**L**inguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

# Compositional Syntax-based Phrase-level Polarity Annotation for German

**Manfred Klenner**
**Simon Clematide**
**Stefanos Petrakis**
**Marc Luder**

Published by CSLI Publications

**LiLT** volume 7, issue 15 January 2012

# Compositional Syntax-based Phrase-level Polarity Annotation for German

Manfred Klenner, *University of Zürich, Institute of Computational Linguistics* Simon Clematide, *University of Zürich, Institute of Computational Linguistics* Stefanos Petrakis, *University of Zürich, Institute of Computational Linguistics* Marc Luder, *University of Zürich, Department of Psychology*

## Abstract

We introduce the task of word and phrase-level polarity annotation for German as part of an attempt to develop a compositional theory of clause-level polarity determination. Thus, annotations should give access to the nested building blocks, the structural strata of polarity composition. Therefore and in contrast to existing polarity-tagged corpora, we annotate not exclusively on the basis of surface strings, but argue that proper polarity annotation of complex phrases requires access to their syntactic structures. We discuss the principles of our treebank design, and present the inter-annotator agreement of our kick-off annotations on a test suite of 270 sentences that was compiled specifically to contain interesting polarity combinations.

**LiLT** Volume 7, Issue 15, January 2012.
*Compositional Syntax-based Phrase-level Polarity Annotation for German.*

## 1 Introduction

Our work is part of a larger initiative that strives to produce a multi-layered (word, phrase and sentence level, but also opinion target and opinion holder) German reference corpus for sentiment analysis[1]. On the basis of the large DeWaC corpus (Baroni et al., 2009), we have started producing a polarity-tagged test suite, where words and phrases are annotated with regard to their sentiment orientation. We sampled 270 sentences according to one of the following criteria: They contain at least (a) an intensifier and a polar word, (b) a shifter without an intensifier, (c) a positive and a negative word within a sentence. The word polarities were taken from our freely available polarity lexicon (Clematide and Klenner, 2010) comprising 8,000 lemmas.

We soon got aware of the fact that the proper recursive segmentation and annotation of complex phrases could only succeed if it were based on a common syntactic annotation framework. This is in contrast to any other existing annotated polarity corpora (e.g. MPQA (Wilson et al., 2005)), where annotations are carried out solely on surface strings. There, only the maximum span of phrases is identified and given a polarity. The polarity of embedded substrings is left undetermined. Such an annotation strategy does not lead to insights regarding the interplay between the building blocks of complex phrases. This, however, is needed in order to better understand, derive or learn a compositional theory of polarity determination. Moreover, adopting and adhering to the TIGER corpus annotation guidelines (Brants and Hansen, 2002) allowed annotators to annotate syntactic structures in a firmer and more consistent way, which became evident from inter-annotator agreement measurements.

The various factors determining the polarity of a phrase or a clause have first been discussed in Polanyi and Zaenen (2006). The authors point out that some words modify the polarity of other words. For instance, *fail to win*, where *fail* inverts the positive polarity of *win* and thus produces a negative verb phrase (VP). In Moilanen and Pulman (2007), a compositional treatment of sentence-level polarity is sketched. However, a thorough discussion of the underlying polarity decisions is missing. Recently, Neviarouskaya et al. (2010) introduced a more elaborate and robust compositional theory for English, especially tailored to the verb level. Less attention is paid to the analysis of the challenges given by complex noun phrases (NPs) and prepositional phrases (PPs).

[1] See http://synergy.sentimental.li for more information about the goals and efforts of this initiative.

## 2 Treebank Design and Annotation Process

We use the term polarity analysis to refer to the task of assigning a positive, negative or neutral sentiment orientation to words, phrases and sentences. Clearly, subjective words (e.g. *wonderful*) and phrases expressing opinions (e.g. *I adore...*) are of interest here, but we are also interested in factual polarity, i.e., whether a word, a phrase or a sentence denotes a positive or negative object or situation. For instance, the sentence *Ten passengers survived* is factual, but it normally also evokes positive emotions (here relief). We see factual polarity as a sentiment orientation which is not opinionated, but which still is emotionally affecting. Most existing approaches only care about subjective expressions, although there are notable exceptions such as the work by Neviarouskaya et al. (2010).

Another difference to most existing approaches is that we believe in compositionality[2], that is, phrase-level polarity is a function of word polarity and clause polarity is a function of phrase-level polarity. We are actually interested in the question of how far could such an approach lead without the need for deeper sentence semantics.

Polarity annotation is a somewhat daring enterprise, since the polarity of expressions often depends on ideological perspective, personal and cultural preference and even the philosophical or religious stance people adhere to. One has to cope with the full range of phenomena, including cases of factual polarity (*punishment of the murderer*), statements related to moral standards (*justified punishment*) and emotions (*longing for happiness*) as well as target-specific polarity (*old wine*).

In order to produce gold standard data with a reasonable inter-annotator agreement one has to make the underlying commitments explicit, i.e., to answer the question of how we can decide which polarity a word and a phrase should bear. As a psychological foundation for our polarity annotations we use a classification schema consisting of eight generic emotion categories, as proposed by Dahl (1978), based on the "decision theory of emotions" by de Rivera and Dahl (1977). The eight generic emotions and their associated polarity are *love* [+], *enthusiasm* [+], *anger* [–], *fear* [–], *satisfaction* [+], *joy* [+], *depression* [–] and *anxiety* [–]. During the annotation process, all words and phrases with an emotional connotation are assigned to an emotional category, thereby obtaining their polarity.

Although such a foundation in psychological terms might be sufficient, more detailed annotation principles facilitate the annotation

[2]Other approaches based on compositionality of sentiment include Moilanen and Pulman (2007) and Neviarouskaya et al. (2010).

| Description | Tag | # Words | Examples | Top Phrases | All Phrases |
|---|---|---|---|---|---|
| | | Word Level | | Phrase Level | |
| positive | + | 335 | hope | 158 | 275 |
| negative | − | 362 | doubt | 180 | 300 |
| intensifier | ^ | 63 | heavy | | |
| diminisher | % | 9 | low | | |
| shifter | ~ | 51 | against | | |
| bipolar | # | | | 21 | 54 |
| neutral | 0 | | | 10 | 12 |

FIGURE 1 Distribution of Polarity Tags for the Development Set

decisions. For instance, the expression *confession of murder* can be considered as positive (disclosure of the truth) or negative (a murder was committed). This might well depend on the context, but both views are possible. Not instructing the annotators which perspective to systematically adopt is bound to lead to inconsistencies and low agreement, as was experienced by Balahur et al. (2010). Consequently we have formulated principles, such as the principle "to reveal the truth is positive" which would apply to the previous example. Other principles are: vitalism (to cherish life is positive), humanism (to be a human being is positive) and egalitarism (to adhere to human rights is positive), but also more detailed instructions such as "to act against evil is positive" (e.g. *accusation of the murderer*). We know how fragile our guidelines still are, but we believe that we are heading in the right direction.

### 2.1 Word and Phrase Level Annotation

We started the annotation of noun and prepositional phrases in our corpus by annotating the word-level polarity first. Only words which appear as part of an NP or PP were annotated. As can be seen in Figure 1, we consider 5 classes which express or modify polarity on the word level. Annotation decisions are based on the meaning of words in context. For instance: *human* in *human gesture* is positive while in *human body* it is neutral.

We decided to annotate the phrase structure boundaries of NPs and PPs according to the TIGER corpus guidelines (Brants and Hansen, 2002). In the following, we use brackets to indicate the structure of phrases. A subphrase is annotated if it contains at least one polar word. Phrasal modifiers of a polar phrase are integrated even if they do not contain a polar word (*[the honest+ compliment+ of the doctor]+*). If a modifier contains a polar word its head phrase is annotated as well (*[the doctor's [honest+ compliment+]+]+*). Relative and infinitival clauses

are only integrated if they are inside a complex phrase, they are currently not included if located at the end of the phrase. Adjective phrase brackets are not inserted and annotated, for the moment.

In order to make the polarity of a complex phrase transparent, the polarity of the embedded phrases must be combined according to the syntactic structure of the complex phrase. So for example in the phrase *die Strafe der Einsamkeit für ihre notorische Nostalgie der kolonialen Zeiten* (*the punishment of solitude for her notorious nostalgia of the colonial periods*), the PP starting with *for* modifies *punishment* not *solitude*. We assign the following structure: *[the punishment-[of solitude]- [for her notorious- nostalgia+ of the colonial periods]-]-* . The PP *of the colonial periods* is not grouped since it contains no polar words.

In order to cope with cases of mixed polarities (e.g. positive and negative NPs in a coordination structure), the symbol # is used. We consider these kinds of phrases as bipolar, e.g. *[[grausamer- Spott-]- und [mitfühlender+ Trost+]+]#* or *[[mercyless- mockery-]- and [compasionate+ consulation+]+]#*.

### 2.2 Inter-Annotator Agreement

Three annotators discussed the decisions for 270 sentences and tried to resolve conflicts. This original set of sentences can thus be regarded as our development set and it was the basis for the annotation guidelines. On the development set we reach full agreement in 91.4% of 641 phrases. For the 369 top-level phrases in our development set, Fleiss' Kappa is 0.86, for all 641 phrases, Kappa is 0.90.

In an evaluation experiment with 30 sentences, we measured the inter-annotator agreement between 3 separate annotators on the word level: In 133 cases at least one annotator assigned a non-neutral word polarity. In 66.2% we reached agreement; Fleiss' Kappa for m raters is 0.69. However, Kappa varies strongly between polarity classes (–: 0.86, ~: 0.84, +: 0.76, ^: 0.61, %: 0.24). If we stick to the 93 cases where all annotators assigned non-neutral polarity we reach 92.5% full agreement. When we adjust these word annotations we reach an upper bound of agreement with our current guidelines; Kappa after this harmonization step is 0.87. For the inter-annotator agreement evaluation on the phrase-level, we used these adjusted word-level annotations. For our 52 top-level phrases, Kappa is 0.79; for all phrases (98 items including the top ones), Kappa is 0.81.

## 3 Related Work

A renowned resource available for sentiment analysis is the MPQA corpus (Wilson et al., 2005). The annotation of sentiment in the MPQA applies to expressions of private states, a major difference to our approach. Another major difference is that they provide labels for contextual polarity of expressions, while we mark the context's influence on the composed level (expression+context), usually one level higher. An example taken from the MPQA Corpus (Wilson et al., 2005) is the subjective expression *the fight against terrorism, violence and intrigues*, which has a negative polarity label resulting from the context's influence. We, on the other hand, would mark it as positive considering the phrases' constituents and let the potential negative evaluation of the sentence that contains this expression occur at a higher level. This difference in annotation philosophy is fundamental and although the final, top-level polarity decision may be the same, the resulting resources are fairly different.

Another quite extensive approach in sentiment annotation is the ICWSM 2010 JDPA Sentiment Corpus (Kessler et al., 2010). This initiative clearly acknowledges the importance of the structure of text and is in that aspect quite similar to ours. They nevertheless also set off from subjective/opinionated language, annotating sentiment that targets specific entities (or mentions of entities). Another difference is their focus on a specific domain, that of reviews of automobiles. We should also mention that they ignore contextual polarity and let it be inferred from the interaction of the constituents, in the same way that we do.

As far as classification schemes for emotion and their application to textual annotation are concerned, we should mention the work done by Volkova et al. (2010). The main difference to our approach is that their annotation is not based on syntactic constituents, merely "...stretches of text where an emotion was to be conveyed..." (Volkova et al., 2010).

The mentioned classification schema of emotions by Dahl (1978) was the background for the development and implementation of the German affective dictionary ADU (Affektives Diktionär Ulm). Items of the ADU are words with an emotional connotation on the single-word level. Raters assigned all 2000 entries of the dictionary to the eight generic emotion categories. The frequency of emotional words in general discourse is about 4%, in psychotherapy discourse about 10%. The ADU was applied to a series of studies concerning the emotional vocabulary of psychotherapy patients; for a summary see Thomä and Kächele (2006). Linguistic Inquiry and Word Count (LIWC), a text analysis ap-

plication developed by the psychologist James Pennebaker, is equally grounded on a single-word dictionary (Pennebaker and Francis, 2001). The dictionary entries are grouped by 80 variables determining linguistic properties and psychological processes, e.g. positive emotion and negative emotion. The English version of the LIWC was applied and validated in several studies (Pennebaker, 2011). There exists also a German version of the LIWC with a German dictionary. It is investigated in two studies; most of the LIWC categories display high equivalence to their English counterparts (Wolf et al., 2008).

## 4 Conclusions and Future Work

Our ultimate goal is the specification of a linguistic theory of compositional polarity determination. In contrast to most approaches in the field of sentiment analysis, we assign a polarity to both facts and opinions. This gives rise to a theory where the agreement of expressed opinions with normative common sense principles can be measured.

We have discussed our annotation guidelines for phrase-level annotation. The focus lies on complex NPs and PPs. In order to understand how compositionality works at this level, we have argued for the need to base annotations of sentiment on syntactic structures as found in treebanks (currently at the phrase level only). The proper annotation of the polarity of complex phrases and their parts is impossible without recourse to their syntactic structure.

Future work is devoted to the refinement of our theory. This goes hand in hand with further annotations, namely on the level of adjective and verb phrases as well as on the clause level. Theory development, in our case, is tightly coupled to treebank creation and annotation. On the technical side we also intend to move from the current textual-based format[3] to a standoff XML format as created by tools like MMAX2 (Müller and Strube, 2006) or PALinka (Orasan, 2003) in order to provide for better sustainable data.

## Acknowledgments

This work is part of a larger initiative whose goal is to produce a multi-layered German reference corpus for sentiment analysis (see http://synergy.sentimental.li for further information).

This work is partly funded by the Swiss National Science Foundation (grant 100015_122546/1).

[3]The annotated test suite in textual-format is available at http://synergy.sentimental.li/Downloads .

## References

Polanyi, Livia and Zaenen, Annie. 2006. Contextual Valence Shifters. In *Computing attitude and affect in text: Theory and applications*, pp. 1–10.

Brants, Sabine and Hansen, Silvia. 2002. Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1643–1649, Las Palmas.

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano and Zanchetta, Eros. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. In *Language Resources and Evaluation*, pp. 209–226.

Clematide, Simon and Klenner, Manfred. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 7–13.

Wilson, Theresa, Wiebe, Janyce and Hoffmann, Paul. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347–354.

Kessler, Jason, Eckert, Miriam, Clark, Lyndsie and Nicolov, Nicolas. 2010. The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain. *International AAAI Conference on Weblogs and Social Media Data Challenge Workshop.*

Balahur, Alexandra, Steinberger, Ralf, Kabadjov, Mijail, Zavarella, Vanni, Van Der Goot, Erik, Halkia, Matina, Pouliquen, Bruno and Belyaeva, Jenya. 2010. Sentiment analysis in the news. *Seventh Conference on Language Resources and Evaluation.*

Moilanen, Karo and Pulman Stephen. 2007. Sentiment Composition. In *Proceedings of the Recent Advances in Natural Language Processing International Conference*, pp. 378–382.

Neviarouskaya, Alena, Prendinger, Helmut and Ishizuka, Mitsuru. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 806–814.

Thomä, Helmut and Kächele, Horst. 2006. Psychoanalytische Therapie: Forschung. Springer.

Dahl, Hartvig. 1978. A New Psychoanalytic Model of Motivation: Emotions as Appetites and Messages. In *Psychoanalysis and Contemporary Thought* 1(3), pp. 373–408. International Universities Press, Inc.

Rivera, Joseph de and Dahl, Hartvig. 1977. A structural theory of the emotions. In *Psychological issues* 10(4). International Universities Press, Inc.

Müller, Christoph and Strube, Michael. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, pp. 197–214.

Orasan, Constantin. 2003. PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pp. 39–43.

Pennebaker, James and Francis, Martha. 2001. Linguistic Inquiry and Word Count. Hillsdale (N.J.): Lawrence Erlbaum Associates.

Pennebaker, James (2011). The secret life of pronouns: What our words say about us. New York: Bloomsbury.

Wolf, Markus, Horn, Andrea, Mehl, Matthias, Haug, Severin, Pennebaker, James and Kordy, Hans. 2008. Computergestützte quantitative Textanalyse. In *Diagnostica* 54(2), pp. 85–98.

Volkova, Ekaterina, Mohler, Betty, Meurers, Detmar, Gerdemann, Dale and Bülthoff, Heinrich. 2010. Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 98–106.