PACE Corpus: a multilingual corpus of Polarity-annotated textual data from the domains Automotive and CEllphone

Christian Hänig
ExB Group
Seeburgstr. 100
04103 Leipzig, Germany
haenig@exb.de

Andreas Niekler
University of Leipzig
Augustusplatz 10
04109 Leipzig, Germany
aniekler@informatik.uni-leipzig.de

Carsten Wünsch
University of Bamberg
An der Weberei 5
96045 Bamberg, Germany
carsten.wuensch@uni-bamberg.de

Abstract

In this paper, we describe a publicly available multilingual evaluation corpus for phrase-level Sentiment Analysis that can be used to evaluate real world applications in an industrial context. This corpus contains data from English and German Internet forums (1000 posts each) focusing on the automotive domain. The major topic of the corpus is connecting and using cellphones to/in cars. The presented corpus contains different types of annotations: objects (e.g. *my car*, *my new cellphone*), features (e.g. *address book*, *sound quality*) and phrase-level polarities (e.g. *the best possible automobile*, *big problem*). Each of the posts has been annotated by at least four different annotators – these annotations are retained in their original form. The reliability of the annotations is evaluated by inter-annotator agreement scores. Besides the corpus data and format, we provide comprehensive corpus statistics. This corpus is one of the first lexical resources focusing on real world applications that analyze the voice of the customer which is crucial for various industrial use cases.

Keywords: Phrase-level Sentiment Analysis, Automotive Corpus, Lexical Resource

1. Introduction

One of the most rapidly growing areas in Natural Language Processing is Sentiment Analysis. With the increasing amount of available textual data in the Internet and especially the Web 2.0, the analysis of subjectivity, sentiments and opinions gained a lot of attention.

The field can be divided into two main challenges: developing algorithms and creating resources (e.g. corpora and dictionaries). Most of the research focuses on English in both fields. While recent algorithms concentrate on more detailed (e.g. the extension from document-level classification to sentence- and phrase-level classification) and comprehensive (e.g. disambiguation of objective and subjective utterances) analyses, only little effort is expended to create non-English resources for Sentiment Analysis (e.g. (Remus and Hänig, 2011)).

Still, many researchers create their own data sets for evaluation because standardized evaluation corpora have not been established, yet. For example, Pang et al. (2002) annotated movie reviews, Hu and Liu (2004) used product reviews to automatically create document level polarity scores and Hoffmann (2005) annotated newspaper texts. Consequently, results of different approaches are not comparable to each other.

Recent comparisons of data from newspapers, Internet forums and domain specific data prove that language models trained on clean newspaper data do not perform well on user-generated data (see (Schierle, 2011)) and consequently, a shift from using newspaper data to using more general web data can be observed (e.g. (Clematide et al., 2012)). Additionally, countless efforts have been undertaken to apply NLP methods to user-generated data (e.g. Twitter, see (Benhardus and Kalita, 2013)), but evaluation corpora for this kind of data is still missing for non-English Sentiment Analysis although it is essential for industrial applications to analyze data from the Web 2.0.

In this paper, we want to contribute a multilingual corpus consisting of user-generated data from Internet forums dealing with automotive issues. This kind of data is relevant for industrial analyses of user's opinions (e.g. quality assurance / perception analyses, see (Bank and Hänig, 2011)). The corpus contains polarity annotations on phrase level along with the target for each expression. The corpus also contains the annotations of four different annotators per text to reflect the humans true perception of polar utterances as good as possible.

2. Related Work

A comprehensive collection of manually (or semi-automatically) created resources for sentiment analysis exists. This includes amongst others: lexicons of polar words (see (Esuli and Sebastiani, 2006)), appraisals (see (Argamon et al., 2009)) and corpora annotated on different linguistic levels (e.g. on phrase-level as in (Agarwal et al., 2009)). An overwhelming part of these resources are available for English.

For German sentiment analysis only few resources exist, e.g. lexicons containing polarity information of words (e.g. (Remus et al., 2010)). Remus and Hänig (2011) published an evaluation corpus which contains polarity annotations on token-, phrase- and sentence-level (477 sentences extracted from German forum entries), (Clematide et al., 2012) augmented an excerpt (270 sentences) of a general-purpose web corpus (see (Baroni et al., 2009)) with polarity on multiple linguistic levels.

Thus, non-English resources for sentiment-related research lack in variety and quantity. Moreover, resources focusing on real world applications barely exist, not even for English. Hence, we provide a corpus that reduces the deficiency in both fields.

3. Data Selection

The posts in this corpus were selected randomly from different Internet forums dealing with automotive issues. Only

posts writing about German premium manufacturers (in alphabetical order: Audi, BMW and Daimler) were selected using a comprehensive database of the respective model names. Further manufacturers are also mentioned, but they do not represent the focus. The selected posts were filtered to ensure that each post contains at least one sentence with the topic telephony in cars. For this purpose, a comprehensive knowledge base as described by Schierle and Trabold (2010) was created containing numerous terms of the cellphone and automotive domain.

4. Corpus Annotation

The corpus was annotated with numerous sentiment annotations. Each annotation basically consists of two different parts – the polarity value and the target object. The polarity value contains the sentiment phrase and its associated value (see Section 4.4.). The opinion target may be an object (see Section 4.5.), a feature (see Section 4.6.) or a combination of these two types (see Section 4.7.). Section 4.2. describes the annotation process in detail.

4.1. Multi-Annotator Approach

Sentiment analysis is not an easy task, not even for humans. Taking this into account, we assigned four different annotators to each post. We introduced the task to the annotators and trained them how to use the graphical user interface (see Section 4.3.) created for the annotation process during training sessions. The annotation guidelines were constituted in a codebook. These guidelines were designed to rise the reliability of the annotations and to formulate a non-biased common knowledge through all the annotators. This annotation process led to up to four annotations per sentiment phrase. We decided to keep all annotations in the data to represent human's perception of sentiment as accurate as possible. Following this approach, the corpus can be employed for a variety of evaluation modes, such as evaluation on:

All annotations Using this mode, it is not possible to achieve 100% F-Score, but this is the annotation distribution representing exactly human perception.

Annotations annotated by n $\geq$ 2 annotators This mode filters utterances that only a portion of all annotators perceived as being polar.

Annotations with unambiguous polarity values This evaluation mode includes only test cases where all four annotators agree. This filter is very restrictive and returns the cases that are easy to decide for humans.

4.2. Annotation Process

The production of coherent annotations requires a mutual understanding of the objects, features, opinion targets and the sentiment polarity values, e.g. the category system of the annotation task. For this reason it is important to define and maintain a codebook to which the annotators can refer to (Krippendorff, 2004).

A list of identifiable objects and features was used to create an initial codebook. This version of the codebook was the basis of a training session with the annotators to ensure appropriate understanding of the annotation guidelines. We encouraged the annotators to produce test annotations which were discussed within a training session afterwards. This was a crucial part of the annotation process in order to ensure understanding of the requirements. The whole annotation process was supervised by a content analysis scholar who instructed and trained the assistants.

Following the training phase we started a pretest to evaluate the initial accuracy and agreement of the annotators' produced annotations according to the codebook. Working with real data showed that there had to be alignments regarding the definition of our objects and features. Furthermore we had to notice that our list of objects and features was incomplete and that we had to update the codebook. We decided to keep the possibility of codebook updates to prevent incomplete annotations within the documents.

In some cases a new object or feature type describes a relation better. Each annotator was given the possibility to suggest new objects and features directly to the instructor. The annotations were reviewed by the instructor and the codebook got an update. If the suggestion could not be accepted, the annotator was instructed to revisit the document and alter the annotation. The annotations were produced redundantly resulting in four independent judgments on each document. The annotation process is shown in Figure 1.

After the pretest we moved on to the generation of the actual annotations. We left the pretest annotations within the corpus. They can be excluded by omitting the first 20 documents for each language.

4.3. Annotation Tool

We created a tool consisting of a data management layer and a graphical user interface to produce the annotations. Within this tool we assigned documents to each of the annotators. The assignment was distributed through the whole collection in order to guarantee the four-fold annotation of each document. We also included the possibility to mark documents in case an annotation needs supervision by the instructor. To ease the annotation process we also included a list of already processed documents and unprocessed documents. The annotators were able to annotate objects, features, opinion targets and sentiment polarity values directly in the text. Words in the text could be assigned to an object or feature category and its related phrases. Figure 2 shows the web-based surface of the annotation tool.

4.4. Sentiment Polarity Value

All phrases are annotated by one of five polarity scores ranging from *very bad* and *bad* over *neutral* to *good* and *very good*.

There are two additional categories for implicit polarity utterances: *implicitely bad* and *implicitely good* as there are indirect opinions and sentiments uttered indirectly sometimes (e.g. *Oh my! What a vehicle.*).

4.5. Objects

Objects are the items being talked about in the text. The texts are selected from Internet forums issuing automotive topics which are reflected by the predefined object categories. Three main objects were defined: cars (*Car*), cell-

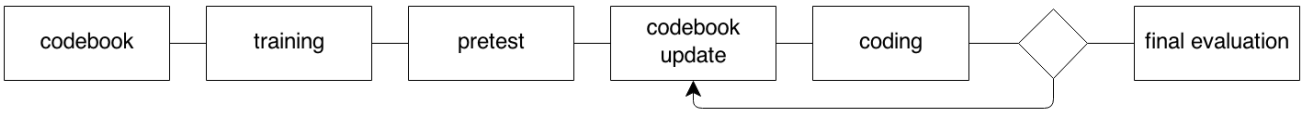

Figure 1: Schema of annotation process.

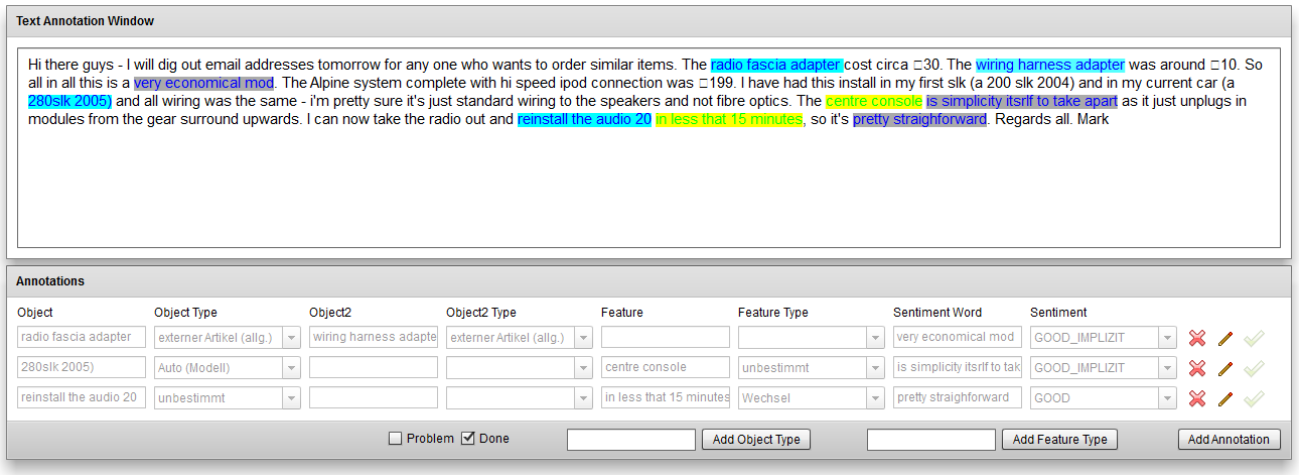

Figure 2: This GUI was used during the annotation process.

phones (*Cellphone*) and MP3 player (*MP3 player*). For each object, three different subtypes are distinguished:

- *Manufacturers* produce this kind of object and are denoted by *(P)*, e.g. *Car (P)* is used to annotate car manufacturers like *Audi*, *BMW*, *Mercedes*, ...
- *Models* (denoted by *(M)*) mark specific models or model types of this object type, e.g. *Car (M)* is used to annotate car models/series like *A8*, *3er* or *S Class*.
- *General* mentions of this object type are marked by *(G)*, e.g. general mentions of cars as in *my car* are annotated by *Car (G)*.

Additionally to these three main object types, three more generalized object types exist:

- *Manufacturer* is used to annotate all other occurrences of manufacturers not covered by the three types above.
- *Model* denotes all other models not being car, cellphone or MP3 player models.
- *General* marks all general mentions of any part / object that is relevant for understanding the sentiment utterance.

The last object type is *Unknown*. In some cases, even humans cannot define the type of each object in a post without the context of the complete forum discussion. In these rare cases, the identifiable objects are annotated as *Unknown*.

4.6. Features

The predefined list of features was created with support of engineers being responsible for integration of convenient telephony and sound solutions into cars. The most frequent feature types were used to construct this corpus (see Section 3.). Thus, the final list contains the most relevant features for connecting cellphones / MP3 players to cars:

adapter / wires	exchange
address book	handling
call quality	hands-free
compatibility	pairing
concept	power
conference call	quality
consumption	service
cost	sound quality
delivery times	unknown
design	unspecified
display	website
documentation	

Table 1: Predefined Features

For all other features not covered by this list, the feature type *Unknown* was used.

4.7. Opinion Targets

Opinion targets may be simple objects, features, or combinations of both types.
All valid possibilities are:

Object The sentiment phrase targets an object.
E.g.: *A* $\underline{\textit{C-class}}_{Car(M)}$ *is a* $\underline{\textit{pretty good car}}_{+}$.

Feature The sentiment phrase targets a feature.
E.g.: *You'll get* $\underline{\textit{more}}_{+}$ $\underline{\textit{gas mileage}}_{Power}$.

Object / Feature A specified feature of a specified object is the target.
E.g.: *Their* $\underline{\textit{designs}}_{Design}$ *are* $\underline{\textit{a little quirky}}_{-}$ *anyway.*

Object / Object The combination of a car with a cellphone/MP3 player is the opinion target. It can also be used to compare different objects.
E.g.: *The $\underline{E}_{Car(M)}$ still isn't an $\underline{LS}_{Car(M)}$.*

Object / Object / Feature The sentiment phrase targets a feature of a combination of car and cellphone / MP3 player.
E.g.: *$\underline{Motorola}_{Cellphone(P)}$ seems to $\underline{work}_{Compatibility}$ the $\underline{best}_{+}$ with $\underline{Mercedes}_{Car(P)}$.*

5. Corpus Statistics and Evaluation

The *PACE* corpus contains numerous objects, features and polar utterances. We describe the extent of the annotations comprehensively in Section 5.1. and present agreement scores of the annotations in Section 5.2..

5.1. Corpus size

The contained number of forum posts, tokens and annotations is provided in Table 2. The selected English forum posts are longer than the German ones on average.

	English	German
# of posts	1000	1000
# of annotations	6.298	5.196
# of tokens	144.948	101.740

Table 2: Corpus size information

As can be seen in Table 3, only 2.9% of all annotations were classified as neutral. Obviously, only people that experienced problems with their cars and/or phones start a discussion in a forum. A similar observation can be made for responders: only persons that have any kind of experience will reply. Either they share the statements of the discussion starter or they report contradicting experiences. In all of these cases, only few posts will be neutral and thus, contain neutral phrases.
Regarding polar annotations, the annotators tend to barely use extreme polarity scores. This effect is called *central tendency* in empirical psychology and occurs whenever humans have to select an answer for indifferent questions or they lack information about the object to be classified (see (Hollingworth, 1910)). In these cases, humans tend to select values around the *central point* (e.g. the median, see (Howell, 2012)) which often is presented in the middle of the scale. Since there are no guidelines when to annotate a phrase as very positive or negative and people perceive such utterances differently, annotation becomes a highly subjective task. While for some people the phrase *My car does not start* is negative, for others it is very negative, because they think of worst case scenarios that could follow (e.g. to miss a flight or an important meeting).
Another interesting fact is that positive phrases are expressed more directly in the corpus than negative ones. Almost 45% of all negative annotations were annotated as implicit polarity, while only 25% of all positive annotations do not contain overt polarity markers. A reason might be the problem description of forum users. Problems lead to negative sentiment, but phrases like *but i can't (or don't know how to) open it* do not contain reliable sentiment-bearing clues that could stand by themselves. In many cases, the complete forum post is necessary for contextual analysis of the phrase's polarity.

	English	German
very positive	547	493
positive	2128	1316
neutral	111	224
negative	1214	1355
very negative	146	151
implicitly positive	982	507
implicitly negative	1170	1150

Table 3: Distribution of polarity annotations

Table 4 shows the distribution of object types. It is obvious that people writing in forums prefer to name the exact model of their cars / cellphones in order to get adequate support from the community. This fact is reflected by the document selection in this corpus – *Car (M)* and *Cellphone (M)* are much more frequent than their respective general versions *(P)* and *(G)*.

	English	German
Car (G)	347	146
Car (M)	2075	1332
Car (P)	588	965
Cellphone (G)	214	185
Cellphone (M)	1425	1942
Cellphone (P)	227	478
MP3 Player (G)	4	15
MP3 Player (P)	132	102
General	607	525
Manufacturer	397	327
Model	980	374
Car accessories		76
unknown		7
unspecified	492	453

Table 4: Distribution of object types

The feature type distribution is outlined in Table 5. This distribution is very fine-grained and a lot of confusions between some feature pairs could occur (e.g. *pairing* and *compatibility*). It is very important for engineers to distinguish different feature types in order to analyze the error as precise as possible. Thus, we decided to not merge certain partially overlapping categories after fruitful discussions with responsible engineers.

5.2. Inter-Annotator Agreement scores

We report scores for various measures to evaluate the annotations as comprehensive as possible. All agreement scores of a document are based on the document's average polarity. This approximation was made to circumvent matching of ambiguous annotations. The boundaries of the annotators' annotations do not always match. There were cases of exact matches and cases with only few overlapping tokens. Sometimes the perception of sentiment was completely dif-

	English	German
adapter / wires	212	247
address book	282	278
call quality	122	151
compatibility	383	576
concept		5
conference call	3	6
consumption		9
cost	206	142
delivery times	38	75
design	299	165
display	94	70
documentation	16	15
exchange	40	15
handling		2
hands-free	138	213
pairing	380	371
power	213	105
quality	466	365
service	229	224
sound quality	181	120
website	42	25
unknown	526	402
unspecified	20	112

Table 5: Distribution of feature types

ferent between two annotators leading to less than four annotations for a particular expression.

To ease this problem we simplified the evaluation according to the assumption that similar annotations should lead to a similar average polarity value throughout the four annotators for the same document. Therefore, we evaluated the agreement on the average sentiment polarity value $\bar{s}_{d,a}$ of a document d edited by annotator a. We used the ordinal scale of the polarity values and transformed them to numbers (1 - very positive, 5 - very negative) to calculate the average value for each annotator in a document. Following our assumption we obtained four different average values $\bar{s}_{d,a}$ for each document. Based on these scores we calculated annotator agreements with the following measures:

- **%-agree:** The measure represents the overlap of similar average annotations in all documents. We assume similarity of two annotator's average polarity scores $\bar{s}_{d,a}$ if $\|\bar{s}_{d,a_1} - \bar{s}_{d,a_2}\| \leq 1$.

- **Kendall's W (W):** All average ratings $\bar{s}_{d,a}$ of an annotator could be transferred into a ranked list. This coefficient measures the agreement by comparing the rankings for each annotator.

- **Krippendorf's α (α):** This measure is the ratio between observed disagreement and expected disagreement (Artstein and Poesio, 2008).

- **Standard Deviation $\sigma(\bar{s}_d)$:** We calculate the standard deviation $\sigma(\bar{s}_d)$ for all annotators of a document to estimate the agreement of the annotators. The overall standard deviation for all annotations and documents within the corpus could be defined as the pooled variances.

We calculated agreement scores for the annotations of both languages combined and separately. Furthermore we also applied the measures to the most agreeing annotators. For this purpose a pairwise / triplewise matching of the average values $\bar{s}_{d,a}$ for each document was done ignoring all but the two or three annotators achieving the closest average polarity value. This additional evaluation was performed to minimize effects of outliers or incorrect annotations. The results are shown in Table 6.

Corpus	$\sigma(\bar{s}_d)$	%-agree	W	α
English$_{all}$	0.38	68.0	0.75	0.68
German$_{all}$	0.40	67.4	0.79	0.71
Complete$_{all}$	0.39	67.7	0.78	0.71
English$_3$	0.28	89.0	0.85	0.77
German$_3$	0.27	94.2	0.88	0.81
Complete$_3$	0.27	91.7	0.88	0.80
English$_2$	0.13	97.7	0.93	0.89
German$_2$	0.11	98.1	0.93	0.90
Complete$_2$	0.12	97.9	0.94	0.90

Table 6: Summarization of the Evaluation results. The results marked with $_2$ / $_3$ were calculated with the annotations of the two / three most agreeing annotators.

6. Conclusions

We presented a comprehensive multi-lingual corpus for evaluation of phrase-level sentiment analysis. The *PACE* corpus consists of real-world user-generated data. It is essential to process this kind of data in authentical quality assurance tasks of the automotive industry.

The steps that were performed to create this gold standard sentiment corpus are described in detail. They included data selection, creation of annotation guidelines and the iterative annotation process including revisions of the codebook. We also described the applied quality preserving strategies of our process.

When taking all annotations into account the agreement scores for W, α and %-agree are lower than 0.8. According to Krippendorff (2004) a value of $\alpha < 0.8$ does not imply a substantial agreement of the annotations. This indicates the complexity of such an annotation task. Especially when annotation of relations between multiple objects / features and associated polarity scores is required, it is very likely that different annotators will not agree on at least one of the parts.

However, agreement scores computed for the most-agreeing pair or triple of annotators certify strong agreement. This may result from the intense work with the annotation guidelines, the revisions of the codebook and the initial training sessions. Other studies report moderate agreements for a similar task of annotating polarity scores on phrase level (see (Remus and Hänig, 2011)).

Finally, we want to encourage other researchers to evaluate their algorithms on real-world and user-generated data to narrow the gap between research and industrial applications.

7. References

Agarwal, A., Biadsy, F., and Mckeown, K. R. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32. Association for Computational Linguistics.

Argamon, S., Bloom, K., Esuli, A., and Sebastiani, F. (2009). Automatically determining attitude type and force for sentiment analysis. In *Human Language Technology. Challenges of the Information Society*, pages 218–231. Springer.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Bank, M. and Hänig, C. (2011). Using the Internet as Sensor for Customer Perception. In *Proceedings of the Fachtagung zum Text- und Data Mining für die Qualitätsanalyse in der Automobilindustrie*, pages 49–55, Leipzig, Germany.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Benhardus, J. and Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., and Wiegand, M. (2012). MLSA A Multi-layered Reference Corpus for German Sentiment Analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3551–3556.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Hoffmann, P. (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*, pages 347–354.

Hollingworth, H. L. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461–469.

Howell, D. (2012). *Statistical Methods for Psychology*. PSY 613 Qualitative Research and Analysis in Psychology Series. Cengage Learning.

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA.

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Sage, Thousand Oaks Calif., 2nd ed. edition.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86.

Remus, R. and Hänig, C. (2011). Towards Well-grounded Phrase-level Polarity Analysis. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 380–392, Tokyo, Japan. Springer.

Remus, R., Quasthoff, U., and Heyer, G. (2010). Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC)*, pages 1168–1171.

Schierle, M. and Trabold, D. (2010). Multilingual Knowledge-Based Concept Recognition in Textual Data. In *Advances in Data Analysis, Data Handling and Business Intelligence*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 327–336. Springer Berlin Heidelberg, Berlin, Heidelberg.

Schierle, M. (2011). *Language Engineering for Information Extraction*. Phd, University of Leipzig.