Guidelines for Annotating Irony in Social Media Text

version 2.0

LT3 Technical Report – LT3 16-01

Cynthia Van Hee, Els Lefever and Véronique Hoste

LT3, Language and Translation Technology Team Department of Translation, Interpreting and Communication - Ghent University *firstname.lastname@ugent.be*

URL: http://www.lt3.ugent.be¹

June 21, 2016

¹The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from http://www.lt3.ugent.be/publications. All rights reserved. LT3, Department of Translation, Interpreting and Communication, Faculty of Arts, Humanities and Law - Ghent University, Belgium.

Contents

1	Introduction	1
2	Evaluative Expressions	3
3	Evaluation Polarity	4
4	Modifiers	5
5	Irony Presence	6
6	Irony Harshness	9
7	Evaluation Target	10
8	Embedded Evaluations	11
9	Annotation Procedure	12
10	Brat Annotation Examples	13

Introduction

With the emergence of Web 2.0, a large part of our daily communication has moved online. As a consequence, social media (e.g., Twitter, Facebook) have become a valuable source of information about the public's opinion for politicians, companies, researchers, trend watchers, and so on (Pang and Lee, 2008). The past decade has seen an increased research interest in text mining on social media data. The frequent use of irony in this genre has important implications for tasks such as sentiment analysis and opinion mining (Maynard and Greenwood, 2014; Reyes, Rosso, and Veale, 2013), which aim to extract positive and negative opinions automatically from online text. To develop or enhance sentiment analysis systems, or more broadly any task involving text interpretation (e.g., cyberbullying detection), it is of key importance to understand the linguistic realisation of irony, and to explore its automatic detection. Most computational approaches to date model irony by relying solely on categorical labels like irony hashtags (e.g., '#irony', '#sarcasm') assigned by the author of the text. To our knowledge, no guidelines presently exist for the more **fine-grained annotation** of irony in social media content without exploiting this hashtag information.

When describing how irony works, theorists traditionally distinguish between situational irony 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. Verbal irony is traditionally defined as expressions that convey an opposite meaning (e.g., Grice (1975), McQuarrie and Mick (1996), Quintilian (1959)) and implies the expression of a feeling, attitude or evaluation (Grice, 1978; Wilson and Sperber, 1992). There has been a large body of research in the past involving the definition of irony and the distinction between irony and sarcasm (Barbieri and Saggion (2014), Grice (1975), Kreuz and Glucksberg (1989), Wilson and Sperber (1992), amongst others). To date, however, experts do not formally agree on the distinction between irony and sarcasm. For this reason, we elaborate a working definition that can cover both expressions described as verbal irony, and expressions described as sarcasm. In the definition, as well as in the remainder of this paper, we refer to this linguistic form as *irony*. In accordance with the traditional definition and that of Burgers (2010), we define irony as an *evaluative expression whose polarity (i.e., positive, neg*ative) is inverted between the literal and the intended evaluation, resulting in an incongruence between the literal evaluation and its context. More concretely, when speaking ironically, one expresses a positive sentiment whereas the implied sentiment is negative, or inversely.

In our working definition, no distinction is made between irony and sarcasm. However, the present 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), since it has been considered a useful feature for distinguishing between irony and sarcasm (Barbieri and Saggion,

2014; Lee and Katz, 1998).

In what follows, we present the different steps in the annotation of verbal irony in online text. All annotation steps can be executed using the brat rapid annotation tool (Stenetorp et al., 2012). The example sentences in the following chapters are taken from a corpus of English tweets collected using the hashtags '#sarcasm', '#irony' and '#not'. It should be noted that not every element of the examples is annotated and discussed. We refer to Chapter 10 for detailed annotation examples.

Evaluative Expressions

The present definition of irony is based on a polarity inversion between two evaluations. Annotators therefore look for expressions of an **evaluation** in the text under investigation. By an evaluation, we understand the entire text span by which someone or something (e.g., a product, an event, an organisation) is evaluated, including **modifiers** (Chapter 4). There are no restrictions as to which forms evaluations take; they can be verb phrases, predicative (adjective or nominal) expressions, emoticons, and so on. Nevertheless, when possible, annotators should include the verb and its apposition in the annotated text span of the evaluation, as well as modifiers (if present). Evaluative expressions can be found in sentences (1) to (4).

- (1) Oh how I love working in Baltimore #not
 → 'Oh how I love' = evaluation
- (2) What a shock. Duke Johnson is hurt in an important game. #sarcasm #canes → 'What a shock' = evaluation
- (3) So glad you'd rather read a book than acknowledge your own kid #not
 → 'So glad' = evaluation
- (4) Interesting visit with Terra Nova yesterday at Stoneleigh, class tent.
 → 'Interesting' = evaluation
 - \rightarrow 'class' = evaluation

As shown in example (4), a text can include more than one evaluation.

Brat howto

In brat, if an evaluative expression consists of several non-consecutive parts (e.g., 'I <u>love</u> this band <u>so much!</u>'), the parts should be linked by means of drag and drop.

Evaluation Polarity

An important subtask of annotating evaluations is **polarity assignment**, which involves determining whether the expressed evaluation is **positive** (e.g., 'love it!') or **negative** (e.g., 'it's a real nightmare'). It is possible that, due to ambiguity or a restricted context, it is not entirely clear whether an evaluation is positive or negative. Such evaluations receive the polarity label 'unknown'. Nevertheless, annotators should indicate a concrete polarity (i.e., positive or negative) as much as possible.

- (5) I hate it when my mind keeps drifting to someone who no longer matters in life. #irony #dislike
 - \rightarrow 'hate' = evaluation [negative polarity]
 - \rightarrow 'no longer matters in life' = **evaluation** [negative polarity]
 - → '#dislike' = **evaluation** [negative polarity]
- (6) First day off for summer...kids wake up at 6:01. Love them but not Awesome. #sleepisfortheweak #not.
 - \rightarrow 'Love' = **evaluation** [positive polarity]
 - \rightarrow 'not Awesome' = evaluation [negative polarity]
 - → '#sleepisfortheweak' = evaluation [negative polarity]
- (7) I'm surprised you haven't been recruited by some undercover agency. #sarcasm
 → "m surprised' = evaluation [unknown polarity]

Like example (4), sentences (5) and (6) contain more than one evaluative expression. In Twitter data, hashtags may also contain evaluations. In this case, annotators are supposed to annotate the hashtag as an entire unit (including the hash sign '#'), even if it is a multiword expression (e.g., '#sleepisfortheweak').

Like words, all hashtags can be annotated, except '#sarcasm', '#irony' and '#not'. They were used to collect the data and are supposed to be left unannotated.

Modifiers

Sometimes, evaluative expressions are modified. This means that their polarity is changed by an element (i.e., a *modifier*) in the text. Modifiers are lexical items that cause a "**shift in the prior polarity** of other nearby lexical items" (Van de Kauter, Desmet, and Hoste, 2015). They can be left out without losing the sentiment expression.

Two types of modifiers are distinguished in our annotation scheme: (i) intensifiers, which increase the intensity of the expressed sentiment and (ii) diminishers, which decrease the intensity of the expressed sentiment (Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2004). The evaluation polarity can be modified by adverbs (e.g., 'absolutely'), interjections (e.g., 'wow'), punctuation marks (e.g., '??!!'), emoticons, and so on. The modifiers in the sentences (8) and (9) are bold-faced.

- (8) The **most** annoying kid lives next to my door!!!
 - → 'most annoying' = **evaluation** [negative polarity]
 - \rightarrow 'most' = **intensifier** of *annoying*
 - \rightarrow '!!!' = **intensifier** of *most annoying*
- (9) Throwing up at 6:00 am is always fun #not
 - \rightarrow 'is always fun' = **evaluation** [positive polarity]
 - \rightarrow 'always' = **intensifier** of *is fun*

Modifiers can, but are not necessarily, syntactically close to the evaluation. When possible, however, they should be included in the annotation span of the evaluation. As shown in example 37 (Chapter 10), modifiers that are part of an evaluative expression should be included in the evaluation span. Modifiers that are not included in the evaluation span (e.g., punctuation marks, emoticons) can be linked to the evaluation they alter by means of drag and drop.

Brat howto

In brat, modifiers should be linked to the evaluative expression they alter by means of drag and drop.

Irony Presence

According to our definition, verbal irony arises from a clash between two evaluation polarities. This can be illustrated with the following example:

(10) I really love this year's summer; weeks and weeks of awful weather.

In sentence (10), the irony results from a polarity inversion between the literal evaluation ('I really love this year's summer'), which is positive, and the intended one ('I hate this year's summer'), which is implied by the context ('weeks and weeks of awful weather').

Irony involves a polarity clash between what is literally said and what is actually meant. What is actually meant can be explicitly mentioned, or it can be implicit. In the former situation, the literal (ironic) evaluation is opposite to another literal evaluation in the text (e.g., 'Yay for school today! hate it...'). In the latter situation, the literal evaluation is opposite to an implied evaluation that can be inferred by common sense or world knowledge, for instance 'I appreciate you sneezing in my face'. Although 'sneezing in my face' is not an evaluation, it evokes a negative sentiment that contrasts with the literally positive statement 'I appreciate'. Chapter 7 elaborates on the annotation of implicit evaluations, or *evaluation targets*.

Annotators should carefully analyse the evaluation(s) expressed in each text and define whether the text under investigation is ironic by means of a clash or not. Additionally, a confidence score (low, medium or high) should be given for this annotation. It is possible, however, that an instance contains another form of irony: there is no polarity clash between what is said and what is meant, but the text is ironic nevertheless (e.g., descriptions of situational irony). These instances should be included in the category **other types of irony**. Instances that are not ironic should be annotated likewise. The three main annotation categories are listed below:

- **Ironic by means of a clash:** the text expresses an evaluation whose literal polarity is opposite to the intended polarity.
- **Other type of irony:** there is no clash between the literal and the intended evaluation, but the text is still ironic.
- Not ironic: the text is not ironic.

Sentences (11) and (12) are examples of ironic texts in which the literally expressed evaluation is opposite to the intended one. In sentence (11), the irony results from a clash between the literally positive 'Yay can't wait!' and 'Exams start tomorrow', which implicitly conveys a negative sentiment. In contrast to sentence (11), the irony in sentence (12) can only be understood by the presence of the hashtag *#not*. Without this hashtag, it is not possible to perceive a clash between what is literally said and what is implied (i.e., 'my little brother is not awesome'). For similar cases, annotators indicate that a hashtag is required to understand the irony.

(11) Exams start tomorrow. Yay, can't wait! #sarcasm

 \rightarrow the message is **ironic by means of a clash**: the polarity of the literally expressed evaluation 'Yay, can't wait!' is positive, whereas the intended evaluation is negative (having exams is generally experienced as unpleasant).

(12) My little brother is absolutely awesome! #not.

 \rightarrow the message is **ironic by means of a clash**: the polarity of the literal evaluation 'is absolutely awesome!' is positive, whereas the intended evaluation is negative.

Brat howto

In brat, if an irony-related hashtag (i.e., '#sarcasm', '#irony' or '#not') is required to understand that the text is ironic by means of a clash, annotators should check the tick box **'hashtag indica-tion needed'**.

Instances that are ironic but not by means of a clash, should be annotated as **other types of irony**. Sentences (13) to (15) are examples that belong to this category. Sentences (14) and (15) present descriptions of situational irony (Chapter 1).

- (13) "@Buchinator_: Be sure you get in all those sunset instagrams before the sun explodes in 4.5 billion years." Look at your next tweet #irony
- (14) Just saw a non-smoking sign in the lobby of a tobacco company #irony
- (15) My little sister ran away from me throwing a water balloon at her and fell into the pool... #irony.

Examples of **non-ironic** messages are presented in sentences (*16*) to (*19*). As non-ironic, we consider instances that do not contain any indication of irony (example (*16*)) or instances that contain insufficient context to understand the irony (example (*17*)). Additionally, the category encompasses tweets in which an irony-related hashtag is used in a self-referential meta-sentence (example (*18*)), or functions as a negator (example (*19*)).

- (16) Drinking a cup of tea in the morning sun, lovely!
- (17) @GulfNewsTabloid Wonder why she decided to cover her head though! #Irony
- (18) @TheSunNewspaper Missed off the #irony hashtag?
- (19) Those that are #Not #BritishRoyalty should Not presume #Titles or do any #PublicDuties

Brat howto

Whether an instance i) is ironic by means of a clash, ii) contains another type of irony or iii) is not ironic at all, should be annotated on the dummy token ¶ preceding each text. The category *other type of irony* is separated into instances describing situational irony (category *situational irony*) and instances expressing other forms of verbal irony (category *other*).

Irony Harshness

Sarcasm is sometimes considered a bitter or sharp form of irony that is meant to ridicule or hurt a specific target (Attardo, 2000; Barbieri and Saggion, 2014; Lee and Katz, 1998). If a tweet is considered ironic by means of a clash, annotators should indicate the **harshness** of the expressed evaluation (i.e., whether the irony is used to ridicule or hurt a person/a company,...). This can be done on a **two-point scale from 0 to 1**, where 0 means that the evaluation is not harsh and 1 that the evaluation is harsh. Additionally, a confidence score (low, medium or high) should be given for this annotation.

- (20) Well this exam tomorrow is gonna be a bunch of laughs #not
 - \rightarrow the message is ironic by means of a clash: the polarity of the literal evaluation
 - is opposite to that of the intended evaluation
 - \rightarrow the ironic evaluation is **not harsh** (score 0)

(21) Yeah you sure have great communication skills #not

 \rightarrow the message is ironic by means of a clash

 \rightarrow the ironic evaluation is **harsh** (score 1), the evaluation is aimed at a person and is ridiculing

Practical Remark

For convenience and to speed up the annotation, a harshness score of 0 need not be annotated explicitly. When there is no harshness score indicated, the message is considered not harsh.

Evaluation Target

As mentioned in Chapter 5, irony often tends to be realised implicitly (Burgers, 2010). This means that one of the opposite evaluations may be expressed in an **implicit** way; its polarity has to be inferred from the context or by world knowledge/common sense. Such text spans are referred to as the **evaluation target**; their implicit sentiment contrasts with the literal evaluation. In brat, targets should always be linked to the evaluative expression(s) they contrasts with.

Like evaluative expressions, the implicit polarity of an evaluation target can be **positive**, **negative** or **unknown**. It can also be **neutral** when the target corefers to another (the actual) target. In sentence (22) for instance, 'you' is a neutral target that refers to '7 a.m. bedtimes', whose implicit polarity is negative given the context. There are no restrictions as to what forms evaluation targets can take: they can be expressed by a complement to a verb phrase (i.e., verb + verb, verb + adverb, verb + noun) (example (23)), or by a noun phrase (e.g., 'Christmas Day', 'school'), etc. Two targets that are connected by a conjunction should be annotated separately (example (24)).

- (22) Ahh 7 a.m. bedtimes, how I've missed you #not #examproblems
 - → "ve missed = **evaluation** [positive polarity]
 - \rightarrow 'you = target of ''ve missed', which refers to the actual target '7 a.m. bedtimes'
- (23) I did so well on my history test that I got an F-!
 - \rightarrow 'did so well' = **evaluation** [positive polarity]
 - \rightarrow 'got an F-' = **target** of 'did so well'
- (24) I just love when the dog of the neighbours barks unstoppably and I can't sleep #not
 - \rightarrow 'just love' = evaluation [positive polarity]
 - \rightarrow 'just' = **intensifier** of 'love'
 - \rightarrow 'the dog of the neighbours barks unstoppably' = target of 'just love'
 - \rightarrow 'can't sleep' = **target** of 'just love'

Brat howto

In brat, evaluation targets should always be linked to the evaluation they contrasts with by means of drag and drop. They cannot cross sentence boundaries. A coreferential relation between two evaluation targets can also be added by means of drag and drop.

Embedded Evaluations

Sometimes, an evaluation is contained by another evaluation (e.g., sentence (25)). This is called an *embedded* evaluation and needs to be annotated as well. Similarly to evaluative expressions, the polarity of embedded evaluations can be **positive**, **negative** (or **unknown** in the case there is not sufficient context), and its prior polarity may be changed by modifiers.

(25) I'm really looking forward to the awful stormy weather that's coming this week.

- \rightarrow "m really looking forward to" = evaluation [positive polarity]
- \rightarrow 'really' = **intensifier** of 'looking forward to'
- \rightarrow 'the awful stormy weather that's coming this week' = **target** of 'really looking forward to'
- → 'awful' = (embedded) evaluation [negative polarity]

Annotation Procedure

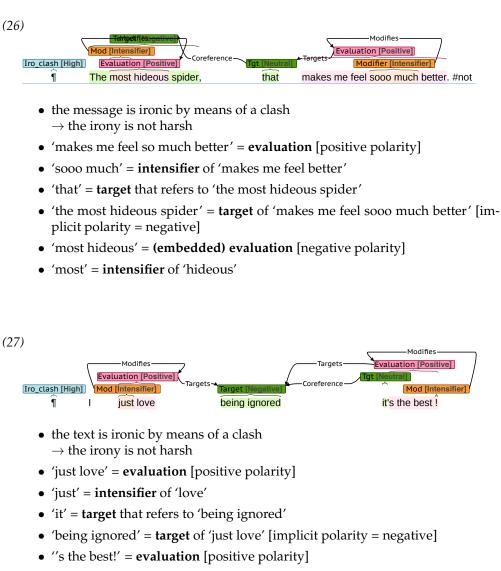
In what follows, we present the different steps in the annotation procedure. It should be noted that, even if a message is not ironic or contains another type of irony than the one based on a polarity clash, annotators should annotate all evaluations that are expressed in the text under investigation. We refer to Chapter 10 for detailed annotation examples in brat.

- 1. Based on the definition, indicate for each text whether it: i) is ironic by means of a clash, ii) contains another type of irony or iii) is not ironic and indicate a confidence score for this annotation.
 - **Ironic by means of a clash**: the text expresses an evaluation whose literal polarity is the opposite of the intended polarity.
 - **Other type of irony:** there is no contrast between the literal and the intended evaluation, however, the text is still ironic (e.g., descriptions of situational irony).
 - Not ironic: the text is not ironic.
- 2. If the text is ironic by means of a clash:
 - In the case of tweets, indicate whether an **irony-related hashtag** (#sarcasm, #irony, #not) is required to recognise the irony.
 - Indicate the **harshness** of the irony on a two-point scale (0-1) and indicate a confidence score for this annotation.

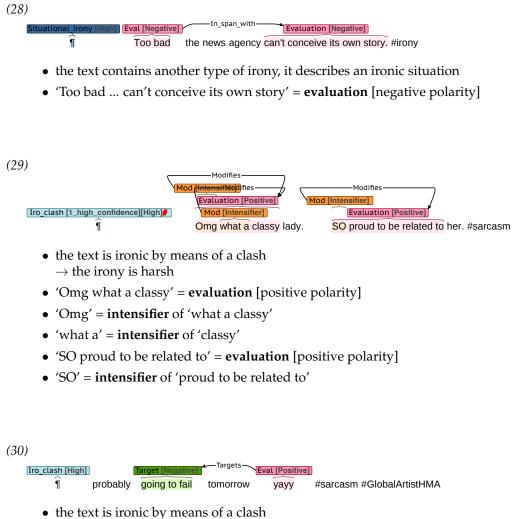
3. Annotate all evaluations contained by the text

- Indicate the polarity of each evaluation.
- If present, annotate modifiers and link them to the corresponding evaluation.
- If present, annotate the evaluation target(s) and link it/them to the evaluation it is in contrast with.
 - * If the target refers to another target, link them by means of a coreferential relation.
 - * Indicate the implicit polarity of the target based on context, world knowledge or common sense.
- 4. If present, annotate embedded evaluations.
- 5. Proceed with the next text.

Brat Annotation Examples



• '!' = intensifier of ''s the best'

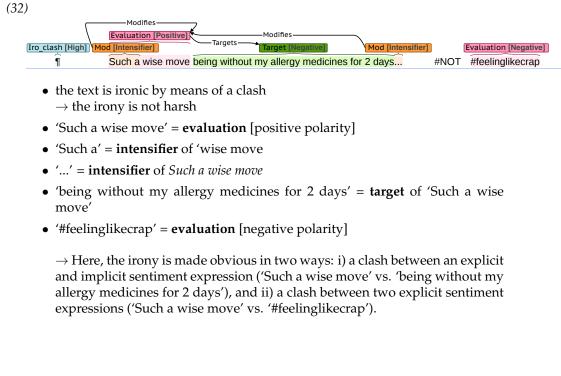


- \rightarrow the irony is not harsh
- 'yayy' = **evaluation** [positive polarity]
- 'going to fail' = **target** of 'yayy' [implicit polarity = negative]

(31)

Non_iro [High

- HOW am I supposed to get over this?! #Not
- the text is not ironic



(33)

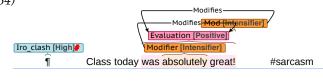
Non iro

¶

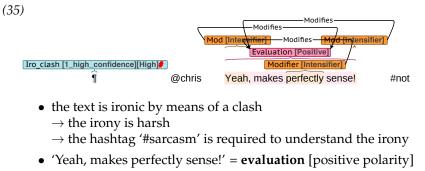
Now i officially look single. Ha the #irony

• the text is not ironic.

(34)



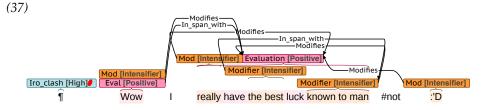
- the text is ironic by means of a clash
 - \rightarrow the irony is not harsh
 - \rightarrow the hashtag '#sarcasm' is required to understand the irony
- 'was absolutely great!' = **evaluation** [positive polarity]
- 'absolutely' = intensifier of 'was ... great!'
- '!' = intensifier of 'was absolutely great'



- 'Yeah' = **intensifier** of 'makes perfectly sense!'
- 'perfectly' = **intensifier** of 'makes ... sense!'
- '!' = **intensifier** of 'makes perfectly sense'



- the text is ironic by means of a clash
 → the irony is not harsh
- 'is a great feeling' = evaluation [positive polarity]
- 'Waking up congested' = target of 'is a great feeling' [implicit polarity = negative]
- 'not being able to breathe' = **target** of 'is a great feeling' [implicit polarity = negative]
- '-_-' = **evaluation** [negative polarity]



- the text is ironic by means of a clash

 → the irony is not harsh
 → the hashtag '#not' is required to understand the irony
- 'really have the best luck known to man' = **evaluation** [positive polarity]
- 'Wow' = **intensifier** of 'really have the best luck known to man'
- 'really' = intensifier of 'Wow ... have the best luck known to man'
- 'the best ... known to man' = intensifier of 'Wow ... really have luck'
- ':'D' = intensifier of 'Wow ... really have the best luck known to man'

References

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

Barbieri, Francesco and Horacio Saggion. 2014. Modelling Irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '14, pages 56–64.

Burgers, Christian. 2010. Verbal irony: Use and effects in written discourse. UB Nijmegen [Host].

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

Grice, H. Paul. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*. Academic Press, New York, pages 41–58.

Kennedy, Alistair and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125.

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

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

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

Maynard, Diana and Mark A. Greenwood. 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), pages 4238–4243.

McQuarrie, Edward F. and David Glen Mick. 1996. Figures of Rhetoric in Advertising Language. *The Journal of Consumer Research*, 22(4):424–438.

Pang, Bo and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Polanyi, Livia and Annie Zaenen. 2004. Contextual Valence Shifters. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications*, AAAI Spring Symposium.

Quintilian. 1959. *The Institutio Oratoria of Quintilian: With an English Translation by H. E. Butler, MA (Vol. 3).* London: William Heinemann (Original work published around AD 95).

Reyes, Antonio, Paolo Rosso, and Tony Veale. 2013. A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

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

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Van de Kauter, Marjan, Bart Desmet, and Véronique Hoste. 2015. The Good, the Bad and the Implicit: A Comprehensive Approach to Annotating Explicit and Implicit Sentiment. *Language Resources and Evaluation*.

Wilson, Deirde and Dan Sperber. 1992. On verbal irony. Lingua, 87(12):53–76.