

Detecting Vocal Irony

*Felix Burkhardt*¹, *Benjamin Weiss*², *Florian Eyben*³, *Jun Deng*³ and *Björn Schuller*³

¹Deutsche Telekom AG, Berlin, Germany

²Technische Universität Berlin, Germany

³audEERING GmbH, Gilching, Germany

felix.burkhardt@telekom.de

Abstract

We describe a data collection for vocal expression of ironic utterances and anger based on an Android app that was specifically developed for this study. The main aim of the investigation is to find evidence for a non-verbal expression of irony. A data set of 937 utterances was collected and labeled by six listeners for irony and anger. The automatically recognized textual content was labeled for sentiment. We report on experiments to classify ironic utterances based on sentiment and tone-of-voice. Baseline results show that an ironic voice can be detected automatically solely based on acoustic features in 69.3 UAR (unweighted average recall) and anger with 64.1 UAR. The performance drops by about 4 % when it is calculated with a leave-on-speaker-out cross validation.

1. Introduction

Verbal irony occurs when someone says something that is obviously not expressing the real intention or meaning; sometimes it is even the opposite. This is usually achieved by gross exaggeration, understatement or sarcasm. The speakers rely on the receiver's knowledge to decode the real meaning, usually because the context of the semantics does not fit (example: saying "beautiful weather" when it starts to rain) or the utterance contains contrasting polarities, as for example in "slow internet is exactly what I need."

Beneath the semantic contrast, this can be also be achieved by contrasting the "tone of voice" with the sentiment of the words. According to the Relevance Theory (Wilson and Sperber, 2012), irony in speech can be considered as an attitude towards the statement, consequently modifying its meaning – for example a verbally positive statement realized with a prosody that indicates a negative attitude. An open question as of yet is whether a purely acoustic

form of irony exists, i.e., an "ironic tone of voice" irrespective of the words being uttered. Studies on this have not been very promising, indicating that prosody alone is insufficient to signal irony (Bryant and Fox Tree, 2005). Still, whereas a "dry" version does not reveal this ironic attitude in prosody, "dripping" irony does differ from sincere utterances in prosody (Bryant and Fox Tree, 2005).

Single target words exhibit acoustic differences between ironic from literal meaning in female German. The ironic versions were lower in fundamental frequency (F0), had longer and hyper-articulated vowels, and more energy (Scharrer et al., 2011). A similar study for French females revealed an expanded F0 range with higher average F0, syllable lengthening, and a raised F0 contour instead of a falling one as discriminating ironic from sincere versions in target words (González-Fuente et al., 2016). A subsequent re-synthesis experiment on the same target words confirmed the effects for all three features combined, for lengthening only, and pitch contour only. For English, particularly a lower average F0, but also lower F0 range, longer durations and higher intensity are found in ironic conditions (Cheang and Pell, 2008; Rockwell, 2000). Directly contrasting adjunct sentences, ironic utterances were slower (Bryant, 2010).

This incongruence in affective prosody with the valence of the semantics (Bryant and Fox Tree, 2002; Woodland and Voyer, 2011) can be even observed in fMRI data (Matsui et al., 2016). From our daily experience probably most of us will tend to agree that this discrepancy between content and prosodic attitude can be easily detected – for the "dripping" case of intentionally signaling this ironic attitude at least. Therefore, an automatic irony detector should be feasible to develop based on sufficient training data, but to our knowledge no such data has been col-

lected and investigated with the goal of automatic classification as of yet.

Irony detection is obviously an important step in human-machine communication as it can reverse the meaning of an utterance. One use-case for example would be the automatic sentiment analysis of vocal social media data for market research, another use-case would be to enhance the semantic processing of robots and software agents with speech interfaces.

This paper is structured as follows. Section 2 describes the Android app that was used to collect data on ironic utterances. Section 3 explains the acoustic classification process, whereas Section 4 does the same for the textual sentiment categorization. In Section 5, the data collection process is described, and in Section 6 the way the data was labeled for emotional expression as well as textual sentiment. It is followed by Section 7, which reports on some experiments we performed. The paper concludes with a summary and outlook in Section 8.

2. The Irony Simulation App

For data collection, an Android app was created which is capable of recording audio, streaming it to audEERING's sensAI API service for analysis of emotional parameters. The app provides various pictures on which the users can comment on to evoke angry or ironic responses. However, these pictures only served as a first stimulation, but users need not to adhere to the pictures. They were free to test the app with any neutral, emotional, ironic, or angry comment they would come up with. Likewise, although the chosen image was logged by the app, it was not used any further.

The main focus of the study was on the detection of irony. The detection of anger was added to compare findings with earlier studies that focused on angry speech detection (Burkhardt et al., 2009).

The app displays the result to user immediately after processing. For the preliminary irony detection module, a text transcription based on the Google Speech Services API and a ternary text sentiment classification was added to audEERING's sensAI API. Further, for the purpose of collecting a spoken irony database, all requests to the API made through this app were saved on the back end in order to build the database from these recordings.

In order to get an estimate on the individual usage



Figure 1: Main screen of the irony data collection app, while a result is being shown.

behavior, the Android ID of the client is submitted and stored on the server. The owners name, however, is not collected. Of course we cannot control that all utterances stemming from one Android device are made by the same person.

Figure 1 displays the main screen of the irony data collection app, while it is showing a recognition result. The results shown in the App include the text transcript, the level of emotional activation (*Aktivierung*), valence (*Valenz*), and anger (*Ärger*), as well as the textual transcription (first line) and the textual Sentiment (positive/neutral/negative). Binary classification results for irony and anger are shown through the buttons at the top. The green button indicates a negative result (no irony/anger), and a flashing red button would indicate a positive result (irony and/or anger). The irony detection in the data collection app is based on rules that detect a mismatch between positive textual sentiment and

rater #		1	2	3	4	5	6	AVG
EWE		0.89	0.68	0.70	0.82	0.87	0.73	0.81
1			0.52	0.59	0.66	0.77	0.67	0.73
2				0.44	0.48	0.45	0.42	0.57
3					0.54	0.59	0.56	0.63
4						0.68	0.58	0.68
5							0.65	0.72
6								0.66
Mean ρ	0.57							
Min ρ	0.42							
Max ρ	0.77							

Table 1: Pairwise cross correlation (Pearson correlation coefficient ρ) for the dimension irony. Mean, Min., Max. values exclude the Evaluator Weighted Estimator (mean) (EWE) values (first row).

rater #		1	2	3	4	5	6	AVG
EWE		0.75	0.78	0.76	0.79	0.73	0.76	0.80
1			0.45	0.59	0.52	0.59	0.58	0.64
2				0.60	0.57	0.56	0.67	0.66
3					0.56	0.59	0.58	0.67
4						0.60	0.57	0.66
5							0.57	0.66
6								0.68
Mean ρ	0.57							
Min ρ	0.45							
Max ρ	0.67							

Table 2: Pairwise cross correlation (Pearson correlation coefficient ρ) for the dimension anger. Mean, Min., Max. values exclude the Evaluator Weighted Estimator (mean) (EWE) values (first row).

negative acoustic valence or a very low acoustic activation alongside a positive sentiment. Binary anger detection is based on a threshold to the anger level, as well as thresholds to activation and valence.

The user can listen to his/her recording again by pressing the play button. Using the pencil icon on the right, the user can open a feedback dialogue, where irony and anger (binary yes/no labels) can be corrected and transmitted to the server. This feedback is optional to the user, however, most trial users were encouraged to make use of this functionality.

3. Acoustic Irony Classification

For transparency and reproducibility, here, we apply the widely used acoustic emotion classification framework to the irony database, which depends on the popular acoustic feature extraction tool openSMILE (Eyben et al., 2013) and Support Vector Ma-

4 energy related LLD	Group
RMS energy, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
55 spectral LLD	Group
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	spectral
6 voicing related LLD	Group
F_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice qual.
log. HNR, jitter (local and δ), shimmer (local)	Voice qual.

Table 3: ComParE acoustic feature set: 65 low-level descriptors (LLD).

chines (SVMs). Two audio feature sets, established in the field of computational paralinguistic, were used;

the extended Geneva Minimalistic Acoustic Parameter Set (eGEMAPS) and the larger-scale Interspeech 2013 Computational paralinguistic Challenge feature set (ComParE) (Schuller et al., 2013).

eGEMAPS is a *knowledge-driven* data set that exploits the first two statistical moments (*mean and coefficient of variation*) to capture the distribution of *low-level descriptors* (LLDs) describing spectral, cepstral, prosodic and voice quality information, creating an 88 dimensional acoustical representation of an utterance. It was specifically designed by a small group of experts to be a basic standard acoustic parameter set for voice analysis tasks including paralinguistic speech analysis. For full details the reader is referred to (Eyben et al., 2016). ComParE, on the other hand, is a large-scale brute forced acoustic feature set which contains 6 373 features representing prosodic, spectral, cepstral and voice quality LLDs. A detailed list of all LLDs for ComParE is given in Table 3. For full details on ComParE the reader is referred to (Eyben et al., 2013).

For classifiers, we use open-source implementations from Scikit-learn (Pedregosa et al., 2011), where SVMs with linear kernels are considered. We scale all features to zero mean and unit standard deviation by using the parameters derived from the training set. For the baseline, we introduce two types of cross validations: 5-fold cross validation (5-fold CV) and leave-one-speaker-out cross-validation (LOSO-CV). By that, it is desired that the results obtained are more representative.

4. Textual Sentiment Classification

The textual sentiment analysis is based on the GATE framework (Cunningham et al., 2011). The name of the grammar formalism used with GATE is Jape. The first steps involve splitting and tokenization. We added the Stanford Part-of-Speech Tagger (Toutanova and Manning, 2000) and a lemmatizer based on a German lexicon derived from the Morpho project (Lezius, 2000). A German polarity lexicon (Waltinger, 2010) is then used to identify tokens that carry positive or negative polarity. Further gazetteers annotate negation tokens (like “*not* great”) and strengthener words (like “*super* bad”). Because tokens can be both polarity words and negation or strengthener words, we use a Jape transducer to neutralize them when they occur in conjunction with a

polarity token. Another set of Jape rules reverses the polarity for polarity words that are negated (example: “not bad”) or used as strengtheners (example “super bad”). Finally, a polarity decision for the whole input sentence is computed by a comparison of the number of positive and negative tokens. The number of polarity tokens with the most frequent polarities, normalized by the number of all tokens, is being used as a confidence measure.

5. Data Collection

We conducted a workshop with lay people to gain experience on how the automatic recognition of sentiment will be perceived by users and also with the aim to collect some user data. Because we were not aware of any “ironic voice” acoustic data collection with the aim of automatic classification, we could not compute a model for irony and the app indicated irony when meeting a discrepancy between textual sentiment and vocally expressed valence, as described in Section 2.

During the workshop, 12 test users got introduced to the app and tried it out. There were nine male and three female participants, aged 25 to 48, average age 37 years, with 7.8 standard deviation, all native Germans currently living in Berlin. The testing subjects installed the app on their private Android phones and were instructed to use it for one week. After this time span, the server that the app needs to analyse the audio was shut down. They were compensated for their time and all of them signed agreement that the data will be recorded, stored and can be used and distributed for scientific research.

After the try-out period a set of 937 labeled samples had been collected, see Section 6 for details on the label process. The number of different Android device IDs is 21. There is 3910 seconds (about one hour) audio in total. The maximum, minimum, and average length are 21.84, 0.84, and 4.18 seconds respectively. The number of recognized words are at most 35 and at least 1 with average value 7.14 and standard deviation 4.75. Using a majority voting on the labelers results in 51.57 % of vocally ironic, 41.91 % angry and 22.15 % both ironic and angry utterances.

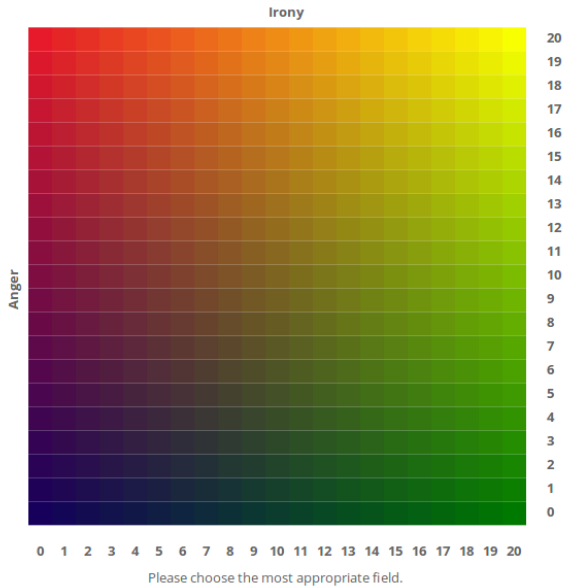


Figure 2: Two dimensional rating matrix used for human labeling of perceived irony and anger levels. Participants were prompted for their response with these instructions: *Please rate the level of anger and irony that you hear from the voice in the recording. Try to ignore the wording and listen only to the sound of the voice! If you are unsure, or no audio is audible, please choose the point (20, 20 → upper right corner) – Do not choose this point otherwise (use 19;19, 20;19, or 19;20 instead if you really need to rate a very high anger and high irony level).*

6. Data Labeling

Two sets of data labels are available in the database: The first is obtained directly from the users who contributed the speech data through the feedback functionality of the app; the second is obtained by manual rating from six labelers.

For rating, the crowd-sourcing platform iHEARu-Play (Hantke et al., 2015) was used. The platform was developed to allow raters to easily work from anywhere they feel comfortable through a web-based interface. No actual crowd-sourcing features which the platform provides (automatic rater trustability, dealing with raters who only rate parts of the database, etc.) were used here. Six student (psychology, linguistics, theatre) and professional (some of the authors of the paper) raters rated the whole database for angry and ironic “sound of voice”. They were instructed to ignore the textual content of the utter-

ances as much as possible and judge solely based on how the speech is expressed non-verbally. The judgements were given as a point in a two dimensional matrix with the dimensions “irony” and “anger”, with twenty points on each axis for quasi-continuous rating. Figure 2 shows the labeling interface in which the raters had to choose one point which reflected anger and irony levels.

In order to obtain, from the six individual ratings, a single gold standard for automatic classification, the Evaluator Weighted Estimator (EWE) is used (see (Grimm and Kroschel, 2005) and (Grimm et al., 2007)). The EWE is a weighted average of the individual ratings, where the weights are based on the average inter-rater reliability of each rater. All evaluations in this paper are based on the EWE.

To compute EWE the following steps are performed:

1. Normalisation of ratings of each rater to 0 mean, and maximum range (1..5)
2. Normal (arithmetic) averaging of all ratings for each segment
3. Computation of EWE weights (r) as Pearson correlation of each rater’s ratings with the average rating from b)
4. Normalisation of EWE weights to sum 1 and min and max. correlations to 0 and 1
5. Computation of final EWE average rating by weighted average using EWE weights

In Tables 1 and 2, the pairwise rater agreements and the agreement of each rater with the EWE (also including this rater) are displayed as cross correlation coefficients. Further, the mean, minimum, and maximum pairwise rater correlations are shown (excluding correlations of raters and EWE).

6.1. Textual data labeling

The textual data was labeled for sentiment by one of the authors with one of the three categories “neutral”, “positive” or “negative”. As stated in Section 2, the textual data is the result of the Google Speech API ASR. It was not manually corrected but used including the errors. An evaluation of the performance by computing WER or BLEU has not been done yet.

We found several situations where the decision for a sentiment was rather difficult. For example when two sentiments were given (example: “this is good but this is bad”), a question was raised (example: “is this bad?”), someone else’s opinion was given (example: “he says it’s bad”), contradictory statements were given (example: “I love it when it’s bad”) or complex statements (“it’d be better to spend the money on crime defense” being rated positive because of the word “better”, whereas “this I’d like to see every day” being rated as neutral because no polarity words are used).

7. Experiments

7.1. Irony and Anger expression

In our following experiments with acoustic irony and anger classification, we unified the EWE values from four raters by using majority voting and then mapping them to two classes. In particular, we use the average value of the unified EWE values to form the binary labels: for acoustic irony classification, if an unified EWE value for one utterance is larger than the average value, then its label is assigned as “irony”, otherwise as “non-irony”. Similarly, we assign the binary anger labels (i.e., “anger” or “non-anger”) for the collected database.

7.2. Sentiment analysis

We evaluated and tuned the textual sentiment analyzer using the data collection. The set of 915 unique sentences was split into 165 samples for a test set and 750 for training. Out-of-the-box, 109 out of the 165 test samples were correctly classified by the Sentiment analyzer (33.93 % error rate). We then went through the training set and corrected wrong classifier decisions by adding rules or editing the polarity lexicon and reached an error rate of 27.27 %. With larger data-sets it would make sense to use machine learning for classification.

7.3. Acoustic Emotion Classification Results

First, Table 4 shows the acoustic irony classification results for the LOSO-CV (leave-one-speaker-out cross validation) and 5-fold CV with different complexity parameters of the linear SVM. It can be seen clearly that ComParE outperforms eGEMAPS by a noticeable margin. Specifically, the baseline

Feature	C (SVM)	UAR	
		LOSO-CV	5-fold CV
ComParE	1e-4	61.4	69.3
ComParE	1e-3	61.3	69.6
ComParE	1e-2	60.6	67.9
ComParE	1e-1	60.3	67.9
ComParE	1.0	60.4	67.9
eGEMAPS	1e-4	48.9	53.4
eGEMAPS	1e-3	53.9	65.5
eGEMAPS	1e-2	54.6	65.6
eGEMAPS	1e-1	54.0	64.1
eGEMAPS	1.0	52.6	64.8

Table 4: Results in terms of Unweighted Average Recall (UAR, %) for the binary irony classification task with ComParE and eGEMAPS acoustic feature sets. LOSO-CV stands for the leave-one-speaker-out cross validation while 5-fold CV stands for the 5-fold cross validation. C indicates the complexity parameter for the linear SVM, which corresponds to the penalty parameter of the error term.

systems using eGEMAPS achieve promising UARs (Unweighted Average Recall) of 54.6 % and 65.6 % for the LOSO-CV and 5-fold CV evaluation schemes. In the meantime, the ComParE-based systems result in UARs of 61.4 % and 69.6 %, which is higher than the one obtained by the corresponding eGEMAPS systems. It is worth noting that the ComParE-based systems are more robust to the choice of the complexity of the linear SVM when compared to the eGEMAPS baseline systems for the acoustic irony classification.

Next, Table 5 presents the experimental results for the acoustic anger classification task, which was previously defined in Section 7.1. Based on Table 5, we can see that the baseline systems obtain representative performance in terms of UARs for both the LOSO-CV and 5-fold CV evaluation schemes. It is surprising for the LOSO-CV scheme that the linear SVM system using the eGEMAPS feature set of only 88 features reaches a UAR of 63.5 %, which is clearly higher than the one obtained by the ComParE feature set of 6 373 features. This suggests that the knowledge-driven feature set of eGEMAPS contain more informative information for the anger classification task in the LOSO situation. For the 5-fold CV scheme, the ComParE baseline system surpasses the eGEMAPS system, just like we found with the acoustic irony classification.

Feature	C (SVM)	UAR	
		LOSO-CV	5-fold CV
ComParE	1e-4	59.0	67.1
ComParE	1e-3	57.5	63.7
ComParE	1e-2	58.8	62.5
ComParE	1e-1	58.6	62.5
ComParE	1.0	58.6	62.5
eGEMAPS	1e-4	62.2	53.1
eGEMAPS	1e-3	63.5	63.8
eGEMAPS	1e-2	63.0	64.1
eGEMAPS	1e-1	61.5	63.3
eGEMAPS	1.0	57.6	61.8

Table 5: Results in terms of Unweighted Average Recall (UAR, %) for the binary anger classification task with ComParE and eGEMAPS acoustic feature sets.

8. Summary and Outlook

We described a data collection for vocal irony and anger expression based on an Android app that was specifically developed for this study. The main aim of the investigation was to find evidence for a non-verbal expression of irony. A data set of 937 utterances was collected and labeled by six listeners for irony and anger (for comparability with earlier studies). The automatically recognized textual content was labeled for sentiment. We conducted experiments to classify ironic utterances based on sentiment and tone-of-voice with machine learning. The results show that irony can be detected automatically solely based on acoustic features in 69.3 UAR (unweighted average recall) and anger with 64.1 UAR. The performance drops by about 4 % when it is calculated with a leave-one-speaker-out cross validation.

It is planned to make the collected data available to the research community to foster the investigation of ironic speech.

There are many ideas to work on this topics, including:

- Enhance the textual sentiment classifier by adding more rules.
- When sufficient data was collected, enhance the textual sentiment classifier by machine learning.
- Investigate the influence of ASR errors on the sentiment classification, for example by contrasting it with a manual transcription.
- Differentiate between categories of vocal and non-vocal expressions of irony and investigate

with which modalities are best suited to detect them.

- Add more modalities, beneath text and voice, to irony detection, for example facial recognition.
- Investigate in how far prosodic expression of irony is culture dependent, as indicated by (Cheang and Pell, 2009), for example by testing the data with non-German speakers from different cultures.
- Validate hypotheses on acoustic expressions of irony by re-synthesis experiments, for example by systematically varying acoustic features of target sentences with speech synthesis and doing a perception test.

9. References

- Gregory A. Bryant and Jean E. Fox Tree. 2002. Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol*, 17(2):99–117.
- Gregory A. Bryant and Jean E. Fox Tree. 2005. Is there an ironic tone of voice? *Language and Speech*, 48(3):257–277.
- Gregory A. Bryant. 2010. Prosodic contrasts in ironic speech. *Discourse Processes*, 47(7):545–566.
- Felix Burkhardt, Tim Polzehl, Joachim Stegmann, Florian Metze, and Richard Huber. 2009. Detecting real life anger. In *Proceedings ICASSP, Taipei; Taiwan*, 4.
- Henry S. Cheang and Marc D. Pell. 2008. The sound of sarcasm. *JSpeech Communication*, 50(5):366–381.
- Henry S. Cheang and Marc D. Pell. 2009. Acoustic markers of sarcasm in cantonese and english. *Journal of the Acoustical Society of America*, 126(3):1394–1405.
- Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

- Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. of ACM Multimedia 2013*, pages 835–838, Barcelona, Spain. ACM.
- Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.
- Santiago González-Fuente, Prieto Pilar, and Ira Noveck. 2016. A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French. In *Proc. Speech Prosody*, pages 1–5.
- Michael Grimm and Kristian Kroschel. 2005. Evaluation of Natural Emotions Using Self Assessment Manikins. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2005*, pages 381–385, Cancun, Mexico, November. IEEE.
- Michael Grimm, Emily Mower, Kristian Kroschel, and Shrikanth Narayanan. 2007. Primitives based estimation and evaluation of emotions in speech. *Speech Communication*, 49:787–800.
- Simone Hantke, Florian Eyben, Tobias Appel, and Björn Schuller. 2015. iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, pages 891–897, Xi’an, P. R. China, September. AAAC, IEEE.
- Wolfgang Lezius. 2000. Morphy - german morphology, part-of-speech tagging and applications. *Proceedings of the 9th EURALEX International Congress*, pages 619–623.
- Tomoko Matsui, Tagiru Nakamura, Akira Utsumi, Akihiro T. Sasaki, Takahiko Koike, Yumiko Yoshida, Tokiko Harada, Hiroki C. Tanabe, and Norihiro Sadato. 2016. The role of prosody and context in sarcasm comprehension: Behavioral and fMRI evidence. *Neuropsychologia*, 87:74–84.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Patricia Rockwell. 2000. Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, 29(5):483–495.
- Lisa Scharrer, Ursula Christmann, and Monja Knoll. 2011. Voice modulations in German ironic speech. *Language and Speech*, 54(4):435–465.
- Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, et al. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. of INTERSPEECH*, pages 148–152, Lyon, France. ISCA.
- Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70.
- Ulli Waltinger. 2010. GermanPolarityClues: A lexical resource for German sentiment analysis. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)*.
- Deirdre Wilson and Dan Sperber. 2012. Explaining irony. In *Meaning and Relevance*, pages 123–146. Cambridge University Press, Cambridge, MA.
- Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.