MLSA – A Multi-layered Reference Corpus for German Sentiment Analysis

Simon Clematide[*], Stefan Gindl[ϒ], Manfred Klenner[*], Stefanos Petrakis[*], Robert Remus[+], Josef Ruppenhofer[ψ], Ulli Waltinger[ϕ], Michael Wiegand[π]

University of Zürich, Institute of Comp. Linguistics,[*], MODUL University Vienna, Department of New Media Technology[ϒ], University of Leipzig, Department of Computer Science, Natural Language Processing Group[+], University of Hildesheim[ψ], University of Bielefeld, Artificial Intelligence Group[ϕ], Saarland University, Spoken Language Systems[π]

{klenner,siclemat,petrakis}@cl.uzh.ch[*], stefan.gindl@modul.ac.at[ϒ], rremus@informatik.uni-leipzig.de[+], josef.ruppenhofer@uni-hildesheim.de[ψ], uwalting@techfak.uni-bielefeld.de[ϕ], michael.wiegand@lsv.uni-saarland.de[π]

Abstract

In this paper, we describe MLSA, a publicly available multi-layered reference corpus for German-language sentiment analysis. The construction of the corpus is based on the manual annotation of 270 German-language sentences considering three different layers of granularity. The sentence-layer annotation, as the most coarse-grained annotation, focuses on aspects of objectivity, subjectivity and the overall polarity of the respective sentences. Layer 2 is concerned with polarity on the word- and phrase-level, annotating both subjective and factual language. The annotations on Layer 3 focus on the expression-level, denoting frames of private states such as objective and direct speech events. These three layers and their respective annotations are intended to be fully independent of each other. At the same time, exploring for and discovering interactions that may exist between different layers should also be possible. The reliability of the respective annotations was assessed using the average pairwise agreement and Fleiss' multi-rater measures. We believe that MLSA is a beneficial resource for sentiment analysis research, algorithms and applications that focus on the German language.

Keywords: Sentiment Analysis, Emotion detection, Lexical resource

1. Introduction

Sentiment analysis is a highly active research area that embraces not only work on the identification of opinions, emotions and appraisals, but also on the construction of corpora and dictionaries. While various approaches and resources have been proposed for polarity or subjectivity classification for English (Pang et al., 2002; Wilson et al., 2005), relatively few benchmark collections and corpora that focus on German have been made available. Moreover, with respect to existing work on corpora for sentiment analysis and opinion mining, most approaches have focused on user-rated product reviews at document-level, even though multiple opinions and factual information may be found within single sentences.

In this paper, we present MLSA, the result from a European research collaboration that aims to provide a publicly[1] available multi-layered reference corpus for sentiment analysis in German. The compilation of the MLSA corpus is based on manual annotation at different layers of granularity (cf. Figure 1.) using a set of 270 sentences. Within Layer 1, each sentence has been analyzed according to the notions of subjectivity/objectivity and their polarity, i.e. positive, negative or neutral. On Layer 2, the word- and phrase-level has been targeted, focusing on aspects of subjective and factual language. Layer 3 covers annotations on the expression-level, using the notions of private state and speech. Included in its annotations are the sources and targets of opinions. Each layer has been annotated by multiple raters, and the annotations' quality has been assessed by two different inter-annotator agreement measures.

The rest of the paper is structured as follows: In Section 2. we present related work. Section 3. describes the multi-layered reference corpus for German-language sentiment analysis and provides an overview of the data representation and the annotation schemata applied. Section 4. presents the assessment of the inter-annotator agreement and finally, Section 5. concludes this paper.

[1]The corpus is publicly available: http://synergy.sentimental.li/Downloads

2. Related Work

A plethora of sentiment-related corpora is available for English. Whereas earlier work strongly focuses on coarse-grained classification tasks, such as document-level polarity classification (Pang et al., 2002) there has lately been a shift of attention towards more fine-grained tasks dealing with polarity and subjectivity on sentence-level, phrase-level or even expression-level. Though for the former labeled data can be automatically generated (Pang and Lee, 2005; Blitzer et al., 2007), for instance by deriving the polarity from user ratings in product reviews, the latter requires manual annotation (Wiebe et al., 2005; Toprak et al., 2010). The increasing significance of sentiment analysis in natural language processing is also reflected by two benchmark tasks: TAC Opinion Question Answering (Dang, 2009) and NTCIR Multilingual Opinion Annotation Task (Seki et al., 2010), providing text collections for their respective tasks as well. Comparing the availability of English-language resources with the few corpora that are currently available for German (e.g. Remus and Hänig (2011)), the need for further resources becomes obvious.

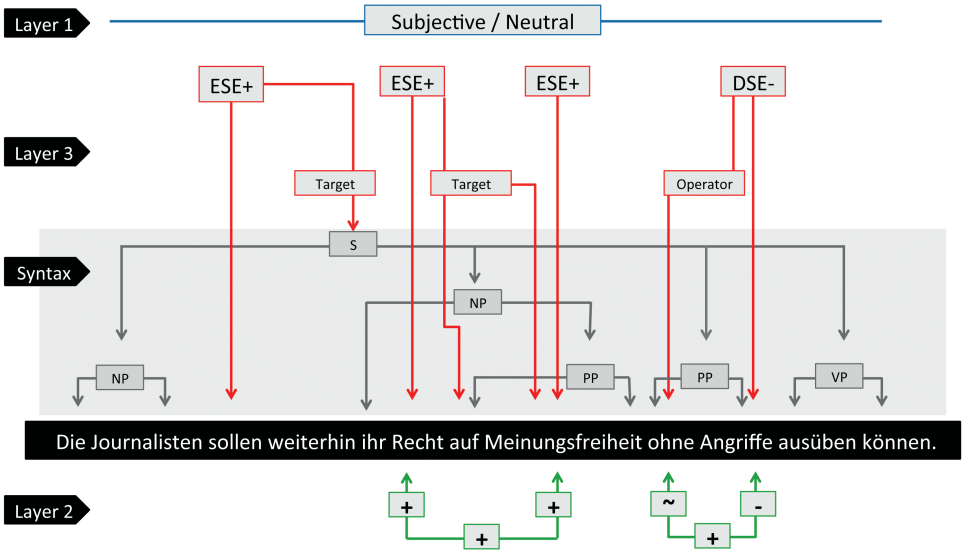

Figure 1: Excerpt from the multi-layered annotation.

3. A Multi-layered Reference Corpus for German Sentiment Analysis

For the construction of the multi-layered MLSA reference corpus, we used a set of sentences extracted from the DeWaC Corpus (Baroni et al., 2009). The DeWaC Corpus is a collection of German-language documents of various genres obtained from the web. DeWaC does include, but does not exclusively consist of, opinionated, expressive or polarity specific language. Its main properties are its generic nature, its sheer size and its concurrent representation of language used on the web. In order to sample sentences that better suit our research goals, we extracted those where negation of, intensification of, as well as contrasts between polar words were detected. Using such simple heuristics allowed for constructing a dataset that was sufficiently biased – up to a certain degree – towards "sentimentality", while still being generic enough. This detection was based on Clematide and Klenner (2010)'s polarity lexicon and resulted in a set of 270 sentences. Consequently, all sentences were manually annotated at three layers of granularity: we now describe each annotation layer in detail.

3.1. Layer 1: Sentence-level Annotations

Sentence-layer annotation is the most coarse-grained annotation in the corpus. We adhere to definitions of objectivity and subjectivity introduced in Wiebe et al. (2005). Additionally, we followed guidelines drawn from Balahur and Steinberger (2009). Their clarifications proved to be quite effective, raising inter-annotator agreement in a sentence-layer polarity annotation task from about 50% to more than 80%. All sentences were annotated with respect to two dimensions, subjectivity and polarity (cf. Table 1, 2). Subjectivity covers the existence of an actual attitude within a statement. Statements with purely informative content and without an explicit attitude are considered as objective, whereas statements with affective content are subjective. Factuality has two possible values, objective vs. subjective. The second dimension is the polarity of a statement. Negative polarity is equal to negative sentiment, positive polarity denotes positive sentiment and neutral polarity either denotes the lack of explicit sentiment or ambiguity within the sentence.
An example of a subjective sentence with negative polarity is:

(1) *"Das Schlimmste aber war eine mir unerklärliche starke innere Unruhe und das gleichzeitige Unvermögen, mich normal zu bewegen."*
["But the worst thing was an inexplicable severe inner restlessness and the concomitant inability to move normal."]

The sentence does not contain any obvious factual information, but only expresses the inner state of a person. An example of an objective sentence without any overt polarity is:

(2) *"Die Bewegung der extrem detaillierten Raumschiffe basiert auf realen physikalischen Gesetzen."*
["The movement of the extremely detailed spaceships is based on real physical laws."]

Non-neutral polarity can also be assigned to an objective sentence. This sounds like an oxymoron in the first place, but it becomes obvious with an example:

(3) *"Die Folge war hohe Arbeitslosigkeit im Textilgewerbe, das hauptsächlich für den Export produzierte."*
["The result was high unemployment in textile industry, which mainly produced for export."]

From a factuality point of view the sentence is objective, since it simply expresses a statement concerning the "high unemployment in textile industry". However, high unemployment is a problem for a society, rendering it's existence a negative matter-of-fact (provided someone does not argue

from an industrialists point of view, where high unemployment decreases production costs). Thus, an objective sentence might also contain a piece of information causing a negative/positive emotional response in a reader.
The different layers of MLSA are not synchronized, i.e. the annotations on one layer cannot be used to derive annotations on a different layer. MLSA contains sentences where the simple aggregation of phrase-layer polarity assessments would deliver results different from the sentence-layer assessment:

(4) *"Wenn du nicht in die Hölle willst, dann sei demütig und ertrage auch die schlimmste Folter ohne Hass auf deine Peiniger, denn es ist letztlich nur um deiner Seele Willen, sie vor der Hölle zu bewahren."*
["If you are not willing to go to hell, then be humble and endure the worst torture without hatred for your tormentors, because ultimately it is only to save your soul from hell."]

The phrase-level annotation lists four negative phrases in total, with only one positive phrase ("without hatred for your tormentors"; the negative phrase "for your tormentors" is embedded in the positive phrase). Such an annotation would suggest a negative annotation on the sentence-level as well. However, only one of the three sentence-level annotators assigned a negative label to this sentence. The same is true for the following sentence:

(5) *"Sie liefert Meldungen über das politische Ortsgeschehen, interessante Bräuche und kulturelle Veranstaltungen oder greift ernste, soziale, kirchliche, lustige oder kuriose Themen auf."*
["It provides news about the local political events, interesting traditions and cultural events or serious takes on social, religious, funny or strange issues."]

Although consisting of only positive phrases this sentence gets an exclusively neutral assessment on the sentence-level.
These "inconsistencies" show the difficulties arising when creating a corpus for sentiment analysis. Annotations from one level cannot be easily transferred or summed up to be used on another level. However, these inconsistencies also emphasize the relevance of MLSA. The annotations on all three levels were done independently, which guarantees that there are no distortions introduced by a transfer from one level to the other. Researchers interested in different aspects of Sentiment Analysis will find different aspects of the corpus useful. Moreover, it also allows for holistic approaches, which have inter-dependencies between different layers as an explicit goal.

3.2. Layer 2: Word- and Phrase-level Annotations

On Layer 2, we are concerned with polarity on the word- and phrase-level (specifically nominal phrases (NPs) and prepositional phrases (PPs)), annotating both subjective and factual language. We exploit the syntactic structure of these phrases and annotate their polarity following the interaction between their structural elements. This is a major difference compared to existing annotation efforts and is driven by what we see as the need for an annotation that is based on the syntactic structure of the textual unit at hand, which in turn could lead to an explicit compositional treatment of the polarity of complex phrases, i.e. a system that learns how to determine the polarity of a complex phrase based on its parts.

Tag	# of tags in consensus
subjective	147
objective	71
no consensus	52

Table 1: Distribution of the subjectivity and objectivity tags annotators reached a consensus on in Layer 1.

Tag	# of tags in consensus
positive	55
negative	78
neutral	75
no consensus	62

Table 2: Distribution of the positive, negative and neutral tags annotators reached a consensus on in Layer 1.

We segment NPs and PPs according to the TIGER guidelines (Brants and Hansen, 2002). Relative clauses and adjective phrase boundaries are not yet marked up as this paper is written. On the phrase-level the following polarity tags are used: + for positive, – for negative, *0* for neutral polarity and # for bipolar phrases. Moreover, phrase borders are indicated by square brackets and respective polarities are attached to the closing brackets. On the word-level three additional tags are used: % for diminishers (low), $\wedge$ for intensifiers (high) and $\sim$ for shifters (inversion). We apply manual word-sense disambiguation as we consider word polarities to be context-dependent, e.g. "menschlich" in "menschliche+ Geste" (*human gesture*) compared to "menschlicher*0* Körper" (*human body*).
We exclusively focus on annotating phrases where – via compositionality – the sentiment of a phrase could be derived from the sentiment of its constituents, either words or phrases. Because of our focus, we only annotate phrases which contain polarized constituents.
An example of our annotation scheme which exhibits the compositional aspects of sentiment is the following:

(6) *"ohne Hass auf deine Peiniger"*
["without hatred for your torturers"]

We start from the word-level, assigning the appropriate polarity tags where applicable, and get:

(7) *"ohne$\sim$ Hass– auf deine Peiniger–"*

We then segment the phrase into NPs and PPs, and assign polarity to the segments:

(8) *"[ohne$\sim$ Hass– [auf deine Peiniger–]–]+"*

Finally, the overall polarity is assigned, which in this case is positive.

Tag	Marker	#Words	Examples	#Top Phrases	#All Phrases
positive	+	335	hope	158	275
negative	–	362	doubt	180	300
intensifier	∧	63	heavy	n.a.	n.a.
diminisher	%	9	low	n.a.	n.a.
shifter	∼	51	against	n.a.	n.a.
bipolar	#	n.a.	n.a.	21	54
neutral	0	n.a.	n.a.	10	12

Table 3: Distribution of the polarity tags in Layer 2.

Another example, following the extact same steps, takes as input the phrase:

(9) *"keine Angst vor dem schrecklichen Phantom"*
["no fear for the horrible phantom"]

and outputs the following annotation with an overall positive polarity:

(10) *"[keine∼ Angst– [vor dem schrecklichen Phantom–]–]+"*

Table 3 provides some descriptive statistics regarding the annotations produced on Layer 2. The Top Phrases column contains the counts for phrases that stand directly below the sentence-level, i.e. if such a phrase was to be composed into a higher level textual unit, that unit would be the sentence at hand. In a similar way, the All Phrases column contains the counts for all possible phrases below the sentence-level that have been annotated with polarity, including the top phrases. As a first general remark we can observe a slight tendency for negativity in our dataset, both on word- and phrase-level, while neutrality is observed seldom. Secondly, we can see that primary examples of compositionality, like the intensification and shifting phenomena also have a significant presence in our dataset. Finally, coming back to neutrality, although it was observed less frequently, we can see how a number of phrases have in fact been assigned an overall neutral polarity although they contain polar words and/or phrases. For example the phrase:

(11) *"Trotz dieser erheblichen Steigerung der absoluten Zahlen"*
["Despite this considerable increase of absolute numbers"]

is assigned an overall neutral polarity despite the presence of shifters and positive words:

(12) *"[Trotz∼ dieser erheblichen+ Steigerung+ der absoluten Zahlen]"*

which provides us with an example where compositionality does not always break through to the top level. In other words, a phrase's overall polarity will not necessarily always be positive, negative or bipolar, although it contains polarized constituents.

	Merged	Annotator 1	Annotator 2
DSE	656	642	638
ESE	734	692	713
OSE	7	7	6

Table 4: Major annotation frame types in Layer 3.

	Merged	Annotator 1	Annotator 2
Source	261	254	249
Target	1124	1053	1074
Operator	60	54	58
Modulation	160	147	155
Polarity	23	23	18
Support	130	126	127

Table 5: Major frame label categories in Layer 3.

3.3. Layer 3: Expression-level Annotations

The annotation scheme of Layer 3 adheres to the main concepts of expression-level annotation of the MPQA corpus (Wiebe et al., 2005). This type of annotation is important for building systems for sentiment-related information extraction tasks, such as opinion summarization or opinion question answering (Stoyanov et al., 2005; Stoyanov and Cardie, 2011). In those tasks, the sentiment towards a specific entity, e.g. a person, an organization or a commercial product, is to be extracted. Sentiment annotation on the sentence-level (Layer 1) or on complex phrases (Layer 2) are less helpful for such applications.
We annotate lexical units denoting frames of private states, i.e. states that are not open to observation and verification and their corresponding frame elements. We distinguish between the three types, *Objective Speech Events (OSEs)*, such as sentence (13), *Direct Speech Events (DSEs)*, such as sentence (14), and *Explicit Subjective Expressions (ESEs)*, such as sentence (15). The latter are used by speakers to express their frustration, wonder, positive sentiment, mirth, etc., without explicitly stating that they are frustrated, etc. (Wiebe et al., 2005).

(13) *"Peter [sagte]OSE, dass es regnete."*
["Peter [said]OSE it was raining."]

(14) *"Peter [schimpfte]DSE über das Wetter."*
["Peter [complained]DSE about the weather."]

	Layer 1	Layer 2	Layer 3
Annotators	3	3	2
Items used for calculation	270 sentences	133 words, 98 phrases	130 events and expressions
Fleiss' Kappa (Average Pairwise Agreement)	Sentence-level subjectivity: 0.721 (87.2%)	Word-level polarity: 0.685 (76.9%)	DSEs, OSEs, ESEs: 0.667 (80.8%)
Fleiss' Kappa (Average Pairwise Agreement)	Sentence-level polarity: 0.765 (84.6%)	Phrase-level polarity: 0.808 (88.4%)	Expression-level polarity: 0.897 (93.8%)

Table 6: Inter-annotator agreements for Layer 1, 2 and 3.

(15) *"Peter trägt eine [furchtbare]ESE Jacke."*
["Peter wears a [terrible]ESE jacket."]

Each frame can be assigned optional frame flags. The flag inventory consists of the *prior polarity* of a frame (i.e. positive, negative, or both) and a label denoting *backgrounded* sentiment. Lexical units conveying such a sentiment entail sentiment information but their primary meaning conveys something else. For example, the verb "ermorden" ("to murder") means "to kill another being" but this usually entails that the perpetrator has a negative sentiment towards its victim.
Typical frame elements are the *source* and the *target* of a frame, *modulation* (i.e. diminishers and intensifiers) and *operator* by which context modification such as negation or modal embedding is captured.

(16) *"[Peter]source [schimpft]DSE [nicht]operator [viel]modulation [über das Wetter]target."*
["[Peter]source does [not]operator [complain]DSE [much]modulation [about the weather]target."]

Another element called *polarity* denotes markers that indicate the polarity towards the target. Note that this is different from the polarity frame flag which indicates the *prior* polarity of the lexical unit evoking the pertaining frame. For example, the verb "criticize" evokes a DSE with a negative polarity frame flag. The noun "Kampagne" ("campaign"), by contrast, evokes a DSE without a polarity flag since "Kampagne" is underspecified for polarity towards its target. Its source can, in principle, have either positive or negative polarity towards the target. Prepositional markers that appear on the dependents of such a predicate, for example "für/gegen" ("for/against") in "Kampagne für/gegen höhere Steuern" ("campaign for/against higher taxes"), are considered a marker indicating the contextual polarity towards the target (as it has not been specified by the target itself). Those markers are assigned the polarity frame element.
Some important descriptive statistics of the annotations on Layer 3 are given in Tables 4 and 5, which represent the counts for each individual annotator as well as of the adjudicated version. As can be seen from Table 4, we have very few instances of OSEs in our data. One important reason for this is that, unlike in the MPQA, we did not annotate frames for the top-level writer's speech event because it is always unexpressed and there is no syntactic predicate for us to target. As Table 5 shows, we have far fewer Source elements annotated than we do Targets. This has two reasons. First, the former often correspond to the implicit writer of the text and thus are not available for annotation. Second, we have a relatively high number of ESEs among the subjective frame types: ESEs by definition cannot realize Sources as syntactic dependents. Another interesting observation (not spelled out in either table) is that specifications of Polarity, though rare overall, are more common with DSEs: only two cases occur with ESEs. The most common type of Polarity element is an adjective such as *positive* or *negative* modifying a noun DSE, as in "negative Reaktionen der Mitmenschen" ("negative reactions by others").

4. Inter-annotator Agreements

In order to measure the reliability of our annotations, we computed inter-annotator agreements by means of two measures for all layers: average pairwise agreement and (Fleiss, 1981)'s multi-rater Kappa. Calculations are based on all sentences for Layer 1 and on a 30 sentence test set for Layer 2 and Layer 3 (cf. Table 6). On all three layers we reached at least "substantial agreement", for phrase-level polarity and expression-level polarity even "almost perfect agreement" (Landis and Koch, 1977).

5. Conclusions

In this paper, we described the creation of MLSA, a multi-layered reference corpus for German sentiment analysis. The corpus contains sentences annotated on sentence-level, word- and phrase-level and expression-level. Due to its multiple layers, it is applicable to various sentiment analysis approaches. Used as a gold standard, such a corpus facilitates comparability and reproducibility.
Moreover, it frees the researcher from the burden to collect and annotate data by themselves. Thus, we believe that establishing our corpus as a standard resource in German-language sentiment analysis will be beneficial for the research field.

6. Acknowledgements

We gratefully acknowledge financial support of the German Research Foundation (DFG) through the EC 277 *Cognitive Interaction Technology* at Bielefeld University, the German Federal Ministry of Education and Research (BMBF) under grant no. "01IC10S01", and of the Swiss National Science Foundation (grant 100015_122546/1).

7. References

A. Balahur and R. Steinberger. 2009. Rethinking sentiment analysis in the news: from theory to practice and back. In *Proceeding of the 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA)*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447.

Sabine Brants and Silvia Hansen. 2002. Developments in the tiger annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation*, pages 1643–1649, Las Palmas.

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

Hoa Trang Dang. 2009. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, MD, USA.

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

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics, July.

Robert Remus and Christian Hänig. 2011. Towards Well-grounded Phrase-level Polarity Analysis. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, number 6608 in LNCS, pages 380–392. Springer.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, , and Noriko Kando. 2010. Overview of Multilingual Opinion Analysis Task at NTCIR-8 - A Step Toward Cross Lingual Opinion Analysis. In *Proceedings of NTCIR-8 Workshop Meeting*.

Veselin Stoyanov and Claire Cardie. 2011. Automatically creating general-purpose opinion summaries from text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 202–209, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the opqa corpus. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 923–930, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden, July. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language ann. *Language Resources and Evaluation*, 39(2/3):164–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.