LiLT volume 14, issue 7 August 2016

Sarcastic Soulmates

Intimacy and irony markers in social media messaging

Koen Hallmann, *Radboud University,* Florian Kunneman, *Radboud University,* Christine Liebrecht, *Tilburg University,* Antal van den Bosch, *Radboud University,* Margot van Mulken, *Radboud University*

Abstract

Verbal irony, or sarcasm, presents a significant technical and conceptual challenge when it comes to automatic detection. Moreover, it can be a disruptive factor in sentiment analysis and opinion mining, because it changes the polarity of a message implicitly. Extant methods for automatic detection are mostly based on overt clues to ironic intent such as hashtags, also known as irony markers. In this paper, we investigate whether people who know each other make use of irony markers less often than people who do not know each other. We trained a machine-learning classifier to detect sarcasm in Twitter messages (tweets) that were addressed to specific users, and in tweets that were not addressed to a particular user. Human coders analyzed the top-1000 features found to be most discriminative into ten categories of irony markers. The classifier was also tested within and across the two categories. We find that tweets with a user mention contain fewer irony markers than tweets not addressed to a particular user. Classification experiments confirm that the irony in the two types of tweets is signaled differently. The within-category performance of the classifier is about 91% for both categories, while cross-category experiments yield substantially lower generalization performance scores of 75% and 71%. We conclude that irony markers are used more often when there is less mutual knowl-

edge between sender and receiver. Senders addressing other Twitter users less often use irony markers, relying on mutual knowledge which should lead the receiver to infer ironic intent from more implicit clues. With regard to automatic detection, we conclude that our classifier is able to detect ironic tweets addressed at another user as reliably as tweets that are not addressed at at a particular person.

Irony markers

The French poet Alcanter de Brahm was the first to propose an ironic sign (a question mark turned backward) to guide readers in the ironic interpretation of an utterance Satterfield (1982). This suggestion was never followed up. One of the reasons may be that this sign would be a spoiler, and ambiguity is precisely one of the goals of ironists. The irony mark would reduce the pleasure of using irony. However, using irony without a sign comes with a risk, because the ironic intention of the communicator may go unnoticed. In order to help the receiver to detect the intention of the communicator, she may use overt signals, irony markers, in the spirit of (but not necessarily as overt as) Alcanter de Brahm's suggestion.

Irony consists of an utterance with a literal evaluation that is implicitly contrary to its intended evaluation. Although irony and sarcasm are not completely synonymous, the phenomena are strongly related (Attardo, 2007, Brown, 1980, Gibbs and O'Brien, 1991, Kreuz and Roberts, 1993, Mizzau, 1984, Muecke, 1969), and are therefore treated as such by researchers (Grice, 1978, Tsur et al., 2010). For the purposes of this article we consider sarcasm to be synonymous with irony and use the terms as interchangeable.

Irony and sarcasm have been the subject of many lively academic debates. The phenomenon has been defined in many different ways (for an overview, see Burgers et al., 2011). Most theories on irony concur that irony is a distinct rhetorical figure and that irony is a property of an utterance that requires the addressee to reconsider the attitude of the communicator (Grice, 1978, Sperber and Wilson, 1995, Clark and Gerrig, 1984, Attardo, 2000a, Giora, 2003). Recently however, some psycholinguistic approaches to irony prefer to consider irony as a broad phenomenon, that encompasses all utterances in a non-serious context and that includes expressions of humor, jocularity and hyperbole (e.g., Colston and Gibbs, 2007; Gibbs, 2000; Pexman et al., 2009). In this paper we use the following definition of irony: Irony is an utterance with "a literal evaluation that is implicitly contrary to its intended evaluation" (Burgers et al., 2011, p. 190).

If an utterance is read ironically, the valence of the evaluation implied in the literal utterance is reversed in the ironic reading (Burgers et al., 2011, p. 190). Some features are essential to irony, which are called irony factors (Attardo, 2000b). If an irony factor is removed from an utterance, this utterance is no longer ironic (Attardo et al., 2003; for a discussion of irony factors, see Burgers et al., 2012a). In contrast, irony markers are meta-communicative clues that can "alert the reader to the fact that an utterance is ironic" (Attardo, 2000b, p. 7), but they are not inherent to irony. An irony marker hints at the receiver that the communicator takes a different stance on the propositional content in the utterance she expresses. Verbal or non-verbal cues that can serve as irony markers may also be used to serve other communicative goals, such as politeness, disagreement, surprise, etc. (Colston, 1997, Colston and O'Brien, 2000). Example 1 contains several irony markers.

(1) *I really can't wait to see everyone's beautiful face in the lucid lights of the hallway at school! #sarcasm*

The intensifier 'really', the hyperbole 'can't wait to see', the indefinite pronoun 'everyone', the positive epithets 'beautiful' and 'lucid', the exclamation mark and the hashtag #sarcasm all signal to the receiver that the communicator is being ironic. However, as long as the discrepancy between the intended meaning and uttered meaning is evident to the receiver, the ironist may refrain from using markers. Had the ironist removed the irony markers from her utterance, and said

(2) *I hope to see you in the lights of the hallway at school*

her utterance would still count as ironic, but the irony would be more difficult to detect (Attardo, 2000b). There must be some discrepancy between the reality and the utterance, but the extent of this discrepancy may vary, and in order to arrive at a successful interpretation of irony, the receiver has to recognize it in order to interpret the utterance as it was intended. Therefore, the communicator may decide to help the receiver and use cues or hints that play a supportive role.

The identification of irony markers has received small but significant attention in the irony literature (Muecke, 1978, Seto, 1998, Burgers et al., 2013). Muecke has been the first to suggest an exhaustive overview of irony markers, discerning between kinesic, graphic, phonic, semantic, and discourse markers. Examples in face-to-face communication are smiles, winks and nudges, pitch, tone of voice and false coughs and air quotes. In written communication, exclamation marks may serve as irony markers, just as dots (...), inverted commas, intensifiers (*very*, *clearly*), superlatives (*best, most, fantastic*), discourse markers (rhetorical questions, *yeah*, *well*) and conventionalized irony, such as *nice* and *fine*. The use of irony markers varies between the medium that is used, both as a result of constraints imposed by the medium (e.g., print does not allow ironic tone-of-voice or facial cues) or convention (e.g., emoticons are accepted in social media, but not in *The New Yorker*). Computer mediated communication (CMC) is often seen as intermediate between written and spoken communication, in that it seeks to incorporate the expressiveness of oral discourse by using cues that guide the interpretations of the utterances. Emoticons, hashtags and typographic markers abound in CMC for signaling irony (Hancock, 2004).

The literature on irony markers is rarely based on empirical research. Burgers et al. (2012b), however, devised a coding scheme for irony markers and then analyzed a corpus of newspaper and magazine columns. They also asked the (human) coders to rate how difficult it was to understand the ironic utterance. Since it is the irony marker's job to hint at ironic intent, they expected that irony accompanied by many irony markers would be easier to understand than ironic utterances with fewer cues. Contrary to expectation, ironic utterances that contained more irony markers were not judged to be less complex than those without markers. Burgers et al. (2012a) then did a follow-up experiment which manipulated the amount of irony markers by adding or deleting them from the original utterances. In this case, they did find the expected effect, and ironic utterances with more markers were easier to understand than the same ironic utterances with fewer irony markers. Burgers et al. (2012a) conclude that ironists use irony markers with particular regard to their estimation of the context of the ironic utterance, including the receiver (see also Burgers, 2010 for a more elaborate discussion).

In sum, ironists can rely on a wide range of cues to signal ironic intent. Since the perception of ironic intent is essential for irony comprehension and since there is no necessity for the explicit signaling of irony (because the discrepancy between what is said and what is meant may be indicative enough for the true intention of the communicator), it can be hypothesized that a lack of familiarity between communicators will increase the probability of the presence of irony markers. Common ground refers to the shared understanding of those involved in the conversation (Clark, 1996). It is the sum of the mutual, common or joint knowledge, beliefs and suppositions of people who engage with each other in a communicative situation.

People who know each other will rely more heavily on common ground, and they will consider irony markers to be spoilers. People who do not know each other, share less common ground and are less inclined to rely solely on the discrepancy between the utterance and the intended meaning. Therefore, they may use more irony markers to avoid the risk of being misunderstood. Communicators who do not know each other should be more confident about irony use when several cues are present (Kreuz, 1996). A solidary relationship between communicators is believed to facilitate the process of understanding irony - the shared common ground and shared thoughts about a particular idea or event in the past make that communicators need less rely on irony markers to get their ironic intent across (Pexman and Zvaigzne, 2004, 146).

Contrary to expectation, however, Caucci and Kreuz (2013) found that communicators use more non-verbal irony markers when talking to friends than when talking to strangers. They provide evidence that irony is signaled by a variety of facial cues, such as movement of the head, eyes, and mouth, and that these cues are more commonly employed by friends than by strangers. It could be the case that facial expressiveness also correlates with familiarity. If this is the case, than it comes as no surprise that non-verbal cues are used more often with friends than with strangers. However, if we are to infer from Caucci and Kreuz's findings that familiarity increases the use of irony markers, then we should find the same tendency in written communication. Therefore, in this paper, we will examine whether people who know each other use more or fewer irony markers than people who do not know each other in social media communication.

Irony detection in social media

Recently, the automatic detection of irony and sarcasm has received a lot of scholarly attention. The field of sentiment analysis and opinion mining aims to automatically tell the polarity of a sentiment. Sarcasm can be a disruptive factor, because it implicitly changes the polarity of a message. The detection of sarcasm is therefore important, if not crucial, for the development and refinement of sentiment analysis systems, but is at the same time a serious conceptual and technical challenge.

Most current approaches, which are mostly statistical and data-driven in nature, test their algorithms on publicly available social media data such as Twitter or product reviews (Carvalho et al., 2009, González-Ibánez et al., 2011, Reyes et al., 2013, Vanin et al., 2013, Davidov et al., 2010, Tsur et al., 2010, Kunneman et al., 2015, Burfoot and Baldwin, 2009) and make use of categorical labels such as hashtags to collect their corpus (for example, Reyes et al., 2013 collected tweets with the hashtag '#irony' and González-Ibánez et al., 2011 collected tweets with '#sarcasm' and '#sarcastic').

Reyes and Rosso (2012) identify humorous and ironic patterns in social media by automatically evaluating features that concern ambiguity, polarity, unexpectedness and emotional scenarios. They show that ironic (and humorous) texts deviate from other messages (political, technical or general tweets). Reyes et al. (2013) propose a set of eight different features, mostly based on the irony literature, to assess potentially ironic statements in different datasets (varying from movie and book reviews to news-wire documents). They include textual features such as punctuation marks and emoticons, emotional scenarios (such imagery and pleasantness) and unexpectedness (based on semantic measures). The authors find that irony is a fairly rare phenomenon in the datasets under investigation. They also find that human annotators, who checked the output of their irony detection algorithm, experience a lot of difficulties in assessing the ironic intent on the basis of isolated fragments. They achieve higher results when the fragments are presented in context.

Recent work on automatic sarcasm detection feed a classifier with more complex features of sarcasm. Riloff et al. (2013) observe that sarcasm is often characterized by a positive sentiment in relation to a negative state or situation. They collect a bootstrapped lexicon of negative situations and positive phrases. Training a machine learning classifier on the co-occurrence of these two yields the best result. Likewise, Joshi et al. (2015) make use of the positive and negative weights of words in a sentiment lexicon to recognize implicit and explicit incongruities in tweets and messages on online fora.

Rajadesingan et al. (2015) and Bamman and Smith (2015) extend the scope to the context outside of a textual unit, and model characteristics of the sender (Rajadesingan et al., 2015, Bamman and Smith, 2015), the addressee and the conversation (Bamman and Smith, 2015) for sarcastic tweets that contain a user mention ('@user'). Several characteristics of the past tweets and user profile of the sender and addressee are included as features. Including all features leads to the best sarcasm detection performance.

While Rajadesingan et al. (2015) and Bamman and Smith (2015) acknowledge that sarcastic tweets with a user mention can be better understood by looking at the relationship between an author and her audience, little is known about the differences in characteristics between sarcastic tweets that are directed towards a specific user and sarcastic tweets that are not. User mentions in social media allow for a distinction between user-directed and general tweets. Tweets with a user mention are directed at a particular addressee, another Twitter user with whom the sender starts or entertains a conversation. General tweets are broadcasted soliloquies, not directed at any person in particular. It is therefore to be expected that twitter users that are already in interaction with each other, or that can refer to prior common knowledge, use fewer irony markers in their sarcastic tweets than twitter users that do not share a common past or conversation.

In line with research on the influence of common ground on the use of sarcastic markers, we treat these two kinds of sarcastic tweets as separate categories and perform a detailed analysis on the types of markers by which they can be recognized as sarcastic. This study is the first to analyze the difference between sarcastic markers in user-directed tweets and tweets without a user-mention. In doing so, we aim to provide insights into the use of sarcasm in different contexts, and thereby contribute to automatic sarcasm detection. Based on the findings of Burgers et al. (2012a), we expect to find that the subset with user mention tweets contains fewer explicit irony markers than the subset with tweets not addressed at a particular user.

Method

To acquire sets of sarcastic markers of sarcastically intended tweets with and without a user mention, we trained a machine learning classifier on both categories and extracted the top-1000 irony predicting elements per category. These elements were subsequently analyzed on the presence of irony markers by two human coders.

Data

In our study we focus on tweets in the Dutch language. We make use of the hashtags #not and #sarcasme (#sarcasm) as a shortcut to collect a large number of sarcastic tweets, and divide them into tweets that contain a user mention (matching for strings prefixed by '@') and tweets that do not.

For the collection of tweets we made use of a database provided by the Netherlands e-Science Centre consisting of IDs of a substantial portion of all Dutch tweets posted from December 2010 onwards (Tjong Kim Sang and van den Bosch, 2013).[1] From this database, we collected all tweets that contained the selected hashtags '#sarcasme' and '#not' until January 31st 2013. This resulted in a set of 644,057 tweets in total. Following Mohammad (2012) and González-Ibánez et al. (2011), we cleaned up the dataset by only including tweets in which the given

[1] http://twiqs.nl/

hashtag was placed at the end or exclusively followed by other hashtags or a url. Hashtags placed somewhere in the middle of a tweet are more likely to be a grammatical part of the sentence than a label (Davidov et al., 2010), and may refer to only a part of the tweet. Applying these filtering steps resulted in 513,547 sarcastic tweets in total as training data, 137,649 tweets containing a @user mention (27% of the total), and 375,898 tweets that do not (73%).[2]

As a background corpus to contrast against sarcastic tweets, we took a sample of tweets in the period from October 2011 until September 2012 (not containing tweets with any of the sarcastic hashtags). To provide the classifier with an equal number of cases for the sarcasm and background categories and thus produce a training set without class skew, 375,898 tweets were selected randomly, equal to the amount of sarcastic tweets without a @user mention.

By leveraging #not and #sarcasm to collect many sarcastic tweets, we deliberately choose for quantity, and are aware of two important implications of this approach. The first is that, as these hashtags are added by users, we can not be sure whether the tweets that contain them are actually sarcastic. Kunneman et al. (2015) annotated a sample of 250 tweets that end with either #sarcasm, #not and #irony, and found that 212 of them, about 90%, were actually sarcastic. Second, #not and #sarcasm, being highly specific markers of sarcasm, will have an influence on the amount and types of markers that are used in the remainder of the tweet. The selected hashtags are very closely related to the definition of sarcasm as a "literal evaluation that is implicitly contrary to its intended evaluation" (Burgers et al., 2011: 190) because they literally imply the examined phenomenon ('#sarcasm') and the intended polarity flip ('#not'). As a result, it is likely that the collected tweets are indeed sarcastic (Kunneman et al. (2015) but on the other hand, less obvious sarcastic tweets with perhaps their own sarcastic characteristics and irony markers are not present in this study (see Filatova (2012) and Walker et al. (2012) for alternative data collection methods). Table 1 presents some examples of sarcastic tweets with or without user mention.

Extraction of sarcastic markers

In order to acquire the irony predicting elements for both user-directed tweets and tweets without a user-mention, we trained a machine-learning classifier to distinguish sarcastic from non-sarcastic utterances in both categories and extracted the top-1000 most predicting elements.

[2] The tweet IDs for both sets of tweets can be downloaded from `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:65746`

TABLE 1 Examples of tweets in the dataset marked with #sarcasme or #not, addressed to a user or not addressed to a user. The tweets are translated from Dutch.

User	No user
@USER It's a shame Lu, they adore me, you know... and @USER as well.. #sarcasm	It's always a pleasure to go to Sneekes #not #sarcasm
@USER otherwise we won't stand a chance, because this is of the utmost importance! #sarcasm	Maybe 2C will have to pay attention during English class #not #sarcasm
@USER Great, right? to travel a couple of hours for that. A nice start time as well, so we can sleep late. This makes me very happy #sarcasm	HAHA YOU ARE SO FUNNY #not #sarcasm
#Buitenhof @USER 'If you don't innovate, you stagnate' - Gosh! what an eye-opener #sarcasm	Wow, it's so awesome to have soap in my eye!! #not #sarcasm #pain
@USER haha she's such a lovely lady #sarcasm	There, put on the outfit again. Will go into the great atmosphere and work until 11 PM #sarcasm

We trained a classifier on the following two datasets:

1. 137,649 sarcastic tweets with a user-mention, labeled as 'sarcastic', equated with a sample of 137,649 of the random tweets, labeled as 'non-sarcastic'.
2. 375,898 sarcastic tweets without a user-mention, labeled as 'sarcastic', equated with the 375,898 random tweets, labeled as 'non-sarcastic'

Before classification, the tweets in both sets were tokenized.[3] Punctuation and emoticons were kept as potential elements to signal sarcasm (Burgers et al., 2012b). We lowercased all tokens, but maintained capitalization for tokens that were completely written in capitals. To further normalize the tweets, we converted each token to their lemma.[4]

[3]Tokenization was carried out with Ucto, `http://ilk.uvt.nl/ucto`
[4]Lemmatization was carried out with Frog, `http://ilk.uvt.nl/frog`

We only extracted word uni-, bi- and trigrams as features (including punctuation and emoticons as separate words), to acquire a largely unbiased set of features for analysis. We removed features containing one of the hashtags '#not' and '#sarcasme', by which the tweets were collected, and features containing a user mention.

As classification algorithm we employed Balanced Winnow (Littlestone, 1988) as implemented in the Linguistic Classification System.[5] This algorithm is known to offer state-of-the-art results in text classification, and produces interpretable per-class feature weights that can be used to inspect the highest-ranking features for one class label. The α and β parameters were set to 1.05 and 0.95 respectively. The major threshold ($\theta+$) and the minor threshold ($\theta-$) were set to 2.5 and 0.5. The number of iterations was bounded to a maximum of three. After training the classifier, we selected the 1000 features with the highest rank for both datasets.[6] The top 20 features for both datasets is presented in Table 2.

Corpus analysis

After gathering the data,[7] a coding scheme was developed based on (Burgers et al., 2012a,b). The coding scheme was pretested on 300 elements out of the total of 2000 and then adapted to best fit the data. The category 'Metaphor' was dropped because no metaphors were present, and the category 'Ambiguity' was added, because in some cases the meaning of an element was unclear.

The final coding scheme consists of ten dichotomous variables representing irony predicting elements of tweets. The first three categories are all binary and concern the nature of the expression: **Evaluation** denotes an evaluation in the element (e.g. "fun"), as opposed to an objective, descriptive meaning ("long"); **Polarity** concerned the polarity of the evaluation, which can be positive or negative; and **Ambiguity** is defined as an element with multiple possible meanings of which the intended meaning cannot be ascertained by the coder. The Evaluative/Descriptive and Ambiguous categories are mutually exclusive. If an element is evaluative (e.g. "fun"), it is not descriptive or ambiguous. Only evaluative elements have a polarity which can be positive or neg-

[5] http://www.phasar.cs.ru.nl/LCS/

[6] The ranked features can be downloaded from `http://cls.ru.nl/~fkunneman/data_sarcastic_soulmates.zip`

[7] All data collected for this study will be made freely available, as well as all annotations of the features by the human coders. The coders were the first author and a student assistant and worked independently. Twitter data will be made available in the form of tweet IDs. This footnote is a placeholder for the URL offering links to the data.

TABLE 2 Top ranked 20 features after training a classifier on the detection of sarcastic tweets with a user-mention and sarcastic tweets without a user mention ('BOS' is the beginning-of-sentence mark)

Tweets addressed to user		Tweets not addressed to user	
Feature	Gloss	Feature	Gloss
gezellig	cosy	#leuk	#fun
#gezellig	#cosy	#jippie	#yippie
#leuk	#fun	#fijn	#nice
#slim	#smart	#heelfijn	#verynice
medelijden	pity	#gezellig	#cosy
LEUK	FUN	#slim	#smart
lekker	nice	#luckyme	#luckyme
leuk	fun	#altijdleuk	#alwaysfun
jippie	yippie	#handig	#practical
beetje_maar	just_a_little	#yay	#yay
ge-wel-dig	great	#yes	#yes
#jippie	#yippie	#altijdfijn	#alwaysnice
lekkerding	hottie	#gaatgoed	#goeswell
BOS_ik_en	BOS_me_and	#jeej	yay
#fijn	#nice	#boeiend	#interesting
que	que	intressant	interesting
slimme	smart	leeeuk	fuun
#grapje	#joke	#jeeeej	#yaay
geweldige	great	LEUK	FUN
joepie	yippie	#joepie	#yippie

ative. The other seven variables concern specific irony markers. These categories are not mutually exclusive because an element can contain multiple markers, for instance the hyperbolic and all capitals "FANTASTIC". The category **Hyperbole** was defined as a word strongly deviating from the semantic average (e.g., "fantastic" was defined a hyperbole but "nice" was not). **Interjections** such as "gee" were defined as words that have no referent, but do have meaning. **Repetitions of letters or vowels** refers to repeating letters (e.g., "grrrrrreat"). **Capitals** was defined as irregular use of capitals, with the exception of the first letter of a word which as defined as regular and lower cased to avoid the use of a capital constituting a separate element (e.g. "Great" and "great" were lower-cased, whereas "GREAT" was not). **Punctuation marks** cover all punctuation marks except for comma's, @ and #. **Hashtags** is the use of # before a word (e.g., #fun). The last category of irony markers are **Emoticons**, defined as simulating facial expressions through punctuation marks.

Both coders [8] were presented with all 2000 elements, but were unaware of the category to which the element belonged. In total, 71 (3.55%) elements were excluded from the analysis because it was unclear what their meaning was, leaving 974 elements of the top-1000 irony predicting elements from tweets not addressed to another Twitter user and 955 elements from tweets that did address another user.[9] All disagreement between coders was resolved by one of the other authors who acted as a third, independent coder. The Cohen's Kappa values indicating agreement between the coders are given in Table 3.

TABLE 3 Cohen's Kappa values indicating agreement for annotating the presence of irony markers.

Evaluation	.66
Postive polarity (if evaluative)	.66
Ambiguity	.43
Interjection	.80
Repetition	.83
Capitals	.86
Punctuation marks	.96
Hashtag	.94
Emoticon	1.00
Hyperbole	.58

Results

To check for differences in the frequency of the presence of irony markers between elements of tweets that address another user and tweets that do not, chi-square analyses were used. Comparing the average amount of irony markers was done with a one-way Analysis of Variance.

Regarding the three categories concerning the nature of sarcastic tweets, elements from tweets that mention another user differed in evaluativeness, $\chi^2(1, N = 1929) = 34.46, p < .000$. Elements from tweets that mention another user were less frequently evaluative (35%) than those from tweets that did not mention another user (48.2%).

The evaluations of elements from tweets that mention another user were as often positive as those that were not addressed at another

[8] We wish to thank Mathilde Blom for assisting with the coding of the predicting elements.

[9] The top ranked features and their annotations can be downloaded from `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:65746`

TABLE 4 Frequencies of presence of irony markers for top-1000 elements of tweets that mention users and those that do not in function of variables. $n = 1929$, except Polarity where $n = 803$. $* = p < .05, ** = p < .01, *** = p < .001$.

	Tweets addressed to user	Tweets not addressed to user
Evaluation***	35%	48.2%
Positive polarity (if evaluative)	95.8%	96.2%
Ambiguity	20.3%	17.2%
Hyperbole	7.7%	9.4%
Repetition***	5.5%	12.6%
Hashtags***	4.1%	13.6%
Capitals*	0.3%	1.1%
Punctuation marks***	2.3%	0.4%
Emoticons***	2.5%	0.5%
Interjections**	10.1%	14.8%

user, $\chi^2(1, N = 1803) = 0.06, p = .081$. The evaluations of elements from tweets that mention another user did not differ in their polarity (95.8% positive, 4.2% negative) from those that did not mention a user (96.2% positive, 3.8% negative). Elements from tweets which mentioned another user did not differ significantly on ambiguity from those that did not mention a user $\chi^2(1, N = 1929) = 2.97, p = .085$.

Regarding the presence of irony markers in the elements, there was an overall significant difference in the sum of the irony markers (Hyperbole, Interjections, Repetition, Hashtag, Capitals, Punctuation Marks and Emoticons) between elements from tweets with and without @user mentions, $F(1, 1928) = 46.54, p < .000$. Elements from tweets that mention another user had an average of .33 ($SD = .56$) irony markers, whereas elements from tweets that did not mention another user had an average of .52 ($SD = .71$) irony markers. The amount of irony markers in an element varied between 0 and 3 for both categories (see Table 5).

TABLE 5 Percentages of elements and their amount of irony markers (range 0-3) as a function of the user mention category, $n = 1929$.

	Amount of markers 0	1	2	3
No user mentioned	59.8%	28.6%	11.0%	0.6%
User mentioned	71.9%	23.7%	4.3%	0.1%

As to the specific irony markers, for almost every irony marker a difference was found between elements from tweets that mention another user and tweets that did not mention another user. There was a significant difference in the use of repetition of letters and vowels, $\chi^2(1, N = 1929) = 29.14, p < .000$. Elements from tweets that mention another user did not feature repetition (5.5%) as frequently as elements from tweets that did not mention another user (12.6%). The use of hashtags (aside from #sarcasme and #not) also differed significantly, $\chi^2(1, N = 1929) = 53.51, p < .000$. Elements from tweets that mention another user did not feature hashtags (4.1%) as often as elements from tweets that did not mention another user (13.6%). There was a significant difference in use of capitals, $\chi^2(1, N = 1929) = 4.45, p = .035$. Elements from tweets that mention another user did not feature capitals (0.3%) as frequently as elements from tweets that did not mention another user (1.1%). The use of punctuation marks also differed significantly, $\chi^2(1, N = 1929) = 13.00, p < .000$. Elements from tweets that mention another user featured punctuation marks (2.3%) more frequently than elements from tweets that did not mention another user (0.4%). The same was found for emoticons, $\chi^2(1, N = 1929) = 13.02, p < .000$. Elements from tweets that mention another user featured emoticons (2.5%) more frequently than elements from tweets that did not mention another user (0.5%). For interjections there was also a significant difference, $\chi^2(1, N = 1929) = 9.91, p = .002$. Elements from tweets that mention another user were less frequently in the form of an interjection (10.1%) than elements from tweets that did not mention another user (14.8%). However, elements from tweets which mentioned another user did not differ significantly on the presence of hyperbole, $\chi 2(1, N = 1929) = 1.77, p = .184$.

Cross-category classification experiments

As irony-predicting elements from tweets that contain a user mention on average contain fewer irony markers than tweets that are not addressed to specific users, we would expect a machine learning classifier that was trained on the former category to perform less well on the latter category than the other way around.

We tested this hypothesis by performing a cross-category classification experiment. We equated the conditions of the 'user' and 'non-user' categories by reducing the (larger) amount of 'non-user' sarcastic tweets to the amount of the 'user' category (137,649 tweets), and selecting non-overlapping samples of 137,649 random tweets for training and testing. This resulted in four different sets for the cross-category classification.

In addition to training on one of the training sets and testing on the contrasting test set, resulting in two classification experiments, we performed a within-category classification by means of 10-fold cross-validation on both train sets. Again, we applied Balanced Winnow for classification, using the features as described in the Method Section.

The classification performance, measured in terms of $F_{\beta=1}$-scores on classifying the 'sarcasm' class, is displayed in Table 6. These scores show that a classifier trained and tested within its own category is better at predicting whether a tweet is sarcastic than a classifier that is tested on the other category. As was shown in the corpus analysis, there are significant differences in the amounts of irony markers between the two categories, which is reflected in the lower cross-category scores.

TABLE 6 F-scores for classifying 'sarcasm' by training on the data set displayed in the row and testing on the column

	Addressed to user	Not addressed to user
Addressed to user	0.91	0.72
Not addressed to user	0.75	0.91

Training on tweets not addressed to a user and testing on the other category leads to a slightly higher F-score of 0.75 than training on tweets addressed to a user and testing on the other category (0.72). Apparently, the higher number of explicit markers that were identified from the former category helps the classifier to better identify sarcastic tweets in the 'user mention' category.

In contrast to our expectations, the two classifiers that are trained and tested on the same category both yield a score of 0.91. We expected a worse performance for the 'user-mention' category with its reduced use of explicit sarcastic markers. Apparently, other elements, such as topical words, are still useful to recognize sarcastic tweets in this category. Of course, if a topic such as 'school' is addressed in an ironic discussion, then human speakers will not see the word 'school' itself as an irony marker (which is a clue to ironic intent that is purposefully used by the sender), but the use of certain topical words such as 'school' may aid automatic detection because they occur relatively frequently in ironic tweets.

Conclusion

The use of irony markers differs significantly between elements from the tweets addressed at specific users and those that are not. Elements from sarcastic tweets not directed at specific users are less often evaluative.

There was no difference when it came to the polarity of the elements. For both categories, elements that were evaluative were equally often positively evaluative. We conclude therefore that the top irony predicting elements from tweets between users who know each other, are less often evaluative and therefore more implicit than elements from tweets addressed at no particular user.

The irony predicting elements from tweets that contain a user mention display less often repetition of letters and vowels, hashtags, capital letters, or interjections and on average contain fewer irony markers than elements from tweets that are not addressed to specific users. Conversely, punctuation marks and emoticons were more frequent in elements from ironic tweets between Twitter users rather than when sarcasm was directed at no one in particular. There were no differences with regard to ambiguity and hyperbole.

Automatic machine-learning-based sarcasm detection in tweets addressed at specific users and tweets that are not, confirms that the sarcasm in the two types of tweets is marked differently. The within-category performance of classifiers trained with the Balanced Winnow learning algorithm is about 91% for both categories, while cross-category experiments yield substantially lower generalization performance scores of 75% and 71%. These results confirm that sarcasm is marked differently between tweets that are directed at a user or not, but also show that machine-learning-based sarcasm detection is still able to detect sarcasm between users as accurately as sarcasm at no user in particular, as long as it has been trained on a specific corpus.

Discussion

In general, the results confirm our hypothesis that irony markers are used more often when there is less mutual knowledge between sender and receiver. Tweets addressing other Twitter users contain less often irony markers, and rely instead on mutual knowledge which should lead the receiver to infer ironic intent. There are, however, three exceptions that ask for some explanation. First, emoticons were used more often in user-mention tweets, which suggests that emoticons are less unequivocal than other irony markers. Indeed, emoticons were judged to be ambiguous far more often (79.3%) than elements of tweets that did not contain emoticons (17.8%), $\chi^2(1, N = 1929) = 70.80, p < .000$. In fact, a winking emoticon only signals that the content of the utterance is not to be taken literally. 19 (79.17%) of the 24 different emoticons that were among the top-1000 elements predicting sarcasm in user-mention tweets were judged to be ambiguous. Punctuation marks were also judged as ambiguous more often (50%) than the average element (18.3%), $\chi^2(1, N = 1929) = 16.87, p < .000$. Also, 11 (50%) of the 22 emoticons that were among the top-1000 elements predicting sarcasm in tweets not addressing another user were judged to be ambiguous.

The second finding that does not match our expectations is the fact that there is no difference in the use of hyperbole between the two categories. However, this finding is in line with research that focuses on the functions of hyperbole in irony (Colston, 1997, Colston and O'Brien, 2000). Hyperbole alters both the literal and the ironic meaning of an ironic utterance. For example, the ironically intended "fantastic job!" is interpreted as more negative than the ironically intended "nice job!". It appears that in ironic tweets, hyperbole is not only used to signal ironic intent, but it serves other purposes as well. This finding converges with Burgers et al.'s (2012a) corpus analysis. Recall that Burgers et al. (2012a) found that ironic utterances with more irony markers than others were not judged to be less complex, because the ironists used irony markers differently between contexts (see Burgers et al. 2012b; Burgers, 2010). The only exception to this was also hyperbole, most likely because it has functions beyond merely signaling ironic intent.

Finally, there was no difference in ambiguity between the two categories. Although around 18% of the top 1000 elements predicting sarcasm were judged to be ambiguous in meaning, this number is probably slightly inflated because coders judged the individual elements rather than entire tweets, and this result is probably a side effect of the chosen coding method.

The performance that was yielded as part of the cross-category classification experiments was high relative to the scores that are reported in other works on sarcasm detection. For example, González-Ibánez et al. (2011) report an accuracy of 75.95 as highest score on distinguishing sarcastic from positive tweets, and Riloff et al. (2013) yield an optimal F-score of 0.51. A probable reason for the high scores in our experiment, optimally an F-score of 0.91, is that the data made the task simpler. For example, sarcasm was contrasted against completely random tweets, rather than tweets with positive sentiment or with a certain topic. Furthermore, the distribution of sarcastically labeled and other tweets was identical during training and testing. Importantly, though, the experiment gave an impression of the difference between sarcastic tweets with and without a user mention in the context of sarcasm detection.

One important drawback in our study is that we focused on tweets in which the sender already made explicit that she was using sarcasm, because she used the hashtags #not or #sarcasme. However, by studying other elements in these tweets, we have shown that when communicators know each other, they make less use of explicit clues to ironic intent than when they do not know each other. We have thus found confirmation for our hypothesis. In sum, the use of irony markers varies as a function of context. These findings converge with those of Burgers et al. (2012a,b). However, the classifier trained on the corpus of tweets that mentioned another user performed equally accurate on tweets that mentioned a user as the non-user mention classifier did on tweets addressed at no one in particular. It appears that even though the tweets addressed at a user contain fewer irony markers, those markers that are used still allow for relatively accurate automatic detection. It appears that in a given time frame certain topics are often discussed ironically, and that the topic in itself is an irony marker. In conclusion, it appears that the tweets analyzed in this study are within the limits of superficial methods of sarcasm detection, although a classifier does need to be trained on a sample that reflects the amount and types of irony markers used - which, as we found, can vary.

Our results may have important implications both for research on irony and sarcasm as well as for computational linguistics that focuses on non-literal language in general.

First of all, our findings may help to build more sophisticated classifiers. The corpus analysis points to prominent sarcastic markers, such as repetition and interjections, that might increase the accuracy of sarcasm detection when explicitly incorporated in the feature space. Secondly, our finding suggests that ironists consciously vary the transparency of their ironic intent as a function of their audience and context. The ironist apparently makes an estimation of the difficulty of the context and varies the number of markers accordingly to accommodate her audience. Ironists wish to achieve different goals when using irony, using it strategically (Bryant, 2012). It might be the case that depending on the goal, the number of irony markers or the type of irony markers varies.

For instance, Kaufer (1977) argues that one of the functions of irony is to induce a sense of pleasure by suggesting that both the ironist and the addressee belong to an inner circle ('wolves'), consisting of those witty enough to comprehend the sender's ironic intent, at the expense of 'sheep' who are none the wiser; see also Gibbs and Izett (2005), van Mulken et al. (2010). Stern (1990) mentions another function; she suggests that irony inherently gives a sense of self-satisfaction, because the receiver is witty enough to be able to see through the ambiguity. Horton (2007) on the other hand claims that understanding figurative language in general both maintains and establishes a degree of intimacy, because the receiver was able to infer what the sender meant based on mutual knowledge, which itself is the result of a degree of intimacy. Given the results of the current research, all these different functions may induce the ironist to vary the use of irony markers accordingly. These are interesting avenues for future research.

With regard to automatic sarcasm detection, the current study emphasizes the importance of context in both the production and comprehension process of non-literal language. Our findings underline what has also been stated by Wallace (2013) and Reyes et al. (2013) who concur that unless an irony predicting algorithm accounts for an explicit model of the communicator and the communicative situation, automatic irony detection will remain a challenge. It follows from our results that in the case of communicative situations where there is a relatively high degree of mutual knowledge, for instance for Instant Messaging services such as Whatsapp and Facebook Messenger, statistical methods for sarcasm detection solely based on explicit cues are very likely to lack accuracy (Wallace et al., 2014). Future research should therefore explore new ways of operationalizing context. For instance, the number of followers a Twitter user has may be correlated to the number of irony markers she uses, simply because it is impossible to share mutual knowledge with all addressees.

Acknowledgements

This research was funded by the Dutch national program COMMIT. We wish to thank Mathilde Blom for assisting with the coding of the predicting elements, Erik Tjong Kim Sang for the development and support of the `http://twiqs.nl` service, and BuzzFeed for the YouTube skit entitled "Sarcastic Soulmates", `https://www.youtube.com/watch?v=Lun7SoXwJk8`.

References

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

Attardo, S. 2000b. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *RASK* 12:3–20.

Attardo, S. 2007. Irony as relevant inappropriateness. In R. W. Gibbs, R. W. G. Jr., and H. Colston, eds., *Irony in language and thought: A cognitive science reader*, pages 135–170. New York, NY: Lawrence Erlbaum.

Attardo, S., J. Eisterhold, J. Hay, and I. Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor* 16(2):243–260.

Bamman, David and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Brown, R. L. 1980. The pragmatics of verbal irony. In R. W. Shuy and A. Shnukal, eds., *Language use and the uses of language*, pages 111–127. Washington, DC: Georgetown University Press.

Bryant, G. A. 2012. Is verbal irony special? *Language and Linguistics Compass* 6(11):673–685.

Burfoot, C. and T. Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164. Association for Computational Linguistics.

Burgers, C., M. van Mulken, and P.J. Schellens. 2012a. Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics* 44(3):231–242.

Burgers, C., M. van Mulken, and P.J. Schellens. 2012b. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology* 31(3):290–310.

Burgers, C., M. van Mulken, and P.J. Schellens. 2013. The use of co-textual irony markers in written discourse. *humor* 26(1):45–68.

Burgers, C., M. van Mulken, and P. J. Schellens. 2011. Finding irony: an introduction of the verbal irony procedure (vip). *Metaphor and Symbol* 26(3):186–205.

Burgers, C. F. 2010. *Verbal irony: Use and effects in written discourse*. Nijmegen, The Netherlands: Ipskamp.

Carvalho, Paula, Luís Sarmento, Mário J Silva, and Eugénio De Oliveira. 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*, pages 53–56. ACM.

Caucci, Gina M. and Roger J. Kreuz. 2013. Social and paralinguistic cues to sarcasm. *Humor* 25(1):1–22.

Clark, Herbert H. 1996. *Using language*. Cambridge university press.

Clark, Herbert H and Richard J Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General* 113(1):121–126.

Colston, H. and R Gibbs. 2007. Irony in language and thought: A cognitive science reader. In R. W. Gibbs and H. L. Colston, eds., *A brief history of irony*, pages 3–21. Psychology Press.

Colston, H. L. 1997. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes* 23(1):25–45.

Colston, H. L. and J. O'Brien. 2000. Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse Processes* 30(2):179–199.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. 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.

Filatova, Elena. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pages 392–398.

Gibbs, Jr., R.W. 2000. Irony in talk among friends. *Metaphor and Symbol* 15(2):5–27.

Gibbs, R. W. and C. Izett. 2005. Irony as persuasive communication. In H. Colston and A. Katz, eds., *Figurative language comprehension: Social and cultural influences*, pages 131–151. New York, NY: Lawrence Erlbaum.

Gibbs, R. W. and J. O'Brien. 1991. Psychological aspects of irony understanding. *Journal of pragmatics* 16(6):523–530.

Giora, R. 2003. *On our mind: Salience, context, and figurative language*. Oxford University Press.

González-Ibánez, Roberto, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 581–586. Stroudsburg, PA, USA: Association for Computational Linguistics.

Grice, H. 1978. Further notes on logic and conversation. In P. Cole, ed., *Pragmatics: syntax and semantics*, pages 113–127. New York, NY: Academic Press.

Hancock, Jeffrey T. 2004. Verbal irony use in face-to-face and computer-mediated conversations. *Journal of Language and Social Psychology* 23(4):447–463.

Horton, W. S. 2007. Metaphor and readers' attributions of intimacy. *Memory & cognition* 35(1):87–94.

Joshi, Aditya, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)* pages 757–762.

Kaufer, David. 1977. Irony and rhetorical strategy. *Philosophy & Rhetoric* 10(2):90–110.

Kreuz, Roger J. 1996. The use of verbal irony: Cues and constraints. *Metaphor: Implications and applications* pages 23–38.

Kreuz, R. J. and R. M. Roberts. 1993. The empirical study of figurative language in literature. *Poetics* 22(1):151–169.

Kunneman, Florian, Christine Liebrecht, Margot Van Mulken, and Antal Van den Bosch. 2015. Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management* 51(4):500–509.

Littlestone, N. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* 2:285–318.

Mizzau, M. 1984. *L'ironia: la contraddizione consentita*. Milan, Italy: Feltrinelli.

Mohammad, Saif M. 2012. #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255. Stroudsburg, PA, USA: Association for Computational Linguistics.

Muecke, D. C. 1969. *The compass of irony*. Oxford University Press.

Muecke, D. C. 1978. Irony markers. *Poetics* 7(4):363–375.

Pexman, Penny M, Lenka Zdrazilova, Devon McConnachie, Kirby Deater-Deckard, and Stephen A Petrill. 2009. "that was smooth, mom": Children's production of verbal and gestural irony. *Metaphor and Symbol* 24(4):237–248.

Pexman, Penny M and Meghan T Zvaigzne. 2004. Does irony go better with friends? *Metaphor and Symbol* 19(2):143–163.

Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.

Reyes, Antonio and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems* 53(4):754–760.

Reyes, Antonio, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47(1):239–268.

Riloff, Ellen, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

Satterfield, L. 1982. The ironic sign. In *Semiotics 1980*, pages 467–474. U.S.: Springer.

Seto, K.-i. 1998. On non-echoic irony. In R. Carson and S. Uchida, eds., *Relevance theory: Applications and implications*, pages 239–255. Amsterdam, The Netherlands: John Benjamins.

Sperber, D. and D. Wilson. 1995. *Relevance: Communication and cognition*. Oxford, UK: Blackwell Publishers, 2nd edn.

Stern, B. B. 1990. Pleasure and persuasion in advertising: rhetorical irony as a humor technique. *Current Issues and Research in Advertising* 12((1-2)):25–42.

Tjong Kim Sang, E.k and A. van den Bosch. 2013. Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal* 3:121–134.

Tsur, O., D. Davidov, and A. Rappoport. 2010. Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169.

van Mulken, Margot, Christian Burgers, and Bram van der Plas. 2010. Wolves, confederates, and the happy few: The influence of comprehension, agreement, and group membership on the attitude toward irony. *Discourse Processes* 48(1):50–68.

Vanin, Aline A, Larissa A Freitas, Renata Vieira, and Marco Bochernitsan. 2013. Some clues on irony detection in tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 635–636. International World Wide Web Conferences Steering Committee.

Walker, Marilyn A, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Wallace, Byron C. 2013. Computational irony: A survey and new perspectives. *Artificial Intelligence Review* pages 1–17.

Wallace, Byron C, Laura Kertz Do Kook Choe, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 512–516.