Exploring the Realization of Irony in Twitter Data

Cynthia Van Hee, Els Lefever and Véronique Hoste

LT[3], Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
cynthia.vanhee, els.lefever, veronique.hoste@ugent.be

Abstract

Handling figurative language like irony is currently a challenging task in natural language processing. Since irony is commonly used in user-generated content, its presence can significantly undermine accurate analysis of opinions and sentiment in such texts. Understanding irony is therefore important if we want to push the state-of-the-art in tasks such as sentiment analysis. In this research, we present the construction of a Twitter dataset for two languages, being English and Dutch, and the development of new guidelines for the annotation of verbal irony in social media texts. Furthermore, we present some statistics on the annotated corpora, from which we can conclude that the detection of contrasting evaluations might be a good indicator for recognizing irony.

Keywords: social media, figurative language processing, verbal irony

1. Introduction

With the arrival of Web 2.0, technologies like social media have become accessible to a vast amount of people. As a result, they have become valuable sources of information about the public's opinion. What characterizes social media content is that it is often rich in figurative language (Maynard and Greenwood, 2014; Reyes et al., 2013). Handling figurative language represents, however, one of the most challenging tasks in natural language processing. It is often characterized by linguistic devices such as humor, metaphor and irony, whose meaning goes beyond the literal meaning and is therefore often hard to capture, even for humans. Effectively, understanding figurative language often requires world knowledge and familiarity with the conversational context and the cultural background of the conversation's participants; information that is difficult to access by machines. Verbal irony is a particular genre of figurative language that is traditionally defined as saying the opposite of what is actually meant (Curcó, 2007; Grice, 1975; McQuarrie and Mick, 1996). Evidently, this type of figurative language has important implications for tasks that handle subjective information, such as sentiment analysis (Maynard and Greenwood, 2014). Sentiment analysis involves the automatic extraction of positive and negative opinions from online texts. It goes without saying that its accuracy can be significantly undermined by the presence of irony, as illustrated by Example 1. (extracted from the SemEval-2015 training corpus for Task 11[1]).

(1) It was so nice of my dad to come to my graduation party. #not

Regular sentiment analysis systems will probably classify this tweet as positive, whereas the intended emotion is undeniably a negative one. The hashtag *#not* indicates the presence of irony in this example. By contrast, in example 2. (taken from Riloff et al. (2013)), there is no explicit indication of irony present. Nevertheless, the irony is noticeable because given our world knowledge, we know that the act of going to the dentist (for a root canal) is typically

an unpleasant situation. This clearly contrasts with the positive expression *yay, I can't wait*. Considering this contrast, one may assume that the author is not sincere about the expressed sentiment, but rather wants to be ironic.

(2) Going to the dentist for a root canal this afternoon. Yay, I can't wait.

If we want to push the state-of-the-art in tasks like sentiment analysis, it is important to build computational models that are capable of recognizing irony so that the actual meaning of an utterance can be understood if it is not the same as the expressed one. In the current research, we aim to understand how verbal irony works and how it can be recognized in social media texts. Based on the classic definition of verbal irony, our hypothesis is that the detection of contrasting evaluations might be a good indicator for recognizing it.

The remainder of this paper is structured as follows. Section 2. gives an overview of existing work on defining and modeling verbal irony. Section 3. presents the data collection and annotation. In Section 3.1., we discuss the different steps of our annotation scheme and include some examples, while Section 3.2. shows the results of an inter-annotator agreement study to assess the annotation guidelines. Section 4. elaborates on the annotated corpus; a number of statistics are presented to provide insight into the data. Finally, Section 5. concludes the paper with some prospects for future research.

2. Related Research

There are many different theoretical approaches to verbal irony. Traditionally, a distinction is made between situational and verbal irony. Situational irony is often referred to as situations that fail to meet some expectations (Lucariello, 1994; Shelley, 2001). Shelley (2001) illustrates this with firefighters who have a fire in their kitchen while they are out to answer a fire alarm. A popular definition of verbal irony is saying the opposite of what is meant (Curcó, 2007; Grice, 1975; McQuarrie and Mick, 1996; Searle, 1978). While a number of studies have stipulated some nuances to and shortcomings of this standard approach (Camp, 2012;

---

[1] http://alt.qcri.org/semeval2015/task11

Giora, 1995; Sperber and Wilson, 1981), it is commonly used in contemporary research on the automatic modeling of irony (Curcó, 2007; Kunneman et al., 2015; Reyes et al., 2013; Riloff et al., 2013).

When describing how irony works, many studies have come across the problem of distinguishing between verbal irony and sarcasm. To date, there still exists significant diversity of opinion about the definition of verbal irony and whether and how it is different from sarcasm. On the one hand, some studies consistently use one of both terms (Grice, 1975; Grice, 1978; Sperber and Wilson, 1981), or consider both as essentially the same phenomenon (Attardo, 2000; Attardo et al., 2003; Reyes et al., 2013). Camp (2012, p. 625), for example, elaborates a definition of sarcasm that "can accommodate most if not all of the cases described as verbal irony". On the other hand, a number of studies claim that sarcasm and verbal irony do differ in some respects and throw light upon the differences between the two phenomena, including ridicule (Lee, 1998), hostility and denigration (Clift, 1999), and the presence of a victim (Kreuz and Glucksberg, 1989). Most of the studies that are discussed here use the term *verbal irony* generally without specifying whether and how it is differentiated from sarcasm. For this reason, our definition does not distinguish between both phenomena, but rather focusses on a particular form of utterance that can cover both expressions described as verbal irony and expressions that are considered sarcastic.

In contrast to linguistic and psycholinguistic studies on irony, computational approaches to modeling irony are less abundant. Utsumi (1996) describes one of the first forays into the computational modeling of irony. However, their approach assumes an interaction between the speaker and the hearer of irony, which makes it difficult to apply it as a computational model for handling online texts. Nevertheless, with the proliferation of Web 2.0 applications and the considerable amount of data that has come available, irony modeling has received increased interest from the domain of natural language processing. Carvalho et al. (2009) explore the identification of irony through oral and gestural clues in text, including laughter, punctuation marks and onomatopoeic expressions. The authors report accuracies ranging from 45 to 85%. Veale and Hao (2010) propose an algorithm for separating ironic from non-ironic similes. Based on commonly used words and frequent structures in ironic similes, the algorithm manages to identify 87% of the ironic similes. Davidov et al. (2010) build a pattern-based classification algorithm to identify irony in Twitter and Amazon[2]. Reyes et al. (2013) build a classifier to distinguish between ironic and non-ironic tweets. Their approach is based on different four types of features (viz., *signatures*, *unexpectedness*, *style*, and *emotional scenarios*) and the results are promising: the classifier obtains F-scores of up to 62% and 76% with an imbalanced and a balanced distribution, respectively. Similarly to the work of Reyes et al. (2013), Barbieri and Saigon (2014) investigate the identification of ironic tweets among a set of tweets about a particular topic (viz., education, humor and politics). The authors cast the automatic detection of

ironic tweets as a classification problem and make use of a series of lexical features carrying information about word frequency, style (i.e., written -vs- spoken), intensity, structure, sentiments, synonyms, and ambiguity. The analysis reveals that their system outperforms a bag-of-words approach when cross-domain experiments are carried out, but not when in-domain experiments are conducted. Kunneman et al. (2015) constructed a Dutch dataset of tweets. By making use of word n-gram features, their system manages to detect 87% of the ironic tweets in the test set. A careful analysis of the system's output reveals that linguistic markers of hyperbole (e.g., intensifiers) are strong predictors for irony.

To date, most computational approaches to model irony are corpus-based and make use of categorical labels, such as *#irony* assigned by the author of the text. To our knowledge, no guidelines have yet been developed for the more fine-grained annotation of verbal irony in social media content without relying on hashtag information. In this research, we focus on irony modeling in English and Dutch Twitter messages[3]. In accordance with the standard definition, we define irony as *an evaluative expression whose polarity (i.e., positive, negative) is changed between the literal and the intended evaluation, resulting in an incongruence between the literal evaluation and its context*. This definition integrates a number of aspects that are identified as inherent to verbal irony across various studies, including its implicit and evaluative character (Attardo, 2000; Grice, 1975; Kreuz and Glucksberg, 1989; Wilson and Sperber, 1992). Our definition is similar to that of Burgers (2010) in that the irony affects the polarity expressed in the literal and the intended meaning. Unlike Burgers (2010), we prefer the term *polarity change* to *polarity reversal* since the expressed sentiment in an ironic statement can also be stronger or less strong (in ironic hyperbole and understatement, respectively) than the intended one.

3. Data Collection and Annotation

Most current approaches to modeling irony make use of publicly available social media data including tweets and product reviews (Carvalho et al., 2009; Davidov et al., 2010; Kunneman et al., 2015). In the case of tweets, classification algorithms are often based on corpora that are built from tweets labeled with hashtag synonyms of *#irony* or *#sarcasm*. For the current research, a Twitter dataset was constructed for two different languages, being English and Dutch. Both corpora were collected with the hashtags *#irony*, *#sarcasm* (*#ironie* and *#sarcasme* in Dutch) and *#not*. We are aware, however, that irony may also be realized in tweets without any irony-related hashtag. Since no annotation guidelines for verbal irony are available to our knowledge, we elaborated a scheme that presents the different steps for annotating irony in social media text (Van Hee et al., 2015). The scheme allows i) to identify ironic instances in which a polarity change takes place and ii) to indicate contrasting text spans that realize the irony. The following three main annotation steps are taken:

---

1. Based on the definition, indicate for each text whether it is ironic, possibly ironic or not ironic.

2. If the text is ironic:

   - Indicate whether an irony-related hashtag (e.g., *#not*, *#sarcasm*, *#irony*) is required to notice the irony.

   - Indicate the harshness of the irony on a two-point scale (0-1).

3. Annotate contrasting evaluations.

Further details for each annotation step are provided in Section 3.1.

3.1. Annotating Irony in Social Media Text

As a first step in the annotation process, the annotators annotated each tweet as ironic, possibly ironic or non-ironic. According to this annotation scheme, a tweet is considered ironic if it presents a polarity change between the literal and the intended evaluation. The evaluation polarity can be changed in three ways: by using (i) opposition (i.e., the literal evaluation is opposite to the intended evaluation), (ii) hyperbole (i.e., the literal evaluation is stronger than the intended evaluation) or (iii) understatement (i.e., the literal evaluation is less strong than is intended). When annotating irony, annotators thus look for contrasting evaluations. This contrast can be realized by explicit and implicit evaluations (i.e., expressions whose polarity can be inferred through contextual clues and common sense or world knowledge), as irony tends to be implicit (Burgers, 2010).

Figures 1 and 2 show the annotation of contrasting evaluations in brat. The contrast is realized by means of an opposition and a hyperbole, respectively.
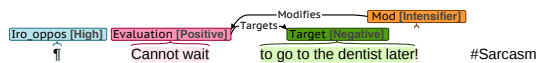

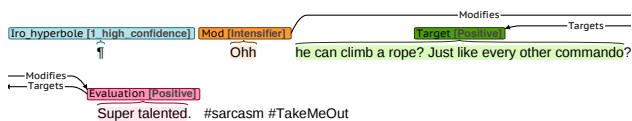Figure 1: Brat annotation of contrasting (opposition) evaluations.

Figure 2: Brat annotation of contrasting (hyperbole) evaluations.

In the first sentence, *cannot wait* is a literally positive evaluation (intensified by an exclamation mark). It is contrasted by the act of going to the dentist, which stereotypically is perceived as unpleasant and thus implicitly carries a negative sentiment. The expression *super talented* in Figure 2 is hyperbolic.

What characterizes Twitter messages is the use of hashtags. Sometimes, these hashtags act as the written equivalent of nonverbal expressions that are used in oral conversations. In the case of irony, the tweet text itself is sometimes not sufficient to notice that the author is being ironic. In other words, only the presence of a hashtag indicates that irony is
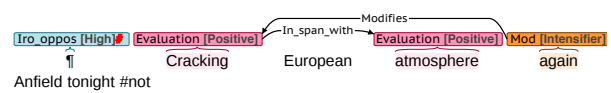

Figure 3: Brat annotation of contrasting evaluations when an irony hashtag is needed.

being used. Figure 3 presents an example where the polarity change is only noticeable by the presence of the hashtag *#sarcasm*. Annotators indicated this with a red mark.

As mentioned earlier, our definition does not distinguish between irony and sarcasm. However, the annotation scheme allows to signal variants of verbal irony that are particularly harsh (i.e., carrying a mocking or ridiculing tone with the intention to hurt someone), as shown in Figure 4.
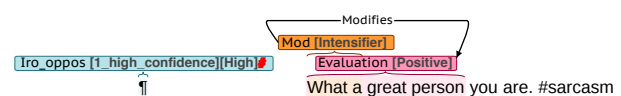

Figure 4: Brat annotation of an evaluation that is harsh.

Tweets that contain some other form of irony than the one described by our definition were annotated as *possibly ironic* (see Figures 5 and 6). This category encompasses two subcategories, being *situational irony* and *other*. The intuition behind this more fine-grained distinction is that there exist other realizations of verbal irony than the ones carrying contrasting evaluations, one of them being descriptions of ironic situations (i.e., *situational irony*). The category *other* encompasses the instances of which the annotators intuitively thought that irony was present, although no polarity change was perceived or no evaluations were present whatsoever.
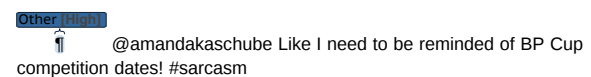

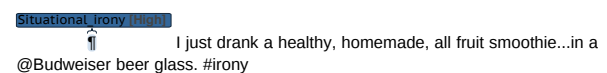Figure 5: Brat annotation of a possibly ironic tweet.

Figure 6: Brat annotation of a possibly ironic tweet describing situational irony.

Despite of having an irony-related hashtag, tweets were considered non-ironic if they do not contain any indication of irony. Additionally, the *non-ironic* category encompasses tweets in which an irony-related hashtag functions as a negator (e.g., *#not*) or is used in a self-referential meta-sentence, as shown in Figure 7.
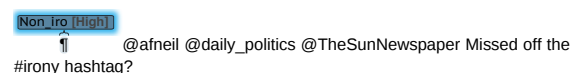

Figure 7: Brat annotation of a non-ironic tweet.

3.2. Inter-annotator Agreement

Both corpora were entirely annotated by students in Linguistics and all annotations were done using the brat rapid annotation tool (Stenetorp et al., 2012). A preliminary inter-annotator agreement study was conducted to assess the guidelines for usability and in particular whether changes or additional clarifications were recommended. We calculated inter-annotator agreement statistics based on four annotations. Firstly, we focused on the identification of a tweet as ironic, possibly ironic or non-ironic (*irony 3-way*). Secondly, we merged the categories *possibly ironic* and *non-ironic* into one category versus the category *ironic* (*irony binary*). Thirdly, we tested the agreement on the decision whether a hashtag is needed to understand the irony. Finally, we calculated the inter-annotator agreement for the identification of contrasting evaluations. Table 1 presents the results of the first inter-annotator agreement study. Based on this study, some changes were made to the annotation scheme (e.g., a refinement of the definition, the instruction to include verbs in the evaluation text span). A second inter-annotator agreement study was conducted on a subset of the corpus to assess whether the annotation scheme can be reliably applied to real-world data. For both the English and the Dutch datasets, a subset of the corpus (100 instances) was independently annotated by three annotators with a background in Linguistics[4].

As shown in Table 2, we averaged the inter-annotator agreement scores of all annotator pairs. We used Cohen's Kappa (Cohen, 1960), which measures inter-rater agreement for categorical labels and takes chance agreement into account. The results of this study show an acceptable level of agreement between the annotators with scores ranging from moderate to substantial.

irony 3-way	irony binary	hashtag indication	polarity contrast
0.55	0.65	0.62	0.57

Table 1: Kappa scores for the first inter-annotator agreement study.

	irony 3-way	irony binary	hashtag indication	polarity contrast
English	0.72	0.72	0.67	0.64
Dutch	0.77	0.84	0.60	0.63

Table 2: Kappa scores for all annotator pairs.

4. Corpus Analysis

In total, 3.000 and 3.179 tweets were annotated for English and Dutch, respectively. Figures 8 and 10 present the statistics for the distribution of the data based on the three main annotation categories, for English and Dutch. As can be inferred from the charts, respectively 57% and 74% of the English and Dutch posts were labeled as ironic according

[4]The annotators for the English corpus also annotated the Dutch corpus. All were Dutch native speakers and Master's students in English.

to our definition. In 99% of the cases, irony was realized by means of an opposition between the intended and the expressed evaluation. In only 1% of the ironic tweets, the polarity change was realized through a hyperbole or an understatement. Figures 9 and 11 show that in about half of the ironic tweets, the irony could not be perceived if there had not been an irony-related hashtag (cf. Figure 3). When we investigate the annotation concerning harshness, we observe that about one third of the English and Dutch ironic tweets (31% and 37%, respectively) is considered to be harsh.
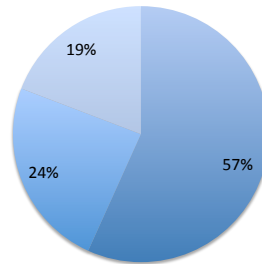

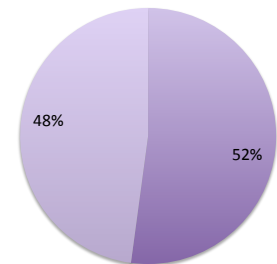Figure 8: Irony annotation (en).

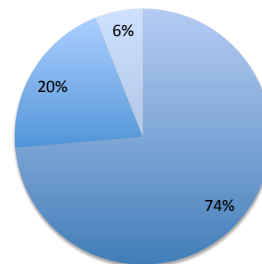Figure 9: Hashtag indication (en).
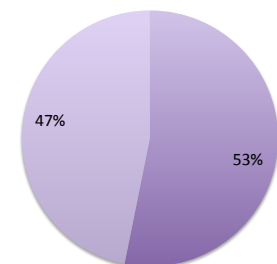

Figure 10: Irony annotation (nl).

Figure 11: Hashtag indication (nl).

Although a major part of the corpus is annotated as ironic, it is interesting to observe that more non-ironic instances were identified in the English corpus when compared to the Dutch corpus (19% versus 6%). We see a number of possible explanations for this difference. Firstly, the hashtag *#not* occurs more frequently (about twice as much) in the non-ironic English tweets than in the non-ironic Dutch tweets. This can be explained by the fact that in English, the word *not* has also a grammatical function (e.g., *Had no sleep and have got school now #not happy*). Secondly, and in line with this observation, their might be an age effect causing some users to write more hashtags than others (for instance *#not* instead of the regular *not*). As a matter of fact, statistics on the Twitter demographics in the Netherlands and the United States[5] show that the largest group of Twitter users in the Netherlands is aged between 45 and 54 years, while the largest group of Twitter users in the US is a

[5]http://www.statista.com

Dutch				
ironic	**possibly ironic**		**non-ironic**	**total**
	situational irony	*other verbal irony*		
2.339	152	496	192	3.179
English				
ironic	**possibly ironic**		**non-ironic**	**total**
	situational irony	*other verbal irony*		
1.703	422	299	576	3.000

Table 3: Statistics of the English and Dutch annotated corpora.

lot younger (between 18 and 24 years)[6]. Finally, for the annotators having Dutch as their native langue, it might have been more difficult to recognize irony in an English corpus when compared to the Dutch corpus. Also noteworthy is the different distribution in the category *possibly ironic* between the two corpora. As is shown in table 3, descriptions of situational irony occur more frequently in the English corpus than in the Dutch one. By contrast, the proportion of tweets labeled as *other* is larger in the Dutch corpus than in the English corpus. Here as well, we might see an effect of the annotators having Dutch as their mother tongue. Hence, they may have more intuitive sense that irony is being used in Dutch text when compared to English text.
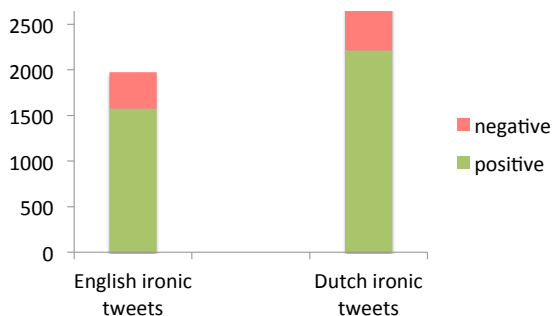

Figure 12: Brat annotation of a non-ironic tweet.

When we investigate the distribution of the literally expressed opinions in ironic tweets, as visualized in Figure 12, we observe that in our corpus, most of the time a literally positive sentiment is expressed, while the intended evaluation is a negative one. The chart also reveals that generally, more evaluations are found in the Dutch than in the English ironic data.

5.   Conclusion and Future Work

We constructed a dataset of about 3.000 tweets for two languages, being English and Dutch, based on a set of irony-related hashtags. We presented a new annotation scheme for the textual annotation of verbal irony in social media text. The guidelines allow to distinguish between ironic, possibly ironic and non-ironic tweets. As the data is human-labeled, we can limit the amount of noise that is introduced when aggregating ironic data based on hashtags.

More concretely, tweets that are not labeled as ironic by human judges (in spite of having a hashtag denoting irony) will be removed from the positive corpus. Another important objective of our guidelines is to identify specific text spans that realize the irony and hence to explore aspects of verbal irony that are susceptible to computational analysis. An inter-annotator agreement study showed an acceptable agreement among the annotators, which asserts the validity of our guidelines. Section 4. elaborates on the annotated datasets and presents some corpus statistics for both the English and the Dutch datasets. The statistics show that in a major part of the corpora, the irony is realized through contrasting evaluations, which means that this might be a good indicator for recognizing irony.

In future work, we will conduct a more elaborate analysis of the annotated corpora to explore i) how verbal irony is realized when no contrasting evaluations are present (i.e., the tweets in the category *other* and ii) whether there exists any consistency in what is perceived as harsh and what is not. Furthermore, we will explore the feasibility to automatically detect ironic utterances based on a polarity contrast, for which the annotated data will serve as a training corpus.

6.   Acknowledgements

7.   Bibliographical References

Attardo, S., Eisterhold, J., Hay, J., and Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor-international Journal of Humor Research*, 16:243–260.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826.

Barbieri, F. and Saggion, H. (2014). Modelling Irony in Twitter, Feature Analysis and Evaluation. In *LREC*, Reykjavik, Iceland.

Burgers, C. (2010). *Verbal irony: Use and effects in written discourse*. UB Nijmegen [Host].

Camp, E. (2012). Sarcasm, Pretense, and The Semantics/Pragmatics Distinction. *Nous*, 46(4):587–634.

Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for Detecting Irony in User-generated Contents: Oh...!!   It's "So Easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, New York, NY, USA.

---

[6]The statistics should be taken with some caution. For the Netherlands, statics were only available from 2013 and for the US from 2015. Moreover, no recent statistics were available for other English- and Dutch-speaking countries such as the UK and Belgium.

Clift, R. (1999). Irony in conversation. *Language in Society*, 28(4):523–553.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Curcó, C. (2007). Irony: Negation, echo, and metarepresentation. *Irony in language and thought*, pages 269–296.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Giora, R. (1995). On irony and negation. *Discourse Processes*, 19(2):239–264.

Grice, H. P. (1975). Logic and Conversation. In Peter Cole et al., editors, *Syntax and Semantics: Vol. 3*, pages 41–58. Academic Press, New York.

Grice, H. P. (1978). Further Notes on Locgic and Conversation. In Peter Cole, editor, *Syntax and Semantics: Vol. 9*, pages 113–127. Academic Press, New York.

Kreuz, R. J. and Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118(4):374.

Kunneman, F., Liebrecht, C., van Mulken, M., and van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management*, 51(4):500–509.

Lee, Christopher J.and Albert N. Katz, A. N. (1998). The Differential Role of Ridicule in Sarcasm and Irony. *Metaphor and Symbol*, 13(1):1–15.

Lucariello, J. (1994). Situational Irony: A Concept of Events Gone Awry. *Journal of Experimental Psychology: General*, 123(2):129–145.

Maynard, D. and Greenwood, M. (2014). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

McQuarrie, E. F. and Mick, D. G. (1996). Figures of Rhetoric in Advertising Language. *Journal of Consumer Research*, 22(4):424–438.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, pages 238–269.

Riloff, E., Qadir, A., Surve, P., Silva, L. D., Gilbert, N., and Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *EMNLP*, pages 704–714. ACL.

Searle, J. R. (1978). Literal meaning. *Erkenntnis*, 13(1):207–224.

Shelley, C. (2001). The bicoherence theory of situational irony. *Cognitive Science*, 25(5):775–818.

Sperber, D. and Wilson, D. (1981). Irony and the Use - Mention Distinction. In Peter Cole, editor, *Radical Pragmatics*, pages 295–318. Academic Press, New York.

Utsumi, A. (1996). A Unified Theory of Irony and Its Computational Formalization. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 962–967, Stroudsburg, PA, USA. Association for Computational Linguistics.

Van Hee, C., Lefever, E., and Hoste, V. (2015). Guidelines for Annotating Irony in Social Media Text, version 1.0. Technical Report LT3 15-02, LT3, Language and Translation Technology Team–Ghent University.

Veale, T. and Hao, Y. (2010). Detecting Ironic Intent in Creative Comparisons. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 765–770, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Wilson, D. and Sperber, D. (1992). On verbal irony. *Lingua*, 87(1):53–76.