TWITTIRÒ: an Italian Twitter Corpus with a Multi-layered Annotation for Irony

Alessandra Teresa Cignarella*
Università degli Studi di Torino
Universitat Politècnica de València

Cristina Bosco**
Università degli Studi di Torino

Viviana Patti†
Università degli Studi di Torino

Mirko Lai‡
Università degli Studi di Torino
Universitat Politècnica de València

Provided the difficulties that still affect a correct identification of irony within the context of Sentiment Analysis tasks, in this paper we describe the main issues emerged during the development of a novel resource for Italian annotated for irony. The project mainly consists in the application on the Twitter corpus TWITTIRÒ of a multi-layered scheme for the fine-grained annotation of irony, as proposed in a multilingual setting and previously applied also on French and English datasets (Karoui et al. 2017). In applying the annotation on this corpus, we outline and discuss the issues and peculiarities emerged about the exploitation of the semantic scheme for Twitter textual messages in Italian, thus shedding some lights on the future directions that can be followed in the multilingual and cross-language perspective too. We present, in particular, an analysis of the annotation process and distribution of the labels of each layer involved in the scheme. This is supported by a discussion of the outcome of the annotation carried on by native Italian speakers in the development of the corpus. In particular, an in-depth discussion of the inter-annotator agreement and of the sources of disagreement is included. The result is a novel gold standard corpus for irony detection in Italian, which enriches the scenario of multilingual datasets available for this challenging task and is ready to be used as a benchmark in automatic irony detection experiments and evaluation campaigns.

1. Introduction

The recognition of irony and the identification of pragmatic and linguistic devices that activate it are known as very challenging tasks to be performed by both humans or automatic tools (Mihalcea and Pulman 2007; Reyes, Rosso, and Buscaldi 2010; Kouloumpis,

* Dipartimento di Informatica, Università degli Studi di Torino
  PRHLT Research Center, Universitat Politècnica de València
  E-mail: cigna@di.unito.it

** Dipartimento di Informatica, Università degli Studi di Torino
  E-mail: bosco@di.unito.it

† Dipartimento di Informatica, Università degli Studi di Torino
  E-mail: patti@di.unito.it

‡ Dipartimento di Informatica, Università degli Studi di Torino
  PRHLT Research Center, Universitat Politècnica de València
  E-mail: lai@di.unito.it

Wilson, and Moore 2011; Maynard and Funk 2011; Reyes, Rosso, and Buscaldi 2012). Considering that the more promising tools for irony detection apply supervised approaches, it is becoming progressively crucial the development of linguistic resources where they can be trained and tested.

The main aim of the work here described is the creation of a currently missing resource, that is an Italian corpus annotated for irony, i.e. TWITTIRÒ, to be used as benchmark for systems addressing the irony detection task within evaluation campaigns of NLP tools for Italian (see the shared task IronITA proposed at EVALITA 2018 (Cignarella et al. 2018b)). Nevertheless, in order to lay the foundation for future comparisons with other languages and resources, we based the annotation on a scheme designed within the context of a multilingual project and described in (Karoui et al. 2017). Considering the complexity of the phenomenon of irony, often described in literature (Grice 1975, 1978; Sperber and Wilson 1981; Wilson and Sperber 2007), this scheme includes different layers and allows a fine-grained description of the addressed phenomenon. It has been successfully applied on French and English corpora, but also on a small dataset for Italian, which can be seen as the preliminary stage of the work presented in (Cignarella, Bosco, and Patti 2017).

The analysis of the disagreement among the native Italian speakers involved in the annotation of the Italian resource confirms that the fine-grained annotation of irony is an especially challenging task, due to aspects related to the subjectivity of the involved annotators. Human annotators, even skilled or domain experts, are always connected to their individual experiences, their individual sense of humor and a certain situational context. Nevertheless, even if they are biased, humans can easily detect the presence of irony when it occurs also in quite early stages of their life. The investigation that we propose here is at a deeper level and concerns the linguistic devices known in pragmatics as signals of irony and their relevance for modeling irony in social media in a computational perspective. Such finer-grained annotation results to be more challenging, also due to some peculiarities of the micro-blog textual genre. Our analysis of the disagreement on the developed corpus aims at shedding some light on the difficulties related to the task.

The paper is organized as follows. The next section surveys the main contribution in this area, while Section 3 describes the collection of the novel dataset for Italian and its organization in different sub-corpora which represent different kinds of Twitter texts. The scheme is the object of Section 4, where each label applied in the annotation is associated to several examples extracted from the different sections of the corpus. Section 5 includes the analysis of the annotation process and it is mainly focused on the distribution of the labels and the analysis of the disagreement among the annotators. The final section summarizes the discussion and the lessons learned.

2. Related work

Communications in social media platforms like Twitter include a significant percentage of linguistic devices for figurative language, such as irony and sarcasm (Davidov, Tsur, and Rappoport 2010; González-Ibáñez, Muresan, and Wacholder 2011; Reyes, Rosso, and Veale 2013), and it has been noticed that this negatively affects the performances of systems for automatic detection of sentiment in microblogging data, which experienced the phenomenon of wrong polarity classification of ironic messages (Bosco, Patti, and Bolioli 2013; Ghosh et al. 2015). Indeed, the presence in a text of ironic devices can flip the polarity of an opinion expressed with positive words to the intended negative meaning – or (rarely) vice versa – working as an unexpected polarity reverser. This can

undermine systems' accuracy. The automatic detection of irony is, therefore, crucial for the development of irony-aware sentiment analysis systems. At the same time it is also an interesting conceptual challenge from a cognitive point of view and it can help to shed some light on how human beings use irony as a communicative tool.

The automatic recognition of irony in texts is, still nowadays, a challenging task to be performed both by human annotators and automatic NLP systems. Different approaches have been developed, the majority of which take advantage only of the textual content itself, since other paralinguistic cues, like e.g. tone or corporal movements, are not considered. Because of its larger availability, Twitter is the most widely used source of samples of ironic texts, which are easy to be collected relying on the behavior of Twitter users, who often explicitly mark their ironic messages by using hashtags such as *#irony* or *#sarcasm*. The pretty good reliability of the user-generated hashtags as golden labels for irony has been experimentally confirmed by some studies (Kunneman et al. 2015). Moreover, it seems that, due to the interaction model underlying the micro-blogging platform, irony expressed here could be somehow easier to analyze. Indeed, Twitter users have to be sharp and short, having only 140 characters for expressing their comments, and most of the times the ironic posts do not require knowledge about the conversational context to be understood. Several works have been carried out using tweets for experimental purposes (Davidov, Tsur, and Rappoport 2010; González-Ibáñez, Muresan, and Wacholder 2011; Reyes, Rosso, and Veale 2013; Wang 2013; Riloff et al. 2013; Barbieri, Saggion, and Ronzano 2014; Ptáček, Habernal, and Hong 2014; Hernández Farías, Benedí, and Rosso 2015; Rajadesingan, Zafarani, and Liu 2015; Bamman and Smith 2015; Joshi, Sharma, and Bhattacharyya 2015; Karoui et al. 2015b). Furthermore, there are some efforts in other social media such as customer reviews from Amazon (http://www.amazon.com/) (Filatova 2012; Buschmeier, Cimiano, and Klinger 2014); comments from the online debate sites such as 4forums.com (http://www.4forums.com/political/) (Abbott et al. 2011; Lukin and Walker 2013) and, recently, Reddit (http://www.reddit.com) (Wallace, Choe, and Charniak 2015).

In most of the studies relying on Twitter data, irony detection has been modeled as a binary classification problem, where mostly tweets labeled with certain hashtags (i.e. *#irony*, *#sarcasm*, *#sarcastic*, *#not*) have been considered as ironic utterances. Following this framework, different approaches have been proposed (Davidov, Tsur, and Rappoport 2010; González-Ibáñez, Muresan, and Wacholder 2011; Reyes, Rosso, and Veale 2013; Riloff et al. 2013; Barbieri, Saggion, and Ronzano 2014; Ptáček, Habernal, and Hong 2014; Hernández Farías, Benedí, and Rosso 2015). The authors proposed models that exploit mainly features related to textual-content such as: punctuation marks, emoticons, part-of-speech labels, discursive terms, specific patterns (e.g., according to (Riloff et al. 2013), a common form of sarcasm in Twitter consists of a positive sentiment contrasting with a negative situation, among others.

Another key characteristic for irony is *unexpectedness* (Attardo 2000). According to many theoretical accounts people infer irony when they recognize an incongruity between an utterance and what is known (or expected) about the speaker and/or the environment. This is something that can be referred to as the pragmatic context. Recent approaches started to address such issue, taking into account information about context (Rajadesingan, Zafarani, and Liu 2015; Bamman and Smith 2015; Wallace, Choe, and Charniak 2015).

For what concerns the affective information, some approaches already use in their models some kind of sentiment and emotional information. (Reyes, Rosso, and Veale 2013) included in their model some features to characterize irony in terms of elements

related to sentiments, attitudes, feelings and moods exploiting the Dictionary of Affect in Language (DAL) proposed by (Whissell 2009). Barbieri et al. (Barbieri, Saggion, and Ronzano 2014) considered the amount of positive and negative words by using Senti-WordNet. (Hernández Farías, Benedí, and Rosso 2015) exploited two widely applied sentiment lexicons (Hu&Liu and AFINN) as features in their model. Recent works focused specifically on studying the role of affective information in a comprehensive manner, by exploring the use of a wide range of lexical resources available for English, reflecting different aspects of a multi-faceted phenomenon (Hernández Farías, Patti, and Rosso 2016).

Few preliminary studies address the task to investigate the differences between irony and sarcasm, which can also be interesting in order to reason on the possibility to observe different polarity reversal patterns behind the two figurative devices. A contribution on this line is given in (Sulis et al. 2016), where authors analyze messages explicitly tagged by users with #*irony*, #*sarcasm* and #*not* in order to test the hypothesis to deal with different linguistic phenomena, with a special focus on the role of features related to the multi-faceted affective information expressed in such texts. There are also other figurative devices, like metaphor, where we can observe that the sentiment polarity of the literal meaning differs from that of the intended figurative meaning (Ghosh et al. 2015). Metaphorical expressions represent a great variety, ranging from conventional metaphors, to poetic and largely novel ones (Shutova 2010). The use of metaphor can sometimes also contribute to give an ironic twist to a sentence and, being pervasive also in social media, represents a further serious challenge for sentiment analysis systems working on this kind of texts.

The majority of the research in irony detection has been addressed in English, although there is some research in other languages, such as: Dutch (Kunneman et al. 2015), Italian (Bosco, Patti, and Bolioli 2013), Czech (Ptáček, Habernal, and Hong 2014), French (Karoui et al. 2015b), Portuguese (Carvalho et al. 2009) and Chinese (Tang and Chen 2014).

The growing interest on this task is attested by the proposal of shared tasks focusing on irony detection and its impact on sentiment analysis in social media, in the context of periodical evaluation campaigns for NLP tools for many languages, see for instance the pilot task on irony detection proposed for Italian in *Sentipolc@Evalita*, in the 2014 and 2016 editions (Basile et al. 2014; Barbieri et al. 2016) and the battery of related tasks proposed for French at *DEFT@TALN2017* (Benamara et al. 2017).

For what concerns English, after a first task at *SemEval-2015* (i.e. *Task 11*) focusing on *Sentiment Analysis of Figurative Language in Twitter* (Ghosh et al. 2015), in 2018 a shared task on irony detection in tweets has been proposed (*SemEval-2018 Task 3: Irony detection in English tweets*)[1]. In the latter, the organizers propose not only the classical binary classification task, where the systems must determine whether a tweet is ironic or not, but also a fine grained multi-class classification task on different types of irony, where the systems must predict one out of four following labels: i) verbal irony realized through a polarity contrast, ii) verbal irony without such a polarity contrast, iii) descriptions of situational irony, and iv) non-irony (Van Hee, Lefever, and Hoste 2018). The setting proposed for the Semeval-2018 is an indication of the growing interest for a deeper analysis of the linguistic phenomena underlying ironic expressions. Such kind of deeper analysis naturally calls for the definition and the exploitation of schemes allowing the

---

[1] https://competitions.codalab.org/competitions/17468

annotation of finer-grained features and resources in order to hopefully improve the performance of automatic systems in this especially challenging task.

With the description and analysis of the TWITTIRÒ corpus we aim at providing a deeper investigation of the issues that arose with the application of the scheme to Italian irony-laden texts, which has been preliminary investigated in (Karoui et al. 2017; Cignarella, Bosco, and Patti 2017). The resulting annotated corpus has been exploited as reference dataset within the context of the IronITA shared task at the EVALITA evaluation campaign[2]. It will be made available to the community and exploitable in the cross- and multi-lingual perspective depicted in (Karoui et al. 2017) from the end of 2018 [3].

3. Data collection

In this section we describe the methodology applied in the collection of tweets, and the internal structure of the dataset.

The methodology applied in the collection of the French and English datasets, which are part of the same multilingual project on irony detection where also TWITTIRÒ collocates, couldn't be applied for Italian. For these other languages indeed tweets were retrieved by using Twitter APIs and filtered through specific *hashtags* exploited by users to self-mark their ironic intention (*#irony*, *#sarcasm*, *#sarcastic*). The exploitation by Italian users of a series of humorous hashtags, but no long-term single hashtag established and shared among them, doesn't pave the way for the application of this methodology for data collection.

Nevertheless, in the last few years several Italian corpora from Twitter, where the presence of irony is marked, have been made available. We extracted from them tweets[4] to be included in TWITTIRÒ according to the distribution presented in Table 1.

• **TW-SPINO** is a portion of SENTI-TUT (Bosco, Patti, and Bolioli 2013) which contains tweets collected from the satirical blog *Spinoza.it*. The language used is grammatically correct and featured by a high register and style, while the topics are variegate with a clear preference for jokes concerning the world of politics and general news.

1. SOURCE: TW-SPINO
   Pubblicata la classifica mondiale della libertà di stampa. Non possiamo dirvi altro. [giga]
   (The world ranking for freedom of printing competition has been published. We cannot say anything else. [giga])

• **TW-SENTIPOLC14** (Basile et al. 2014) contains tweets generated by common users and therefore it is less homogeneous than TW-SPINO, with a frequent use of creative hashtags, mentions, repetitions of laughters. We selected here the political tweets with reference to the government of Monti between 2011 and 2012.

2. SOURCE: TW-SENTIPOLC14
   Mario Monti? non era il nome di un antipasto? #FullMonti #laresadeiconti #elezioni #308.
   (Mario Monti? Wasn't it the name of an entree? #FullMonti #laresadeiconti #elezioni #308.)

• **TW-BS** (Stranisci et al. 2015, 2016) contains tweets on the debate of the reform of Italian School "Buona Scuola". Devices typically exploited in computer-mediated communica-

---

2 `http://di.unito.it/ironita18`
3 `https://github.com/IronyAndTweets/Scheme`
4 A portion of these tweets (400 messages) has already been exploited and analyzed in (Karoui et al. 2017).

tion (CMC) are shown and, being the reform of the education system a highly criticized one, the use of sentences written in ALL CAPS (to decode shouting) is wide.

3. SOURCE: TW-BS
   @fattoquotidiano Quest'anno è peggio del solito: oltre all'amianto c'è anche #labuonascuola.
   (@fattoquotidiano This year is worse than usual: in addition to asbestos there is also #labuonascuola.)

Table 1
Distribution of TWITTIRÒ's tweets across the three different source datasets.

Source corpus	Number of tweets
TW-SPINO	378
TW-SENTIPOLC14	527
TW-BS	519
TWITTIRÒ	**1,424**

The different styles and composition of the three sub-corpora TW-SPINO, TW-SENTIPOLC14, and TW-BS reveal interesting research paths to be followed. We will, therefore, address this issue in Section 5.1.

4. A Multi-layered Annotation Scheme

The main goal of the scheme proposed in (Karoui et al. 2017) and applied in TWITTIRÒ is to provide a fine-grained representation of irony inspired by the issues raised in literature about this topic. For achieving this goal, the scheme includes four different levels of annotation organized as follows.

LEVEL 1: CLASS.
It concerns the classification of tweets into **ironic** or **not ironic**, but it does not apply in principle to our case where the corpus only includes ironic tweets because of the methodology applied in collection.

LEVEL 2: ACTIVATION TYPE.
As stated from various linguistic theories (Grice 1975; Sperber and Wilson 1981; Clark and Gerrig 1984), irony is often exhibited through the presence of a clash or a contradiction between two elements. In tweets, these elements, henceforth named P1 and P2, can be found both as two lexicalized clues belonging to the internal context or can be one in the utterance and the other outside, as part of some pragmatic context external to the tweet. According to (Karoui et al. 2015a), we annotate the **activation type** such that, if the contradiction relies exclusively on the lexical clues internal to the utterance, as EXPLICIT, while if the contradiction that combines lexical clues with an additional pragmatic context external to the utterance, as IMPLICIT.

Explicit contradiction:
It can involve a contradiction between proposition P1 and proposition P2 that have e.g. opposite polarities, like in the example below where the opposition is between *liberate* (freed) and *processate* (processed).

4. SOURCE: TW-SPINO
   [**Liberate**]$_{P1}$ Greta e Vanessa. Saranno [**processate**]$_{P2}$ in Italia. [@maurizioneri79]
   *(Greta and Vanessa have been [**freed**]$_{P1}$. They will [**undergo trial**]$_{P2}$ in Italy. [@maurizioneri79].)*

Implicit contradiction:
The irony occurs because the writer believes that his audience can detect the disparity between P1 and P2 on the basis of contextual knowledge or common background shared with the writer.

5. SOURCE: TW-SPINO
   ["Se davvero abbiamo pagato è uno schifo" ha detto Salvini guardando la laurea di Renzo Bossi.]$_{P1}$ [faro]
   ("It's a shame that we really paid for this" said Salvini, looking at Renzo Bossi's Masters degree. [faro])
   $\rightarrow P_2$: Renzo Bossi got his Master's degree by paying with his father's party money (Lega Nord).

6. SOURCE: TW-BS
   La [buona scuola e le **sillabe**]$_{P1}$ - http:t.conS42fRjAKp
   (The buona scuola and the syllables - http:t.conS42fRjAKp)
   $\rightarrow P_2$: The official document that presented the school reform had several hyphenation mistakes.

There are cases in which irony is activated in multiple ways inside a tweet. It might occur that on one superficial layer irony is explicitly activated from lexicalized cue words, and on a second "hidden" layer there is a deeper level of irony, inferable only through additional pragmatic knowledge. In this case, the tweet has to be annotated as ironic and the activation type as IMPLICIT.

LEVEL 3: CATEGORIES.
Both forms of contradictions can be expressed through different rhetorical devices, patterns or features that are grouped under different labels (i.e. analogy, euphemism, false assertion, oxymoron/paradox, context shift, hyperbole, rhetorical question and other).

Analogy:
In this category are summoned also other figures of speech that comprehend mechanisms of comparison, such as *simile* and *metaphor*. In (7) an analogy is drawn between the footballer Lionel Messi and the Italian minister Maria Elena Boschi because of their authoritarian fathers. **Pure** (furthermore) serves as clue word. In (8) the analogy is activated by the construction "*x **reminds me of** y*". Instead, in (9) the clue words are "*there is x and **also** y*".

7. SOURCE: TW-SPINO
   Leo Messi: "Firmo quello che mi dice papà". Pure la Boschi. [notturnoconcertante]
   (Leo Messi: "I sign what daddy tells me". Also Minister Boschi. [notturnoconcertante])

8. SOURCE: TW-SENTIPOLC14
   Il governo #Monti mi ricorda la corazzata kotiokmin.
   (Monti's government reminds me of the kotiokmin battleship.)

9. SOURCE: TW-BS
   @fattoquotidiano Quest'anno è peggio del solito: oltre all'amianto c'è anche

#labuonascuola.
(@fattoquotidiano This year is worse than usual: other than asbestos there is also #labuonascuola.)

Euphemism:
It is a figure of speech which is used to reduce the facts of an expression or an idea considered unpleasant in order to soften the reality. In (10) is exploited the use of the partitive expression "*a few/a little*". Instead in (11) irony through euphemism is activated by the expression **senza fretta** (no haste). In (12) is used the common device of punctuation such as quotations to soften one's way to express their own opinion.

10. SOURCE: TW-SPINO
    Nel 2006 Charlie Hebdo aveva pubblicato delle vignette satiriche su Maometto. Ci hanno messo un po' a capirle. [nicodio]
    (In 2006 Charlie Hebdo published some satirical cartoons about Muhammad. It took them a while to understand them. [Nicodio])

11. SOURCE: TW-SENTIPOLC14
    @palazzochigi Professor Monti, c'è una certa attesa per la lista #ministri. Senza fretta, però tenga conto della nostra ansia...
    (@palazzochigi Professor Monti, there is a certain wait for the list of #ministers. No haste, but please take into account our anxiety ...)

12. SOURCE: TW-SENTIPOLC14
    """ buona scuola """.
    (""" good school """)

False assertion (implicit only):
Indicates that a proposition, fact or an assertion fails to make sense against the reality. The speaker expresses the opposite of what he thinks or something wrong with respect to a context. External knowledge is fundamental to understand the irony (it is, in fact, implicit only). In the following examples, the sentences written in correspondence of the right arrow are propositions, which are not lexicalized, but must be inferred from the reader to understand irony.

13. SOURCE: TW-SPINO
    "Potrei non opporre veti a un presidente del Pd", ha detto Berlusconi iscrivendosi al Pd. [CONTINUA su http://t.co/oDPUtxTq-U7]
    ("I could not oppose vetos to a the PD President", said Berlusconi, while subsctibing to the PD. [CONTINUES on http://t.co/oDPUtxTq-U7])
    $\longrightarrow$ Berlusconi never subscribed to the PD party.

14. SOURCE: TW-SPINO
    Vedo che c'è molta disinformazione sul referendum del 17 maggio. [@MisterDonnie13]
    (I see there is a lot of misinformation on the referenum of May, 17th. [@MisterDonnie13])
    $\longrightarrow$ The referendum is, in fact, on April, 17th, not May.

15. SOURCE: TW-SENTIPOLC14
    Totoministri per il governo Monti: Gelmini ai lavori pubblici, farà il tunnel dei neutrini!
    (Minister betting for the Monti government: Gelmini to public works, will make the neutrino tunnel!)
    $\longrightarrow$ Minister Gelmini was never in charge of public work administration. It is also a reference to an erroneous statement about neutrinos that the Minister had previously uttered.

Oxymoron/paradox (explicit only):
This category is equivalent to the category FALSE ASSERTION except that the contra-

diction, this time, is explicit. Also in this subsection, the sentences written in correspondence of the right arrow are propositions, which are not lexicalized, but must be inferred from the reader to understand irony.

16. SOURCE: TW-SPINO
Legge elettorale, il Pd si divide. Non vedono l'ora di provarla. [@maurofodaroni]
(Electoral law, the Democratic Party is divided. They can not wait to try it. [@maurofodaroni])
⟶ It is absurd to think that the Italian political party PD has undergone an internal division in order to try a new electoral law they promoted.

17. SOURCE: TW-SPINO
Individuata una mafia tipicamente romana. Prima di mezzogiorno non prendeva appuntamenti.
(Identified a typically Roman mafia. Before noon it did not take appointments.)
⟶ It is common knowledge that people from Rome are often late, thus the paradox of creating a criminal organization that is also often late.

18. SOURCE: TW-SENTIPOLC14
Brunetta sostiene di tornare a fare l'economista, Mario Monti terrorizzato progetta di mollare tutto ed aprire un negozio di pescheria.
(Brunetta claims to go back to being an economist, Mario Monti terrified plans to give up everything and open a fishmonger's shop.)
⟶ It is paradoxical that the Italian Prime Minister would leave the government to open a fishmonger's.

Context shift (explicit only):
It occurs by the sudden change of the topic/frame in the tweet, as in (7), where the first sentence is about pupils' pocket money, while the second is about the price of ice cream. The same happens in (20), where the first clue word is about a Roma encampment, while the second about a safari journey.

19. SOURCE: TW-BS
@matteorenzi Più che la **#labuonascuola** direi #carascuola visto che ci vogliono più di 800 euro a pischello....quasi quanto **5 kg di gelato**
(@matteorenzi More than the#labuonascuola I'd say #carascuola being that more than 800 euros are needed for each kid....almost like 5 kilograms of ice-cream.)

20. SOURCE: TW-SPINO
L'auto di Salvini assalita al **campo rom**. Rovinato il **safari**. [@paniruro]
(Salvini's car assaulted at the Roma camp. The safari is runined. [@paniruro])

Hyperbole/exaggeration:
It is a figure of speech which consists in expressing an idea or a feeling with an exaggerated way. It can be expressed through the use of superlative adjectives such as in (21) or (22). Either with the use of hyperbolic expression as **aberrazione** (aberration) in (23), or indefinite collective adjectives (or pronouns) such as **tutti** (everyone) in (24).

21. SOURCE: TW-SPINO
Il trionfo di Tsipras in Grecia è il più grande successo della sinistra italiana dai tempi di Zapatero in Spagna. [@gmbugs]
(Tsipras' triumph in Greece is the greatest success of Italian Left from the times of Zapatero in Spain. [@gmbugs])

22. SOURCE: TW-BS
#labuonascuola "Sarà bellissimo ascoltare la voce di tutti,...." @matteorenzi la senti questa voce?....(cit. coro da stadio)
(#labuonascuola "It will be great to listen everyone's voice...." @matteorenzi do you hear this

voice?....(cit. football chant))

23. SOURCE: TW-SENTIPOLC14
@masechi Si è già assistito a Porta a Porta alla simulazione di un governo Monti con ministri La Russa e Bindi. Aberrazione audiovisiva.
(@masechi At Porta a Porta we have already seen a simulation of Monti's government with La Russa and Bindi ministers. Audiovisual aberration.)

24. SOURCE: TW-BS
#M5S #Renzi, se tra un anno non ci saranno 170 mila insegnanti di ruolo in più, te li porto tutti a @Palazzo_Chigi #labuonascuola.
(#M5S #Renzi, if in a year there will not be 170 thousand teachers more, I'll take them all to @Palazzo_Chigi #labuonascuola.)

Rhetorical question:

It is a figure of speech in the form of a question asked in order to make a point rather than to elicit an answer. It can be direct and explicit as in (25) and in (26). Or it can be an indirect rhetorical question as in (27)

25. SOURCE: TW-SPINO
Giovanardi applaude la sentenza Cucchi. <u>Cosa vi aspettavate da una frase che inizia con "Giovanardi"</u>? [CONTINUA su http://t.co/oDPUtx2DvV].
(Giovanardi applauds the Cucchi ruling. What did you expect from a sentence that starts with "Giovanardi"? [CONTINUES on http://t.co/oDPUtx2DvV].)

26. SOURCE: TW-SENTIPOLC14
Mario Monti? <u>non era il nome di un antipasto</u>? #FullMonti #laresadeiconti #elezioni #308.
(Mario Monti? Wasn't it the name of an entree? #FullMonti #laresadeiconti #elezioni #308.)

27. SOURCE: TW-BS
#renzi blocca gli stipendi per 10 anni, noi dobbiamo fare #labuonascuola <u>non so se sanno che anche gli insegnanti hanno una famiglia</u>.
(#renzi blocks the salaries for 10 years, we have to do #labuonascuola I don't know if they are aware that also teachers do have a family.)

Other:

This last category represents ironic tweets, which can not be classified under one of the other seven previously described categories. It can occur, for example, in case of humor or situational irony. It is also applied when there is a number of overlapping categories, and thus, it is hard to define which one should be tagged first. In some cases this category is in practice a way for introducing hints about the presence of a class that is not included in our current schema but can be added in the future, e.g. pun or alliteration.

In (28), for example, we notice a pun in bold. In fact, *larga intesa* is a common collocation in Italian to describe a good shared agreement, while *BANCA INTESA* is a well known Italian credit institution. Furthermore, the graphic device of caps lock is used. Instead, in (29) irony is expressed through the use of another type of pun, exploiting assonance "buono-scuola" (monetary ticket used to buy books or office products) vs. "buona scuola" (the reform).

28. SOURCE: TW-SENTIPOLC14
Il Governo Monti parte....c'è **larga** BANCA **INTESA**. http://t.co/x0-u6nt7b.
(The Monti Government starts .... there is a large BANCA INTESA. http://t.co/x0-u6nt7b.)

29. SOURCE: TW-BS

    Dal buono-scuola alla buona scuola renziana: buone ragioni per lo sciopero http://t.co/kduFO6MLdd #10_ottobre #quota_96 #sciopero.
    (From the school vouchers to Renzi's good school: good reasons for the strike http://t.co/kduFO6MLdd # 10_ottobre #quota_96 #sciopero.)

30. SOURCE: TW-SPINO

    L'Euro entrò in vigore l'1 gennaio 2002. Trasformandolo immediatamente nel 2 gennaio 4004. [CONTINUA su http://t.co/pJhE2CMi70]
    (Euro came into force the 1st January 2002. Immediately turning it into 2nd January 4004. [CONTINUES on http://t.co/pJhE2CMi70])

31. SOURCE: TW-SPINO

    Sicilia, arriva barcone di migranti e a bordo c'è anche un gatto. Vengono a rubarci i nostri like. [@LughinoViscorto]
    (Sicily, a boat of migrants arrives and there is also a kitten aboard. They come here to steal our likes. [@LughinoViscorto])

In the tweet (30) we have the overlapping of a *false assertion* (what it is stated is not true), an *hyperbole* (the mathematical doubling of a date) and the *need of external knowledge* (prices doubled in all Europe with the introduction on the new communitarian currency).

In the example (31) is shown the application of racism, trough the exploitation of a shared knowledge, almost an idiomatic expression: "*foreigners come here and they steal X from us*". Furthermore, to understand a second layer of irony, the additional knowledge of social media is needed: cats are a subject that makes people on the net turn crazy, and hence, they receive a big amount of LIKES (Facebook jargon).

LEVEL 4: CLUES.

Clues represent words that can help annotators to decide in which category belongs a given ironic tweet, such as **like** for analogy, **very** for hyperbole/exaggeration. Clues include also negation words, emoticons, punctuation marks, interjections, named entity (and mentions). Since the extraction of the information about this level can be done, to a great extent by automatic tools, we did not addressed this specific task by manual annotation.

5. Annotation and Disagreement

In this section we describe the steps of the annotation phase. We first propose a brief discussion on the distribution of tags on the three sub-corpora. Then, we address the issues that arose within the disagreement analysis and we discuss the results.

5.1 Annotation Process and Label Distribution

The application of the scheme involved three skilled annotators, henceforth named as *A1*, *A2* and *A3*. While the first two have independently applied the schema from scratch to the whole dataset, the work of the third one has been used to solve the disagreement detected by comparing the first two independent annotations.

The annotation process allowed the achievement of an agreement for 1,024 tweets out of the 1,200 tweets collected and annotated. Together with the 400 tweets analyzed in (Karoui et al. 2017) they can be considered as a novel gold standard for Italian, that is the TWITTIRÒ corpus, consisting of almost 1,500 tweets in total. Taking into consideration

the diverse nature of the three sub-corpora which compose the TWITTIRÒ corpus, in the following we will present the resulting distribution of the labels.

Figure 1 shows the distribution according to the type of activation among the three sub-corpora (TW-SPINO, TW-SENTIPOLC-14, and TW-BS). As it can be seen, in the majority of tweets, irony is triggered by an EXPLICIT contradiction in all the three datasets (82% in TW-SPINO, 75% in TW-SENTIPOLC14, and 87% in TW-BS). This confirms the findings in (Karoui et al. 2017) on the first section of the corpus, which highlight that Italian massively displays a different behavior concerning the preference for EXPLICIT activation type. In fact, languages such as French and English seem to favor IMPLICIT activation type, as shown in (Karoui et al. 2017).
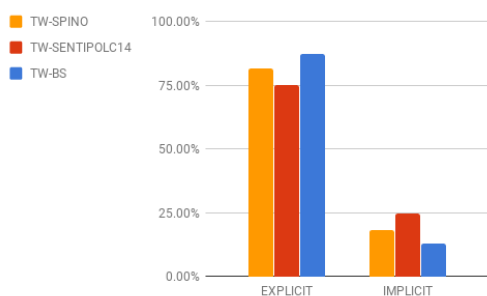
Figure 1
Distribution of types
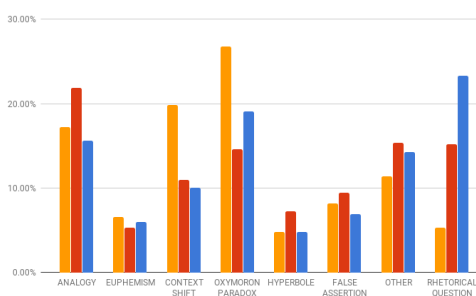
Figure 2
Distribution of categories

Figure 2 shows how the devices that trigger irony (i.e. labels for LEVEL 3) are distributed in the full corpus. As far as the distribution in the three sub-corpora is concerned, we observed some differences and in particular the higher frequency of OXYMORON/PARADOX (25%) and CONTEXT SHIFT (20%) in the TW-SPINO sub-corpus, thus suggesting that these devices are often exploited by the contributors of the Spinoza's blog for creating a sense of surprise. In TW-SENTIPOLC14 a higher frequency of ANALOGY features (21%) has instead been registered, while the TW-BS, composed by tweets about the reformation of the Italian School, contains a higher frequency of RHETORICAL QUESTION tags (22%) tags, which we can link to the dissatisfaction of the people for the proposed reform.

An extensive description of the three sub-corpora together with the distribution of categories featuring them is present in (Cignarella et al. 2018a), where the discussion also focuses on the correlations between categories and topics characterizing the tweets from the different collections.

5.2 Disagreement Analysis

The first steps of the disagreement analysis have been performed on the first tranche of the TWITTIRÒ corpus (1,200 tweets), and considering the effort of two independent annotators (A1 and A2). As far as the activation type of irony is involved, the inter-annotator agreement (IAA) between A1 and A2 for the labeling of IMPLICIT vs. EXPLICIT, calculated with Cohen's coefficient, is $\kappa = 0.41$ (Artstein and Poesio 2008), and the distributions of the labels of this layer for each annotator are reported in Table 2.

Instead as far as the annotation of the categories is involved, the IAA is slightly higher, $\kappa = 0.46$. Providing that our work is mainly focused on category tags, their exploitation

Table 2
Inter-annotator agreement on type tags

		A2		
		implicit	explicit	TOTAL
A1	implicit	**104**	136	240
	explicit	63	**897**	960
	TOTAL	167	1,033	**1,200**

and distribution, the following discussion will be mainly focused on the tweets where A1 and A2 were in disagreement and the need of the A3's annotation was required (577 tweets). As support, Table 3 shows the distribution of category tags exploited by A1 and A2.

Table 3
Inter-annotator agreement on category tags

		A2								
		analogy	euphemism	false assertion	oxymoron paradox	context shift	hyperbole	rhetorical question	other	TOTAL
A1	analogy	**131**	4	9	13	16	8	7	23	211
	euphemism	4	**33**	8	7	10	5	1	6	74
	false assertion	6	1	**53**	21	7	4	0	9	101
	oxymoron paradox	10	8	34	**121**	21	3	4	21	222
	context shift	9	2	4	31	**62**	8	2	14	132
	hyperbole	7	4	13	19	4	**29**	1	14	91
	rhetorical question	8	5	6	25	17	2	**127**	8	198
	other	19	7	22	10	16	4	3	**90**	171
	TOTAL	194	64	149	247	153	63	145	185	**1,200**

The analysis of the disagreement detected in this dataset supports the following ideas. Firstly, observing the tag distribution between A1 and A2 (see Table 3), the tag OXYMORON/PARADOX is the more frequently exploited, followed by ANALOGY, by both annotators. On the other hand EUPHEMISM and HYPERBOLE are the least exploited ones. To further validate our intuitions, we calculated the agreement of A1 and A2 on each category tag, through Cohen's kappa coefficient.
As it can be seen in Figure 3 the category tags OXYMORON/PARADOX ($\kappa = 0.40$) and FALSE ASSERTION ($\kappa = 0.36$) are among the three worst categories in agreement, together with HYPERBOLE ($\kappa = 0.34$). This means that even though annotators exploit the OXYMORON/PARADOX tag most of the times, they rarely agree on its correct application. The categories with the highest inter-annotator agreement are instead RHETORICAL QUESTION ($\kappa = 0.70$) and ANALOGY ($\kappa = 0.56$).

In general, as it is summarized in Table 4, annotators A1 and A2 reach a moderate agreement, $\kappa = 0.46$, on category tags. Moreover, it is interesting to see whether the
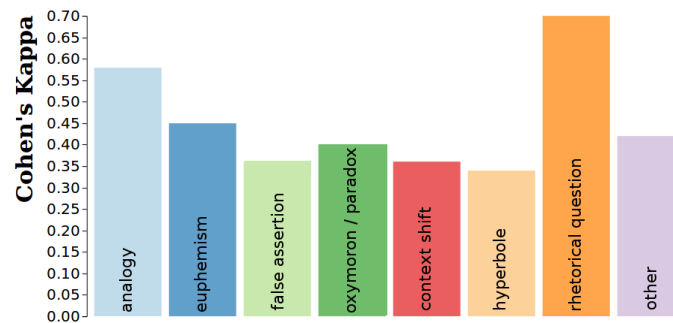
Figure 3
IAA between A1 and A2 on each category tag

value of kappa varies if we take into consideration each portion of the three sub-corpora composing TWITTIRÒ. In fact, the total average value is increased by the good IAA score of the annotation on the sub-corpus of TW-SPINO ($\kappa = 0.57$), and at the same time it is lowered by the poor IAA score obtained on the portion of TW-SENTIPOLC14 ($\kappa = 0.34$).

Our intuition is that the use of correct grammar, good writing style and punctuation, revised by the authors of the satirical blog *Spinoza.it*, improves the precision of the annotation. This fact does not apply to the sub-corpora of TW-SENTIPOLC14 and TW-BS which present a more heterogeneous shape and style.

Table 4
Cohen's kappa (A1 and A2) on each sub-corpus

n# of tweets	sub-corpus	IAA
1,200	ALL	$\kappa = 0.46$
300	TW-SPINO	$\kappa = 0.57$
312	TW-SENTIPOLC14	$\kappa = 0.34$
588	TW-BS	$\kappa = 0.47$

As we can see from Table 4 the inter-annotator agreement, calculated through Cohen's kappa coefficient, ranges from $\kappa = 0.57$ on the higher end of the spectrum and $\kappa = 0.34$ on the lower end of it. Considering that one of the main objectives of the present research is that of providing a new resource for the training of supervised machine leaning methods on Italian, a higher concordance of judgment between only two annotators, would be expected.

Competitions in NLP and Sentiment Analysis tasks tend to use different measuring systems (e.g. accuracy, precision, recall, F1-score, etc...) with respect to those used in annotations tasks in Linguistics (e.g. Cohen's kappa, Krippendorff's alpha, etc...). In order to compare, according to a same measure, the accuracy of humans and that of automatic systems, in Table 5 we provide a sort of accuracy calculated by considering one human annotator as an automatic system and the other as the gold standard.

As we already mentioned, also after the application of a third human independent annotation (A3) on 577 tweets, we still did not reach an agreement for classifying 176 tweets according to our scheme. Since a deeper observation of tweets in disagreement can give interesting information about the different nuances through which irony is

Table 5
Accuracy of A1 against A2

	TYPE	CATEGORY
A1 vs. A2	0.834	0.538

produced and about the complexity of the task, a further discussion on such hard cases is given in (Cignarella et al. 2018a).

Throughout a deeper analysis, the following main issues emerged. The choice between the category tags OXYMORON/PARADOX and FALSE ASSERTION seems to be strongly influenced by personal biases (see Table 3). For example, in the following tweet A1 tagged as explicit OXYMORON/PARADOX, while A2 as implicit FALSE ASSERTION:

32. Adesso ho capito perché ci son così pochi #presepi in giro. La gente ha paura che il #Governo #Monti faccia pagare l'#ICI anche su quelli...
    (Now I get why there are so few Christmas cribs around. People are worried that Monti will put a tax also on them...)

This is partially due to the fact that a relationship exists between the category FALSE ASSERTION (only IMPLICIT) and the category OXYMORON\PARADOX (only EXPLICIT). In fact, according to our multi-layered scheme for irony, those two categories, cover similar types of irony. Often the decision of LEVEL 3 (CATEGORIES) is triggered from the previous decision of LEVEL 2 (TYPES). For example, the category FALSE ASSERTION can be chosen only when we label the tweet as IMPLICIT. On the other hand, the category tags CONTEXT SHIFT and OXYMORON/PARADOX can occur only if Level 2 presents an EXPLICIT type of irony activation.

Another issue we want to address is that of the strong overlapping of RHETORICAL QUESTION with any other tag. As we can see from the following example, it is true that a rhetorical question is made, but the trigger of irony are the paradox and absurdity of the question itself.

33. Ma secondo voi super #Mario #Monti riuscirà a tassare anche la felicità?
    (What do you think, will Monti manage to put a tax also on happiness?)

The problem is caused by the fact that RHETORICAL QUESTION is a category tag that pertains to the linguistic level of pragmatics, which can coexist with semantic or lexical category tags such as ANALOGY or OXYMORON/PARADOX. An improvement in agreement could be that of allowing the presence of one or more categories at the same time.

6. Conclusions

Provided the increasing relevance of the identification of irony within the context of sentiment analysis tasks, in the present work we have described the application of a fine-grained annotation scheme for irony (Karoui et al. 2017). Focusing on the analysis of the disagreement and discussing it by providing several sample tweets, we conducted quantitative and qualitative study on the annotated data. A second point of interest has been the study of style and composition of the three sub-corpora TW-SPINO, TW-SENTIPOLC14, and TW-BS, revealing interesting paths to be followed in future work.

Our research definitely confirms how the task of irony detection is challenging, and how the application of finer grained schemes can contribute to shed some light on this phenomenon. Moreover, we provided a novel Italian resource which extends the scenario in (Karoui et al. 2017) and can be exploited in future experiments on irony detection in a multilingual perspective[5].

Within the 6th evaluation campaign EVALITA 2018[6] we exploited this new gold standard, releasing the TWITTIRÒ corpus as part of the training set of the task on *Irony Detection in Italian Tweets (IronITA)*[7] (Cignarella et al. 2018b), in order to explore the activation of different types of irony (Sulis et al. 2016; Van Hee, Lefever, and Hoste 2016), with a special focus on sarcasm (Wang 2013).

Furthermore, we are planning to enrich our actual dataset with additional syntactic information such as Part-of-Speech tags and syntactic relations in *Universal Dependencies* (UD) format, in order to investigate the role that syntax plays in the activation of irony in Italian, especially in social media short messages.

References

Abbott, Rob, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media, LSM 2011*, pages 2–11, June 23, 2011, Portland, Oregon. Association for Computational Linguistics (ACL).

Artstein, Ron and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

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

Bamman, David and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, pages 574–577, May 26-29, 2015, Oxford, UK. AAAI.

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, December 5-6, 2016, Napoli, Italy. CEUR-WS.org.

Barbieri, Francesco, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the Fifth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, June 27, 2014, Baltimore, Maryland, USA. Association for Computational Linguistics (ACL).

Basile, Valerio, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the EVALITA 2014 Sentiment Polarity Classification Task. In *Proceedings of the Fourth evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, pages 50–57, December 11, 2014, Pisa, Italy. Pisa University Press.

---

5 The dataset is available at: https://github.com/AleT-Cig.

6 See: http://www.evalita.it/2018.

7 See also: http://di.unito.it/ironita18.

Benamara, Farah, Cyril Grouin, Jihen Karoui, Véronique Moriceau, and Isabelle Robba. 2017. Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du Défi Fouille de Textes DEFT2017. In *Actes de l'atelier DEFT2017 associé à la conférence TALN*, pages 1–12, 26 June 2017, Orléans, France. Association pour le Traitement Automatique des Langues (ATALA).

Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Buschmeier, Konstantin, 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 Fifth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, June 27, 2014, Baltimore, Maryland, USA. Association for Computational Linguistics (ACL).

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 First International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion,TSA '09*, pages 53–56, November 6, 2009, Hong Kong, China. ACM.

Cignarella, Alessandra Teresa, Cristina Bosco, and Viviana Patti. 2017. TWITTIRÒ: a Social Media Corpus with a Multi-layered Annotation for Irony. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, volume 2006 of *CEUR Workshop Proceedings*, December 11-13, 2017, Rome, Italy. CEUR-WS.org.

Cignarella, Alessandra Teresa, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018a. Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018)*, pages 4204–4211, May 7-12, 2018, Myazaky, Japan. European Language Resources Association (ELRA).

Cignarella, Alessandra Teresa, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018b. Overview of the Evalita 2018 Task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA'18)*, volume 2263 of *CEUR Workshop Proceedings*, December 10-13, 2018, Turin, Italy. CEUR-WS.org.

Clark, Herbert H. and Richard J. Gerrig. 1984. *On the Pretense Theory of Irony.* American Psychological Association.

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-2010*, pages 107–116, July 15-16, 2010, Uppsala, Sweden. Association for Computational Linguistics (ACL).

Filatova, Elena. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eighth Language Resources and Evaluation Conference (LREC 2012)*, pages 392–398, May 21-27, 2012, Istanbul, Turkey. European Language Resources Association (ELRA).

Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*, page 470–478, June 4-5, 2015, Denver, Colorado. Association for Computational Linguistics (ACL).

González-Ibáñez, Roberto, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the Forty-ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2011*, pages 581–586, June 19-24, 2011, Portland, Oregon. Association for Computational Linguistics (ACL).

Grice, Paul H. 1975. Logic and Conversation. *Syntax and Semantics 3: Speech Arts*, pages 41–58.

Grice, Paul H. 1978. Further Notes on Logic and Conversation. *Pragmatics*, 1:13–128.

Hernández Farías, Delia Irazú, José-Miguel Benedí, and Paolo Rosso. 2015. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In *Pattern Recognition and Image Analysis*, volume 9117 of *Lecture Notes in Computer Science*. Springer International Publishing, June 17-19, 2015, Santiago de Compostela, Spain, pages 337–344.

Hernández Farías, Delia Irazú, Viviana Patti, and Paolo Rosso. 2016. Irony Detection in Twitter: The Role of Affective Content. *ACM Transaction of Internet Technology*, 16(3). Issue in progress.

Joshi, Aditya, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the Fifty-third Annual Meeting of the Association for*

Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 757–762, July 27-31,2015, Beijing, China. Association for Computational Linguistics (ACL).

Karoui, Jihen, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015a. Towards a Contextual Pragmatic Model to detect Irony in Tweets. In *Proceedings of the Fifty-third Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, July 27-31, 2015, Beijing, China. Association for Computational Linguistics (ACL).

Karoui, Jihen, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich-Belguith. 2015b. Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In *Proceedings of the Fifty-third Annual Meeting of the Association for Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 644–650, July 27-31, 2015, Beijing, China. Association for Computational Linguistics (ACL).

Karoui, Jihen, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. In *Proceedings of the Fifteenth European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 262–272, April 3-7, 2017, Valencia, Spain. Association for Computational Linguistics (ACL).

Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the ICWSM: International AAAI Conference on Web and Social Media*, pages 538–541, July 17-21, 2011, Barcelona, Spain. AAAI Press.

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.

Lukin, Stephanie and Marilyn Walker. 2013. Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, June 13, 2013, Atlanta, Georgia, June. Association for Computational Linguistics (ACL).

Maynard, Diana and Adam Funk. 2011. Automatic Detection of Political Opinions in Tweets. In *Proceedings of the ESWC: Extended Semantic Web Conference (ESWC 2011)*, volume LNCS 7117, pages 88–99, May 29-June 2, 2011, Heraklion, Crete, Greece. Springer-Verlag.

Mihalcea, Rada and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 337–347, February 18-24, 2007, Mexico City, Mexico. Springer-Verlag.

Ptáček, Tomáš, Ivan Habernal, and Jun Hong. 2014. Sarcasm Detection on Czech and English Twitter. In *Proceedings of the Twenty-fifth International Conference on Computational Linguistics (COLING 2014)*, pages 213–223, August, 23-29, 2014, Dublin, Ireland. Dublin City University and Association for Computational Linguistics (ACL).

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, WSDM 2015*, pages 97–106, January 31 – February 6, 2015, Shanghai, China. ACM.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2010. Finding Humour in the Blogosphere: the Role of Wordnet Resources. In *Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, pages 56–61, January 31 – February 4, 2010, Mumbai, India.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74:1–12.

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, (EMNLP 2013)*, pages 704–714, October 18-21, 2013, Seattle, Washington, USA. Association for Computational Linguistics (ACL).

Shutova, Ekaterina. 2010. Models of Metaphor in NLP. In *Proceedings of the Forty-eighth Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 688–697, July 11–16,

2010, Uppsala, Sweden. Association for Computational Linguistics (ACL).

Sperber, Dan and Deirdre Wilson. 1981. Irony and the Use-Mention Distinction. *Philosophy*, 3:143–184.

Stranisci, Marco, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC 2016)*, pages 2892–2899, May 23-28, 2016, Portorož, Slovenia. European Language Resources Association (ELRA).

Stranisci, Marco, Cristina Bosco, Viviana Patti, and Delia Irazú Hernández Farías. 2015. Analyzing and Annotating for Sentiment Analysis the Socio-political Debate on #labuonascuola. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 274–279, December, 3-4, 2015, Povo, Trento, Italy. CEUR-WS.org.

Sulis, Emilio, D. Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative Messages and Affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143.

Tang, Yi-jie and Hsin-Hsi Chen. 2014. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of the Twenty-fifth International Conference on Computational Linguistics (COLING 2014)*, pages 1269–1278, August 23-29, 2014, Dublin, Ireland. Association for Computational Linguistics (ACL).

Van Hee, Cynthia, Els Lefever, and Veronique Hoste. 2016. Exploring the Realization of Irony in Twitter Data. In *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC 2016)*, pages 1795–1799, May 23-28, 2016, Portorož, Slovenia. European Language Resources Association (ELRA).

Van Hee, Cynthia, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)*, pages 39–50, June 5-6, 2018, New Orleans, LA, USA. Association for Computational Linguistics (ACL).

Wallace, Byron C., Do Kook Choe, and Eugene Charniak. 2015. Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In *Proceedings of the Fifty-third Annual Meeting of the Association for Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, July 27-31, 2015, Beijing, China. Association for Computational Linguistics (ACL).

Wang, Angela P. 2013. #Irony or #Sarcasm – A Quantitative and Qualitative Study Based on Twitter. In *Proceedings of the Twenty-seventh Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 349–356, November 22-24, 2013, Taipei, Taiwan. Department of English, National Chengchi University.

Whissell, Cynthia. 2009. Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Languages. *Psychological Reports*, 2(105):509–521.

Wilson, Deirdre and Dan Sperber. 2007. On verbal irony. *Irony in language and thought*, pages 35–56.