TWITTIRÒ:
a Social Media Corpus with a Multi-layered Annotation for Irony

Alessandra Teresa Cignarella, Cristina Bosco and Viviana Patti
Dipartimento di Informatica, Università degli studi di Torino
alessandra.cignarell@edu.unito.it
{bosco,patti}@di.unito.it

Abstract

English. In this paper we describe our work concerning the application of a multi-layered scheme for the fine-grained annotation of irony (Karoui et al., 2017) on a new Italian social media corpus. In applying the annotation on this corpus containing tweets, i.e. TWITTIRÒ, we outlined both strengths and weaknesses of the scheme when applied on Italian, thus giving further clarity on the future directions that can be followed in the multilingual and cross-language perspective.

Italiano. *In questo articolo descriviamo la creazione di un corpus di testi estratti da social media in italiano e l'applicazione ad esso di uno schema multilivello per l'annotazione a grana fine dell'ironia sviluppato in (Karoui et al., 2017). Nell'applicare l'annotazione a questo corpus composto da messaggi di Twitter, i.e.* TWITTIRÒ*, abbiamo discusso i punti di forza ed i limiti dello schema stesso, in modo da evidenziare le direzioni da seguire in futuro anche in prospettiva multilingue e cross linguistica.*

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 et al., 2010; Kouloumpis et al., 2011; Maynard and Funk, 2011; Reyes et al., 2012; Hernández Farías et al., 2016). Our goal, was to create an annotated Italian corpus through which we could address some issues concerning formalization and automatic detection of irony. This work collocates, therefore, in the context of a multilingual project for studying irony and for developing resources to be exploited in training NLP tools for sentiment analysis.

Providing that irony detection is a field that has been growing very fast in the last few years (Maynard and Greenwood, 2014; Ghosh et al., 2015; Sulis et al., 2016), and also taking into account that generation of irony (whether it is spoken or written) may also depends on the language and culture in which it is expressed, the main aim of this work is that of replying to the following research questions: *Is it possible to formally model irony? If so, how?*

Through the present paper indeed, we aim at contributing to the study of irony not only in Italian, but rather in a multilingual and cross-linguistic perspective. Our hope is that, on the one hand, studying the use of figurative language in Italian social media texts, will help us to better understand the developing of this figure of speech itself -irony- and its relations with humor. On the other hand, the study will lead us to the discovery of features and patterns that can be shared and confronted with similar projects in other languages.

2 Data collection

In this section we describe the methodology applied in the collection of tweets, and the internal structure of the dataset. Our work is part and extends a wider joint project with other research groups working on English and French (Karoui et al., 2017). In the French and English datasets, where the same annotation scheme for irony has been applied, tweets were retrieved by using Twitter APIs and filtered through specific *hashtags* exploited by users to self-mark their ironic intention (#irony, #sarcasm, #sarcastic). Providing that Italian users exploit a series of humorous hashtags, but no long-term single hashtag is established and shared among them, the same procedure could not be applied.

Some corpora from Twitter, where the presence

of irony is marked, have been made available for Italian in the last few years, and we extracted from them tweets to be included in TWITTIRÒ according to the distribution presented in Table 1[1].

Corpus	Number of tweets
TW-SPINO	400
SENTIPOLC	600
TW-BS	600
TWITTIRÒ	**1,600**

Table 1: Tweet distribution in TWITTIRÒ

As it is shown in Table 1 the tweets were collected from three different pre-existent datasets.

• **TW-SPINO** is a portion of SENTITUT (Bosco et al., 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. 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])*

• **SENTIPOLC** (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. Mario Monti? non era il nome di un antipasto? #FullMonti #laresadeiconti #elezioni #308.
   → *(Mario Monti? Wasn't it the name of a starter? #FullMonti #laresadeiconti #elezioni #308.)*

• **TW-BS** (Stranisci et al., 2015; Stranisci et al., 2016) contains tweets on the debate of the reform of Italian School "Buona Scuola".

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

3 A multi-layered annotation scheme

The main goal of the scheme proposed in (Karoui et al., 2017) is to provide a fine-grained representation of irony and to achieve this goal it includes four different levels of annotation 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.

LEVEL 2: CONTRADICTION 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, see example below, 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., 2015), we annotate the contradiction that relies exclusively on the lexical clues internal to the utterance as *explicit*, while 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* (free) and *processate* (process).

4. [**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. La [buona scuola e le **sillabe**]$_{P1}$ - http:t.conS42fRjAKp
   → *(The [buona scuola and the **syllables**]$_{P1}$ - http:t.conS42fRjAKp)*[2] 2

LEVEL 3: CATEGORIES. Both forms of contradictions can be expressed through different rhetorical devices, patterns or features that are grouped under different labels.

Analogy: In this category are summoned also other figures of speech that comprehend mechanisms of comparison, such as *simile* and *metaphor*.

[1] A portion of these tweets (400 messages) has already been exploited and analyzed in (Karoui et al., 2017).

[2] The official document that presented the school reform had hyphenation mistakes.

5. Il governo #Monti mi ricorda la corazzata kotiokmin.
   → *(Monti's government reminds me of the Battleship Kotiokmin)*

Hyperbole/exaggeration: It is a figure of speech which consists in expressing an idea or a feeling with an exaggerated way.

6. #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 one year at least 170,000 teachers will not be employed, I will bring them **all** to @Palazzo_Chigi #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.

7. 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 comic stips regarding Mohammad. It took them **a while** to understand them.)*

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.

8. Mario Monti? **non era il nome di un antipasto?** #FullMonti #laresadeiconti #elezioni #308.
   → *(Mario Monti? **Wasn't it the name of an appetizer?** #FullMonti #laresadeiconti #elezioni #308.)*

Context shift (explicit only): It occurs by the sudden change of the topic/frame in the tweet.

9. @matteorenzi Più che la **#labuonascuola** direi #carascuola visto che ci vogliono più di 800 euro a pischello....quasi quanto **5 kg di gelato**
   → *(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**.)*

Register changing: (sub-category of the former) in which the "context shift" is due to a sudden change of linguistic style, exploitation of vulgarities or, on the contrary, a rather pompous style. In Italian tweets, users often recur to the exploitation of dialectal expression:

10. Mario, Monti sulla **#cadrega**.
    → *(Mario, Monti on the **#chair.**)*

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).

11. Totoministri per il governo Monti: **Gelmini ai lavori pubblici, farà il tunnel dei neutrini!**
    → *(Footbal pools of ministers for the Monti's government: **Gelmini at public works' ministry, she will build the tunnel of neutrinos!**)*[3]

Oxymoron/paradox (explicit only): This category is equivalent to the category FALSE ASSERTION except that the contradiction, this time, is explicit.

12. Individuata una mafia tipicamente romana. **Prima di mezzogiorno non prendeva appuntamenti.**
    → *(Identified a typical Rome's mafia. **It did not fixed appointments before midday.**)*[4]

Other: This last category represents ironic tweets, which can not be classified under one of the other seven previous categories. It can occur in case of humor or situational irony.

13. Sicilia, arriva barcone di migranti e a bordo c'è anche un gatto. Vengono a rubarci i nostri like. [@LughinoViscorto]
    → *(Sicily, a big boat full of refugees arrives. There's also a kitty on board. They come here and steal our likes.)*

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.

4 Annotation and Disagreement

Given the complexity of irony attested in literature, it is not surprising that the task of annotating irony often leads to disagreement between annotators, which are connected to their individual experience, sense of humor and situational context (Grice, 1975; Grice, 1978; Sperber and Wilson, 1981; Wilson and Sperber, 2007; Reyes et al., 2010; Fink et al., 2011; Reyes et al., 2012).

In our work, the annotation process involved three people previously trained in similar tasks. Since we are aiming at testing the value of the

[3] Minister Gelmini was never in charge of public work administration. It is a reference to an erroneous statement about neutrinos that the Minister had previously uttered.

[4] It is common knowledge that people from Rome are often late, thus the paradox of creating a criminal organization that is also often late.

annotation scheme, the 1,200 new tweets were tagged by two independent annotators (A1 and A2) and by a third (A3) only where a disagreement is detected between A1 and A2.
According to (Karoui et al., 2017), the annotators were asked to apply the second and third levels of the scheme, thus classifying each tweet as featured by implicit or explicit contradiction and selecting for it a category tag between the eight proposed.

4.1 Disagreement Analysis

The inter-annotator agreement (IAA) between A1 and A2 for the labeling of implicit vs. explicit, calculated with Cohen's coefficient, is $\kappa = 0.41$ (moderate agreement), and the distributions of these labels for each annotator are reported in Table 2. Our data analysis, for the moment, seems to corroborate the results of Karoui et al. (2017) where the annotation for the pair EXPLICIT vs. IMPLICIT, obtained a kappa of *0.65* (substantial agreement).

It is interesting to note that while in French implicit activation is the majority (76.42%), in Italian the majority is represented by the explicit type. 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. Further studies are surely needed about the activation type of irony for Italian.

		A2		
		implicit	explicit	TOTAL
A1	implicit	**104**	136	240
	explicit	63	**897**	960
	TOTAL	167	1033	**1200**

Table 2: Inter-annotator agreement on type tags

The IAA regarding category tags is slightly higher, $\kappa = 0.46$ (moderate agreement), as we will examine in detail later. The comparison with the French dataset (Karoui et al., 2017) shows a slightly higher inter-annotator agreement: $\kappa = 0.56$ (still moderate). For the second time a clearer identification of pragmatic devices is encountered in French, overcoming the results obtained between Italian annotators.
It is also interesting to mention that, Karoui et al. (2017) operated some calculations when similar devices were grouped together and the scores showed an increment to $\kappa = 0.60$.

Since our work is mainly focused on category tags, their exploitation and distribution, we will discuss in particular on the tweets where A1 and A2 were in disagreement and the need A3's annotation was required (579 tweets). As support, Table 3 shows the distribution of category tags exploited by A1 and A2.

The analysis of the disagreement detected in this new experimental dataset supports the following ideas. Firstly, observing the tag distribution between A1 and A2, the tag OXYMORON/PARADOX is the more frequently exploited, followed by FALSE ASSERTION (see charts in Fig. 1). Concerning the latter, it is also observed a stronger bias from A1 towards that category tag (15.9%) compared to A2 choices (8.4%).

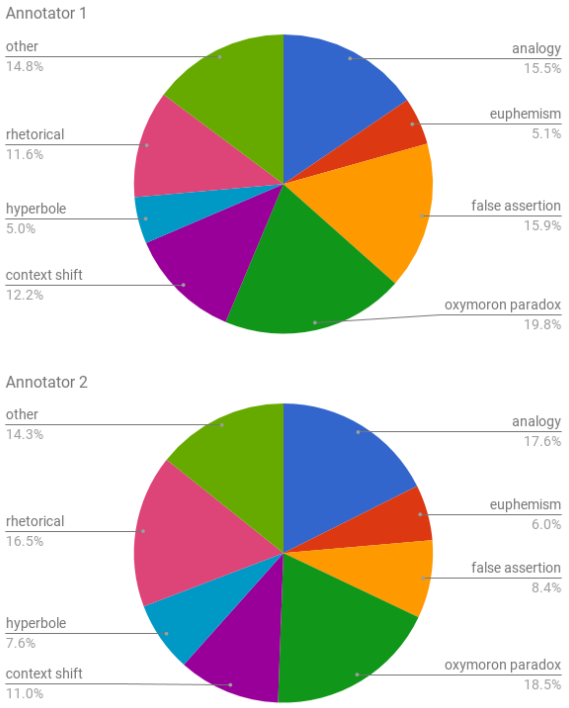

Figure 1: Category tags exploited by the two annotators

The comparison with the annotation results obtained on the French dataset furthermore triggers the need of a deeper research on the application of the scheme in a cross-linguistic perspective.

5 Discussion

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

		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	1200

Table 3: Inter-annotator agreement on category tags

Table 3). In the annotation guidelines it is indeed stated that the labels represent the same category but the former as realized in the context of an explicit contradiction, and the latter when an implicit contradiction happens. For example, in the following tweet A1 tagged as explicit OXYMORON/PARADOX, while A2 as implicit FALSE ASSERTION.

14. 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...)

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.

15. Ma secondo voi super #Mario #Monti riuscirà a tassare anche la felicità?
→ (What do you think, will super #Mario #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 semantical 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.

We have also noticed the exploitation of a common pattern, which we believe should constitute a new category on its own. We named it **false logical conclusion**, most of the time is an EXPLICIT CONTRADICTION, and it expresses which kind of relationship exists between a $P1$ and $P2$. In 45 out of 82 cases, when a false logical conclusion was signaled by at least one annotator (54.88%), the category was tagged as OTHER. We can interpret this as a statistically relevant signal of unsatisfaction of annotators towards the available seven applicable category-tags. Finally, we noticed a high presence of negative words in the whole corpus.

6 Conclusions and future work

The paper describes our work concerning the application of a fine-grained annotation scheme for pragmatic phenomena. In particular, it has been used to annotate the rhetorical device of irony in texts from Twitter. It confirms how this task is challenging, it contributed to shed some light on linguistic phenomena and to significantly extend the resource in (Karoui et al., 2017) with new Italian annotated data to be exploited in future experiments on irony detection in a multi-lingual perspective[5]. The disagreement in the annotation of irony in the three sub-corpora TW-SPINO, SENTIPOLC and TW-BS, which are featured by different characteristics, is a further issue to be addressed. In future work, we plan therefore to investigate the differences in the disagreemente detected across the three portions of TWITTIRÒ providing in-depth analysis of currently available and new linguistic data.

Acknowledgements

The work of Cristina Bosco, Alessandra Teresa Cignarella and Viviana Patti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, project S1618_L2_BOSC_01).

References

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the EVALITA 2014 Sentiment Polarity Classification Task. In *Proceedings of the 4th evaluation cam-*

[5]The dataset is available at: `https://github.com/IronyAndTweets/Scheme`

paign of Natural Language Processing and Speech tools for Italian (EVALITA'14).

Cristina Bosco, 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.

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

Clayton R. Fink, Danielle S. Chou, Jonathon J. Kopecky, and Ashley J. Llorens. 2011. Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. *Johns Hopkins APL Technical Digest*, 30(1):22–30.

Aniruddha Ghosh, 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 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

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

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

Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Trans. Internet Techn.*, 16(3):19:1–19:24.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, and Cristina Bosco. 2017. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Efthymios Kouloumpis, 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*.

Diana Maynard and Adam Funk. 2011. Automatic Detection of Political Opinions in Tweets. In *Proceedings of the ESWC: Extended Semantic Web Conference*.

Diana Maynard and Mark A. Greenwood. 2014. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2010. Finding Humour in the Blogosphere: the Role of Wordnet Resources. In *Proceedings of the 5th Global WordNet Conference*.

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

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

Marco Stranisci, 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 CLiC-it: Italian Conference on Computational Linguistics*.

Marco Stranisci, 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 International Conference on Language Resources (LREC 2016)*.

Emilio Sulis, Delia 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. New Avenues in Knowledge Bases for Natural Language Processing.

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