004). Learning subjective language. Computational Linguistics 30 (3).

Vossen, P., I.Maks, R. Segers and H. van der Vliet (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In *Proceedings of LREC-2008*, Marrakech, Morocco