Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: [http://oatao.univ-toulouse.fr/](http://oatao.univ-toulouse.fr/)
Eprints ID: 18921


To link to this article URL: [http://eacl2017.org/images/site/Proceeding/book_long.pdf](http://eacl2017.org/images/site/Proceeding/book_long.pdf)


Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr
Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study

Jihen Karoui1, Farah Benamara1, Véronique Moriceau2, Viviana Patti3, Cristina Bosco3, and Nathalie Aussenac-Gilles1

1IRIT, CNRS, Université de Toulouse, France
2LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, France
3Dipartimento di Informatica, University of Turin, Italy
1{karoui,benamara,aussenac}@irit.fr
2{moriceau}@limsi.fr
3{patti,bosco}@di.unito.it

Abstract

This paper provides a linguistic and pragmatic analysis of the phenomenon of irony in order to represent how Twitter’s users exploit irony devices within their communication strategies for generating textual contents. We aim to measure the impact of a wide-range of pragmatic phenomena in the interpretation of irony, and to investigate how these phenomena interact with contexts local to the tweet. Informed by linguistic theories, we propose for the first time a multi-layered annotation schema for irony and its application to a corpus of French, English and Italian tweets. We detail each layer, explore their interactions, and discuss our results according to a qualitative and quantitative perspective.

1 Introduction

Irony is a complex linguistic phenomenon widely studied in philosophy and linguistics (Grice et al., 1975; Sperber and Wilson, 1981; Utsumi, 1996). Glossing over differences across approaches, irony can be defined as an incongruity between the literal meaning of an utterance and its intended meaning. For many researchers, irony overlaps with a variety of other figurative devices such as satire, parody, and sarcasm (Clark and Gerrig, 1984; Gibbs, 2000). In this paper, we use irony as an umbrella term that includes sarcasm, although some researchers make a distinction between them, considering that sarcasm tends to be more aggressive (Lee and Katz, 1998; Clift, 1999).

Different categories of irony have been studied in the linguistic literature such as hyperbole, exaggeration, repetition or change of register (see section 3 for a detailed description). These categories were mainly identified in literary texts (books, poems, etc.), and as far as we know, no one explored them in the context of social media. The goal of the paper is thus four folds: (1) analyse if these categories are also valid in social media contents, focusing on tweets which are short messages (140 characters) where the context may not be explicitly represented; (2) examine whether these categories are linguistically marked; (3) test if there is a correlation between the categories and markers; and finally (4) see if different languages have a preference for different categories.

This analysis can be exploited in a purpose of automatic irony detection, which is progressively gaining relevance within sentiment analysis (Maynard and Greenwood, 2014; Ghosh et al., 2015). In particular, it will bring out the most discriminant pragmatic features that need to be taken into account for an accurate irony detection, therefore helping systems improve beyond standard approaches that still heavily rely on features gleaned from the utterance-internal context (Davidov et al., 2010; Gonzalez-Ibanez et al., 2011; Liebrecht et al., 2013; Buschmeier et al., 2014; Hernández Fariñas et al., 2016).

To this end, informed by well-established linguistic theories of irony, we propose for the first time:

- A multi-layered annotation schema in order to measure the impact of a wide-range of pragmatic phenomena in the interpretation of irony, and to investigate how these phenomena interact with context local to the tweet. The schema includes three layers: (1) irony activation types according to a new perspective of how irony activation happens-- explicit vs. implicit, (2) irony categories as defined in previous linguistic studies, and (3) irony markers.
- A multilingual corpus annotated according
to this schema. As the expression of irony is very dependent on culture, we chose, for this first study, three Indo-European languages whose speakers share quite the same cultural background: French, English and Italian. The corpus is freely available for research purposes and can be downloaded here http://github.com/IronyAndTweets/.

- A qualitative and quantitative study, focusing in particular on the interactions between irony activation types and markers, irony categories and markers, and the impact of external knowledge on irony detection. Our results demonstrate that implicit activation of irony is a major challenge for future systems.

The paper is organised as follows. We first present our data. Sections 3 and 4 respectively detail the annotation scheme and the annotation procedure. Section 5 discusses the reliability study whereas Section 6 the quantitative results. In Section 7, we compare our scheme to already existing schemes for irony stressing the originality of our approach and the importance of the reported results for automatic irony detection. Finally we end the paper by showing how the annotated corpora are actually exploited in automatic irony detection shared tasks.

2 Data

The datasets used in this study are tweets about hot topics discussed in the media. Our intuition behind choosing such topics is that the pragmatic context needed to infer irony is more likely to be understood by annotators compared to tweets that relate personal content. We relied on three corpora in French, English and Italian, referred to as F, E and I respectively. Table 1 shows the distribution of ironic vs. non ironic tweets in the data.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Ironic</th>
<th>Not Ironic</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>2,073</td>
<td>16,179</td>
</tr>
<tr>
<td>E</td>
<td>5,173</td>
<td>6,116</td>
</tr>
<tr>
<td>I</td>
<td>806 (Sentipolc)</td>
<td>5,642 (Sentipolc)</td>
</tr>
<tr>
<td></td>
<td>+ 2,273 (TW-SPINO)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Distribution of tweets in each corpus.

The selection of ironic vs non-ironic tweets has been based on partly different criteria for the three addressed languages in order to tackle their features. In English and French, users employ specific hashtags (#irony, #sarcasm, #sarcastic) to mark their intention to be ironic. These hashtags have been often used as gold labels to detect irony in a supervised learning setting. Although this approach cannot be generalized well since not all ironic tweets contain hashtags, it has however shown to be quite reliable as good inter-annotator agreements (kappa around 0.75) between annotators’ irony label and the reference irony hashtags have been reported (Karoui et al., 2015). Nevertheless, irony corpus construction through hashtag filtering is not always possible for all languages. For instance, both in Czech and Italian, Twitter users generally do not use the sarcasm (i.e. ‘#sarkasmus’, in Czech; ‘#sarcasmo’ in Italian) or irony (‘#ironie’ in Czech or ‘#ironia’ in Italian) hashtag variants to mark their intention to be ironic, thus in such cases relying on simple self-tagging for collecting ironic samples is not an option (Ptáček et al., 2014; Bosco et al., 2013). Similar considerations hold for Chinese (Tang and Chen, 2014). For what concerns Italian, we observe that even if occasionally Italian tweeters do use creative hashtags to explicitly mark the presence of irony, no generic shared hashtags have been used for long-time which can be considered as firmly established indicators of irony like those used for English.

The corpora built for English and French are new datasets built using the Twitter API as follows. We first selected 9 topics (politics, sport, artists, locations, Arab Spring, environment, racism, health, social media) discussed in the French media from Spring 2014 until Autumn 2015 and in the American media from Spring 2014 until Spring 2016. For each topic, we selected a set of keywords with and without hashtag: politics (e.g. Hollande, Obama), sport (e.g. #Zlatan, #FI-FAworldcup), etc. Then, we selected ironic tweets containing the topic keywords and the French (English) ironic hashtags. Finally, we selected non ironic tweets that contained only the topic keywords without the ironic hashtag. We removed duplicates, retweets and tweets containing pictures which would need to be interpreted to understand the ironic content. For English, since we were interested in ironic tweets for our annotation purpose, we stopped collecting messages when the number of ironic tweets was sufficient; this ex-
explains the fact that classes of Ironic and Not Ironic tweets in the English dataset are pretty balanced, i.e. the amount of ironic tweets is not very low compared with the amount of not ironic ones.

Italian data are instead extracted from two existing annotated data: the SentiPolc corpus, released for the shared task on sentiment analysis and irony detection in Twitter at Evalita 2014 (Basile et al., 2014), and TW-SPINO which extends the Spinoza section of the Senti-TUT corpus (Bosco et al., 2013). The SentiPolc dataset is a collection of Italian tweets derived from two existing corpora Senti-TUT and TWITA (Basile and Nissim, 2013). It includes Twitter data exploiting specific keywords and hashtags marking political topics. In SentiPolc, each tweet has an annotation label among five mutually exclusive labels: positive opinion, negative opinion, irony, both positive and negative, and objective. TW-SPINO instead is from the Twitter section of Spinoza¹, a popular collective Italian blog that publishes posts with sharp satire on politics. Since there is a collective agreement about the fact that these posts include irony mostly about politics, they represent a natural way to extend the sampling of ironic expressions. Moreover, while SentiPolc collects tweets spontaneously posted by Italian Twitter users, Spinoza’s posts are selected and revised by an editorial staff, which explicitly characterize the blog as satiric. Such difference will possibly have a reflection on the types and variety of irony we detect in the tweets.

3 A multi-layered annotation schema for irony in social media

To define our annotation schema, we analyzed the different categories of irony studied in the linguistic literature. Several categories have been proposed, as shown in the first column of Table 2. Since all these categories have been found in a specific genre (literary texts), the first step was to check their presence on a small subset of 150 ironic tweets from our corpus. Three observations resulted from this first step, regarding irony activation, irony categories, and irony markers.

3.1 Irony activation

We observed that incongruity in ironic tweets often consists of at least two propositions (or words) $P_1$ and $P_2$ which are in contradiction to each other (i.e. $P_2 = \text{Contradiction}(P_1)$). It is the presence of this contradiction that activates irony. This contradiction can be at a semantic, veracity or intention level. $P_1$ and $P_2$ can be both part of the internal context of an utterance (that is explicitly lexicalized), or one is present and the other one implied. We thus defined two types of irony activation: EXPLICIT and IMPLICIT.

In EXPLICIT activation, one needs to rely exclusively on the lexical clues internal to the utterance, like in (1) where there is a contrast between $P_1$ that contains no opinion word, and $P_2$ which refers to a situation which is commonly judged as being negative, but in a communicative context which is clearly unsuitable w.r.t. to the one expressed in $P_1$.

(1) L’Italia [attendire spiegazioni]$P_1$ da così tanti paesi che comincio a pensare che le nostre richieste [finiscano nello spam]$P_2$.

(Italy is [waiting for explanations]$P_1$ from so many countries that I suspect our requests are being [labeled as spam]$P_2$.)

Example (2) shows another example of explicit semantic contradiction between $P_1$ and $P_2$.

(2) Ben non! [Matraquer et crever des yeux]$P_1$, [ce n’est pas violent et ça respecte les droits]$P_2$ !!! #ironie

(Well, no! [Clubbing and putting up eyes]$P_1$, [it is not violent and it does respect human rights]$P_2$ !!! #irony)

On the other hand, IMPLICIT activation arises from a contradiction between a lexicalized proposition $P_1$ describing an event or state and a pragmatic context $P_2$ external to the utterance in which $P_1$ is false, not likely to happen or contrary to the writer’s intention. The irony occurs because the writer believes that his audience can detect the disparity between $P_1$ and $P_2$ on the basis of contextual knowledge or common background shared with the writer. For example, in (3), the negated fact in $P_1$ helps to recognize that the tweet is ironic.

(3) La #NSA a mis sur écoutes un pays entier. Pas d’inquiétude pour la #Belgique: [ce n’est pas un pays entier]$P_1$ #ironie

(The #NSA wiretapped a whole country. No worries for #Belgium: [it is not a whole country]$P_1$ #irony)

$\rightarrow$ P2: Belgium is a country.

¹http://www.spinoza.it/
3.2 Irony categories

Both explicit and implicit activation types can be expressed in different ways which we call irony categories. After a thorough inspection of how categories have been defined in linguistic literature, some of them were grouped, like hyperbole and exaggeration, as we observed that it is very difficult to distinguish them in short messages. We also discarded others, since we considered them as markers rather than irony categories (see the last row in Table 2). We finally retain eight categories, as shown in Table 2: Five are more likely to be found in both types of activation (marked Both) while three may occur exclusively in a specific type (marked Exp for explicit or Imp for implicit).

Categories are not mutually exclusive. Example (5) shows a case of implicit irony activation where the user uses a false assertion P₁ and two rhetorical questions.

(5) @infos140 @mediapart Serge Dassault? Corruption ? Non ! Il doit y avoir une erreur. [C’est l’image même de la probité en politique]P₁ #ironie.

Table 2: Irony categories in our annotation schema.

<table>
<thead>
<tr>
<th>State of the art irony categories</th>
<th>Our categories</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metaphor (Ritchie, 2005; Burgers, 2010)</td>
<td>Analogy\textsuperscript{\textit{both}}, Metaphor and Comparison</td>
<td>Covers analogy, simile, and metaphor. Involves similarity between two things that have different ontological concepts or domains, on which a comparison may be based</td>
</tr>
<tr>
<td>Hyperbole (Berntsen and Kennedy, 1996; Mercier-Leca, 2003; Didio, 2007)</td>
<td>Hyperbole/Exaggeration\textsuperscript{\textit{both}}</td>
<td>Make a strong impression or emphasize a point</td>
</tr>
<tr>
<td>Exaggeration (Didio, 2007)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euphemism (Muecke, 1978; Seto, 1998)</td>
<td>Euphemism\textsuperscript{\textit{both}}</td>
<td>Reduce the facts of an expression or an idea considered unpleasant in order to soften the reality</td>
</tr>
<tr>
<td>Rhetorical question (Barbe, 1995; Berntsen and Kennedy, 1996)</td>
<td>Rhetorical question\textsuperscript{\textit{both}}</td>
<td>Ask a question in order to make a point rather than to elicit an answer (P₁: asking a question to have an answer, P₂: no intention to have an answer because it is already known)</td>
</tr>
<tr>
<td>Context shift (Haiman, 2001; Leech, 2016)</td>
<td>Context Shift\textsuperscript{\textit{exp}}, \textsuperscript{\textit{imp}}</td>
<td>A sudden change of the topic/frame, use of exaggerated politeness in a situation where this is inappropriate, etc.</td>
</tr>
<tr>
<td>False logic or misunderstanding (Didio, 2007)</td>
<td>False assertion\textsuperscript{\textit{imp}}</td>
<td>A proposition, fact or an assertion fails to make sense against the reality</td>
</tr>
<tr>
<td>Oxymoron (Gibbs, 1994; Mercier-Leca, 2003)</td>
<td>Oxymoron/paradox\textsuperscript{\textit{exp}}</td>
<td>Equivalent to “False assertion” except that the contradiction is explicit</td>
</tr>
<tr>
<td>Paradox (Tayot, 1984; Barbe, 1995)</td>
<td>Other\textsuperscript{\textit{both}}</td>
<td>Humor or situational irony (irony where the incongruity is not due to the use of words but to a non intentional contradiction between two facts or events)</td>
</tr>
<tr>
<td>Situational irony (Shelley, 2001; Niogret, 2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surprise effect, repetition, quotation marks, emoticons, exclamation, capital letter, crossed-out text, special signs (Haiman, 2001; Burgers, 2010)</td>
<td>Markers</td>
<td>Words, expressions or symbols used to make a statement ironic</td>
</tr>
</tbody>
</table>

Note that inferring irony in both types of activation requires some pragmatic knowledge. However, in case of implicit, the activation of irony happens only if the reader knows the context. To help annotators identify irony activation type, we apply the following rule: if P₁ and P₂ can be found in the tweet, then explicit, otherwise implicit.

3.3 Irony markers

As shown in Table 2, linguistic literature considers other forms of irony categories, such as surprise effect, repetition, etc. Having a computational perspective in mind, we preferred to clearly distinguish between categories of irony which are pragmatic devices of irony as defined in the previous section, and irony markers which are a set of tokens (words, symbols, propositions) that may activate irony on the basis of the linguistic content of the tweet only. This distinction is also motivated by the fact that markers can either be present in distinct irony categories, not present at all, or present in non ironic tweets as well.

Eighteen markers have been selected for our study. Some of them have shown their effectiveness when used as surface features in irony detection such as punctuation marks, capital letters, reporting speech verbs, emoticons, interjections,
negations, opinion and emotion words (Davidov et al., 2010; Gonzalez-Ibanez et al., 2011; Reyes et al., 2013; Karoui et al., 2015). We investigate in addition novel markers (cf. Table 5): discourse connectives as they usually mark oppositions, argumentation chains and consequences; named entities and personal pronouns, as we assume they can be an indicator of the topic discussed in the tweet (media topic vs. a more personal tweet); URLs as they give contextual information that may help the reader to detect irony; and finally false propositions. These last four markers might be good features for an automatic detection of implicit irony, for example by detecting that an external context is needed. For example, in (2) markers are negations (no, not), punctuation (!, !!!), opinion word (violent) whereas in (3) markers are named entities (NSA, Belgium), negation (no, not) and false proposition (it is not a whole country).

4 Annotation procedure

For each tweet \(t\), the annotation works as follows\(^2\):

(a) Classify \(t\) into Ironic/Not ironic. In case annotators do not understand the tweet because of cultural references or lack of background knowledge, \(t\) can be classified into the No decision class. Note that this third class concerns only French and English corpora since the Italian corpus already has annotations for irony (cf. Section 2).

(b) If \(t\) is ironic, define its activation type: Can \(P_1\) and \(P_2\) be found in the tweet? If yes then explicit, otherwise implicit. Then specify the pragmatic devices used to express irony by selecting one or several categories.

(c) Identify text spans within the tweet that correspond to a pre-defined list of linguistic markers. Markers are annotated whatever the class of \(t\). This is very important for analyzing the correlation between ironic (vs. non ironic) readings and the presence (vs. absence) of these markers.

Linguistic markers were automatically identified relying on dedicated resources for each language (opinion and emotion lexicons, intensifiers, interjections, syntactic parsers for named entities, etc.).

In case of missing markers or erroneous annotations, automatic annotations were manually corrected. Also, to ensure that the annotations were consistent with the instructions given in the manual, common errors are automatically detected: ironic tweets without activation type or irony category, absence of markers, etc. Annotators were asked to correct their errors before continuing to annotate new tweets.

In order to evaluate the stability of the schema regarding language variations, we considered first the French set with a total of 2,000 tweets. Such tweets have been randomly selected from the ones collected as described in Section 2. In order to be sure to have a significant amount of ironic samples, 80\% of the total tweets be manually annotated were selected from the ironic set (i.e. tweets explicitly marked with hashtags like #ironic and #sarcasme)\(^3\). Three French native speakers were involved. The annotation of the French corpus followed a three-step procedure where an intermediate analysis of agreement and disagreement between the annotators was carried out. Annotators were first trained on 100 tweets, then were asked to annotate separately 300 tweets (this step allows to compute inter-annotator agreements, cf. next section), to finally annotate 1,700 tweets. In the last step, a revised version of the schema was provided. The adjudicated annotations performed in the second step are part of the corpus.

Then we annotated the English and Italian sets in two steps. First, a training phase (100 tweets each) and then the effective annotation, with respectively 550 and 500 tweets. Four native speakers were involved: two for English and two for Italian. All annotators are skilled in linguistics, researchers and PhD students in computational linguistics.

5 Qualitative results

We report on the reliability of the annotation schema on the French data. Among 300 tweets, annotators agreed on 255 tweets (174 ironic and 63 not ironic), among which 18 have been classified as No decision. We get a Cohen’s Kappa of 0.69 for Ironic/Not ironic classification which is a

\(^2\)The annotation manual is available at: github.com/IronyAndTweets/Scheme

\(^3\)Notice that at this stage such hashtags have been removed, and manual annotation have been applied to 2,000 tweets for all the layers foreseen by our schema. In this way, the reliability of self-tagging has been confirmed, and it was possible to identify the presence of irony also in tweets where it was not explicitly marked by hashtags.
very good score. When compared to gold standard labels, we also obtained a good Kappa measure (0.62), which shows that French irony hashtags are quite reliable. We also noticed that more than 90% of the tweets annotated as No decision due to the lack of external context, are in fact ironic according to gold labels. We however decided to keep them for the experiments.

For Explicit vs. Implicit, agreement on activation type knowing the tweet ironic obtained a Kappa of 0.65. It was interesting to note that implicit activation is the majority (76.42%). We observed the same tendency in the other languages too (cf. next section). This is an important result that shows that annotators are able to identify which are the textual spans that activate the incongruity in ironic tweets, whether explicit or implicit, and we expect automatic systems to do as good as humans, at best.

Finally, for irony category identification, since the same ironic tweet can belong to several irony categories, we computed agreements by counting, for each tweet, the number of common categories and then dividing by the total number of annotated categories. We obtained 0.56 which is moderate. This score reflects the complexity of the identification of pragmatic devices. When similar devices are grouped together (mainly hyperbole/exaggeration and euphemism, as they are used to make the intended meaning either stronger or weaker), the score increases to 0.60.

6 Quantitative results

The main aim of our corpus-based study is to verify if the different linguistic theories and definitions made on irony can be applied to social media, especially to tweets, and to study its portability to several languages. Besides standard frequencies, we provide the correlations between irony activation types and markers and between categories and markers in order to bring out features that could be used in a perspective of automatic irony detection. In each corpus, all the frequencies presented here are statistically significant from what would be expected by chance using the $\chi^2$ test ($p < 0.05$).

Table 3 gives the total number of annotated tweets and the activation type for ironic tweets. We observe that most irony activation types in the French and English corpora are implicit with respectively 73.01% and 66.28% while in the Italian corpus, explicit activation is the majority. Notice that the fact the analysis of the Italian dataset results in a different tendency on this respect can be possibly related to the absence of user-generated ironic hashtags, while user explicitlty mark the intention to be ironic (see Section 2).

<table>
<thead>
<tr>
<th></th>
<th>Ironic</th>
<th>Non Ironic</th>
<th>No decision</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>394</td>
<td>1066</td>
<td>380</td>
<td>160</td>
</tr>
<tr>
<td>E</td>
<td>144</td>
<td>283</td>
<td>99</td>
<td>24</td>
</tr>
<tr>
<td>I</td>
<td>260</td>
<td>140</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Number of tweets in annotated corpora in French (F), English (E) and Italian (I).

Table 4 gives the percentage of tweets belonging to each category of irony split according to explicit vs. implicit activation, when applicable. Higher frequencies are in bold font. We note that oxymoron/paradox is the most frequent category for explicit irony in French, English and Italian. Concerning implicit irony, false assertion and other are the most frequent categories in French and English (other is the most frequent one in English because a majority of implicit ironic tweets use situational irony, e.g. Libertarian Ron Paul condemns Bill Clinton for taking advantage of 20y/o but would not support any law to protect her: #Monica). In Italian, false assertion, analogy and other are the most frequent categories. As classes are not mutually exclusive, there are 64/38 tweets (resp. in French and English) that belong to more than one category for explicit contradiction. The most frequent combinations are oxymoron/rhetorical question and oxymoron/other for both English and French; oxymoron/hyperbole for French and oxymoron/analogy for English. Concerning implicit activation, there are 134/62 tweets (resp. in French and English) that belong to more than one category. The most frequent combinations are false assertion/other and false assertion/hyperbole for both English and French; and analogy/other for English.

Table 5 provides the percentage of tweets containing markers for ironic (explicit or implicit) and non ironic tweets (row in gray). In French, intensifiers, punctuation marks and interjections are more frequent in ironic tweets whereas quotations are more frequent in non ironic tweets. In English, discourse connectors, quotations, comparison words and reporting speech verbs are twice as
frequent in ironic tweets as in non ironic tweets whereas is it the opposite for personal pronouns. Note that there is no English ironic tweet containing URL since they were all annotated as no decision because of a lack of knowledge from the annotators who did not understand the tweet and the Web page pointed by the URL. In Italian, most of markers are more frequent in ironic tweets, while some, like quotations and URL, are more frequent in non ironic tweets.

Our study of negation as an irony marker actually considers negation words like no and not as well as periphrastic forms of negation such as ne pas in French. We however excluded lexical negations such as unreliable, unhappy, etc. We will further refine our analysis by considering more words that introduce negation. Also, regarding personal pronouns, they are more common in French and English than in Italian. Italian being a pro-drop language can in part motivate the difference detected with respect to pronouns.

Then, we investigated the correlation between irony markers and irony activation types (resp. between irony markers and irony categories). Our aim is to analyze to what extent these markers can be indicators for irony prediction. Using the Cramer’s V test (Cohen, 1988) on the number of occurrences of each marker, we found a statistically significant (p < 0.05) large correlation between markers and ironic/not ironic class for French (V = 0.156, df = 14) and Italian (V = 0.31, df = 6); between medium and large for English (V = 0.132, df = 9). We also found a large correlation between markers and irony activation types for French (V = 0.196, df = 16), between medium and large for Italian (V = 0.138, df = 5) and medium for English (V = 0.083, df = 12).6

We also analyzed the correlations per marker (df=1). The markers which are the most correlated to ironic/non ironic class are: negations, interjections, named entities and URL for French (0.140 < V < 0.410); negations, discourse connectors and personal pronouns for English (0.120 < V < 0.170); and quotations, named entities and URL for Italian (0.310 < V < 0.416). The markers which are the most correlated to explicit/implicit activation are: opposition markers, comparison words and false assertion for French (0.140 < V < 0.190); opposition markers and discourse connectors for English (0.110 < V < 0.120); and discourse connectors, punctuation and named entities for Italian (0.136 < V < 0.213). Note that even if opinion words are very frequent in ironic tweets, they are however not correlated with either irony/non irony classification or explicit/implicit activation (V < 0.06), as many non ironic tweets also contain sentiment words.

Finally, when analyzing which markers are correlated to irony categories, the more discriminant markers are: intensifiers, punctuation, false assertion and opinion words for French (large Cramer’s V); negations, discourse connectors and personal pronouns for English (medium Cramer’s V); and punctuation, interjections and named entities for Italian (medium Cramer’s V).

### 7 Related work

Most state of the art approaches rely on automatically built social media data collections to detect irony using a variety of features gleaned from the utterance-internal context going from n-gram models, stylistic, to dictionary-based features (Burfoot and Baldwin, 2009; Davidov et al., 2010; Tsir et al., 2010; Gonzalez-Ibáñez et al., 2011; Liebrecht et al., 2013; Joshi et al., 2015; Hernández Farías et al., 2015). In addition to the above more lexical features, many authors point out the contribution of pragmatic features, such as the use of common vs. rare words or synonyms (Barbieri and Saggion, 2014). Recent work explores other kinds of contextual information like author profiles, conversational threads, or querying external sources of information (Bamman and Smith, 2015; Wallace et al., 2015; Karoui et al.,
Compared to automatic irony detection, little efforts have been done on corpus-based linguistic study of irony. Most of these efforts focus on analyzing the impact of irony in feeling expressions and emotions, by manually annotating tweets at both sentiment polarity and irony levels. E.g. Van Hee et al. (2016) distinguish between ironic, possibly ironic, and non-ironic tweets in English and Dutch. For ironic statements, polarity change that causes irony was annotated to specify whether the change comes from an opposition explicitly marked by a contrast between a positive situation and a negative one, an hyperbole, or an understatement. Stranisci et al. (2016) recently extend the Italian Senti-TUT schema (cf. Section 2) to mark the aspects of the topic being discussed in the tweet, as well as the sentiment expressed towards each aspect. Bosco et al. (2016) propose a second extension with the annotation of French tweets using three labels: positive irony, negative irony, and metaphorical expression. Current state of the art corpus-based studies are mainly oriented to a sentiment analysis perspective on irony, focusing almost exclusively on capturing tweet’s overall sentiment, explicit polarity change, or syntactic irony patterns. We argue in this paper that irony should instead be an object of study by its own by proposing a more linguistic perspective in order to provide a deeper inspection of what are the inferential mechanisms that activate irony, either explicit or implicit, and the correlations between irony types and irony markers. As far as we know, this is the first study that investigates the portability of a wide-range of pragmatic devices in the interpretation of irony to social media data from a multilingual perspective.

### 8 Exploiting the annotated corpus for automatic irony detection

The French and Italian parts of the annotated corpus have been respectively exploited as datasets for the first irony detection shared tasks DEFT@TALN2017 and for the SENTIPOLC@Evalita shared task on irony detection in both 2014 and 2016 editions (Basile et al., 2014; Barbieri et al., 2016). In particular, currently only the first layer of the annotation scheme has been

---

Table 5: Markers in ironic (Exp or Imp) and non-ironic (NI) tweets in French, English and Italian (in %). Markers with an * have not been studied in irony literature.

<table>
<thead>
<tr>
<th>Negation</th>
<th>Discourse</th>
<th>Humour #a</th>
<th>Intensifier</th>
<th>Punctuation</th>
<th>False prop.a</th>
<th>Surprise</th>
<th>Modality</th>
<th>Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Im</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

---

Table 6: Percentage of tweets in each ironic category containing markers in French, English and Italian.

<table>
<thead>
<tr>
<th>Negation</th>
<th>Discourse</th>
<th>Humour #a</th>
<th>Intensifier</th>
<th>Punctuation</th>
<th>False prop.a</th>
<th>Surprise</th>
<th>Modality</th>
<th>Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Im</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

---

1 https://deft.limsi.fr/2017/

2 http://di.unito.it/sentipolc16
exploited aiming at detecting if a given tweet is ironic or not. The French task is ongoing. For what concerns Italian, in Sentipolc the irony detection task is one three related but independent sub-tasks focusing on subjectivity, polarity and irony detection, respectively. All tweets of the campaign are, therefore, annotated by a multi-layered annotation scheme including tags for all the three dimensions and available on the Task’s website. In 2016 SENTIPOLC has been the most participated EVALITA task with a total of 57 submitted runs from 13 different teams. Not surprisingly, results of the 12 systems evaluated for irony detection seem to suggest that the task appears truly challenging. However, organizers observe that its complexity does not depend (only) on the inner structure of irony, but on unbalanced data distribution in Sentipolc (1 out of 7 examples is ironic in the training set, as they reflect the distribution in a realistic scenario) and on the overall availability of a limited amount of examples (probably not sufficient to generalise over the structure of ironic tweets). The plan is to organize an irony detection dedicated task including a larger and more balanced dataset of ironic tweets in future campaigns. In this perspective, it will be also interesting to investigate if the finer-grained annotation layers for irony proposed here can have a role in the annotation scheme proposed for the new task data.

9 Conclusion and future work

In this paper, we proposed a multi-layered annotation scheme for irony in tweets and a multilingual corpus-based study for measuring the impact of pragmatic phenomena in the interpretation of irony. The results show that our schema is reliable for French and that it is portable to English and Italian, observing relatively the same tendencies in terms of irony categories and markers. We observed correlations between markers and ironic/non-ironic classes, between markers and irony activation types (explicit or implicit) and between markers and irony categories.

These observations are interesting in a perspective of pragmatically and linguistically informed automatic irony detection, since it brings out the most discriminant features. On this line, we plan to accomplish a validation of the schema based on the definition of an automatic classification model built upon such annotated features. Moreover, an interesting challenge could be to apply the annotation schema to a new language also less culturally close to those addressed in this work.

Finally, another perspective is to investigate how the application of our schema can contribute to shed light on the issue of distinguishing between irony and sarcasm. This issue is challenging, and only recently addressed from computational linguistics. In particular, new data-driven arguments for a possible separation between irony and sarcasm emerged from recent work on Twitter data (Sulis et al., 2016). It could be interesting to see the relation between the finer-grained and pragmatic phenomena related to irony investigated in the present study and the higher-level distinction between irony and sarcasm.

References


Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland, June. ACL.


Christine Liebrecht, Florian Kunnenman, and Antal Van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June. Association for Computational Linguistics.


