Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ

Alessandra Teresa Cignarella[1,2], **Cristina Bosco**[1], **Viviana Patti**[1], **Mirko Lai**[1,2]
[1]Dipartimento di Informatica, Università degli Studi di Torino, Italy
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
{cigna,bosco,patti}@di.unito.it
mirko.lai@unito.it

Abstract

In this paper we describe the main issues emerged within the application of a multi-layered scheme for the fine-grained annotation of irony (Karoui et al., 2017) on an Italian Twitter corpus, i.e. TWITTIRÒ, which is composed of about 1,500 tweets with various provenance. A discussion is proposed about the limits and advantages of the application of the scheme to Italian messages, supported by an analysis of the outcome of the annotation carried on by native Italian speakers in the development of the corpus. We present a quantitative and qualitative study both of the distribution of the labels for the different layers involved in the scheme which can shed some light on the process of human annotation for a validation of the annotation scheme on Italian irony-laden social media contents collected in the last years. This results in a novel gold standard for irony detection in Italian, enriched with fine-grained annotations, and in a language resource available to the community and exploitable in the cross- and multi-lingual perspective which characterizes the work that inspired this research.

Keywords: irony, figurative language processing, corpora, social media, Italian

1. Introduction

The automatic recognition of irony is, still nowadays, a challenging task to be performed both by human annotators and automatic NLP systems (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; Sulis et al., 2016). 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 interesting 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 for the first time (*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 multiclass classification task on different types of irony, where the systems must predict one out of four labels describing: 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, 2017; Van Hee et al., in press 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 annotation of finer-grained features and resources in order to hopefully improve the performance of automatic systems in this especially challenging task.

This work aims at the creation of a novel resource for irony detection in the Italian language called TWITTIRÒ. We considered as starting point for this work the scheme provided in (Karoui et al., 2017), which was initially applied to a set of 400 Italian tweets. In particular, we would like to highlight how the complexity of a pragmatic device such as irony, also attested in literature (Grice, 1975; Grice, 1978; Sperber and Wilson, 1981; Wilson and Sperber, 2007; Reyes et al., 2010; Fink et al., 2011; Reyes et al., 2012), makes the annotation task particularly challenging, as will be discussed in a deep analysis of the disagreement between the native Italian speakers involved in the development of the resource. Human annotators, even skilled or domain experts, are indeed always connected to their individual experience, 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. Our investigation concerns the linguistic devices known in pragmatics as signals of irony and their relevance for modeling irony in a computational perspective.

As we will explore in detail in the following section, the TWITTIRÒ corpus consists of three sub-corpora characterized by linguistic differences and peculiarities. With the description and analysis of the current release 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 (Cignarella et al., 2017; Karoui et al., 2017). Since the resulting annotated corpus will be exploited as reference dataset within the context of the next Evalita evaluation campaign[2], it will be made available to the community and exploitable in the cross- and multi-lingual perspec-

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

[2]http://www.evalita.it/2018

tive depicted in (Karoui et al., 2017) from the end of 2018[3]. The paper is organized as follows: in the next sections we describe the dataset on which we worked, focusing on the collection, the annotation process and the annotation scheme. Section 4. is centered on the analysis of the disagreement detected during annotation and presents some hints about the quantitative results. Finally we show, in Section 5., a selection of cases especially difficult to deal with our annotation scheme.

2. Building the Corpus

This project aims at developing an Italian Twitter corpus, to be used as language resource in the training of NLP tools and to become a benchmark in evaluation campaigns for this language, for what concerns irony detection. Since a preliminary resource composed of 400 Italian tweets was available, where a very interesting scheme for describing irony at a fine-grain level (Karoui et al., 2017) has been applied, we considered it as a starting point for our work. In order to extend the corpus, we collected new data (i.e. 1,200 tweets), whose balancing is coherent with that applied in this small existing corpus, and we applied the same scheme in order to build TWITTIRÒ, which thus now includes 1,600 tweets as shown in Table 1[4].

In this section we describe the methodology applied in the collection of these new tweets, and the internal structure of the novel dataset TWITTIRÒ. Some Italian corpora containing Twitter data, where the presence of irony is marked, have been made available to the community in the last few years, thus we extracted from them the new 1,200 tweets to be included in TWITTIRÒ. In particular, the tweets were collected from the following three different pre-existent datasets: TW-SPINO, TW-SENTIPOLC14 and TW-BS.

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 featured by a high register and style, while the topics are variegate with a preference for jokes concerning politics and news.

TW-SENTIPOLC14 (Basile et al., 2014) contains tweets generated by common Twitter users and therefore it is less homogeneous than TW-SPINO. The use of grammar is sometimes very poor, colloquial expressions and vulgarities typical of Computer-Mediated Communication (CMC) appear, such as the frequent use of creative hashtags, mentions, repetitions of laughters. We selected here the political tweets with reference to the government of Mario Monti between 2011 and 2012.

TW-BS (Stranisci et al., 2016) contains tweets on the debate of the reform of Italian School "Buona Scuola". Similarly to TW-SENTIPOLC14, also here devices typically exploited in CMC are shown. For instance, being the reform of the education system a highly criticized one, the use of sentences written in ALL CAPS (to decode shouting) is wide. Part of this corpus has been included in the test set within the Sentipolc 2016 evaluation campaign (Barbieri et al., 2016).

Table 1 shows the composition of TWITTIRÒ and the distribution of tweets over the three sub-corpora.

TW-SPINO	TW-SENTIPOLC14	TW-BS
400	**600**	**600**

Table 1: Tweet distribution in the TWITTIRÒ corpus

The original Italian resource described in (Karoui et al., 2017) was part of a project for studying irony in a multilingual perspective and including also similar English and French datasets annotated with the same scheme. As for what concerns the French and English datasets, 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.

In the following section we focus on the novel 1,200 tweets only, provided that the collection, annotation and disagreement analysis of the previously developed 400 tweets corpus has been discussed in (Karoui et al., 2017) within the context of multilingual experiments. Also for what concerns the details and guidelines of the annotation schema applied in TWITTIRÒ, we refer to the same paper and to (Cignarella et al., 2017), where we discussed a preliminary stage of development of the novel resource.

3. Annotation Process

The annotation process of the 1,200 tweets corpus, coherently of what was done in (Karoui et al., 2017), involved three people previously trained in similar tasks: all tweets were tagged by two independent annotators (A1 and A2) and by a third (A3) only for the tweets where a disagreement was detected.

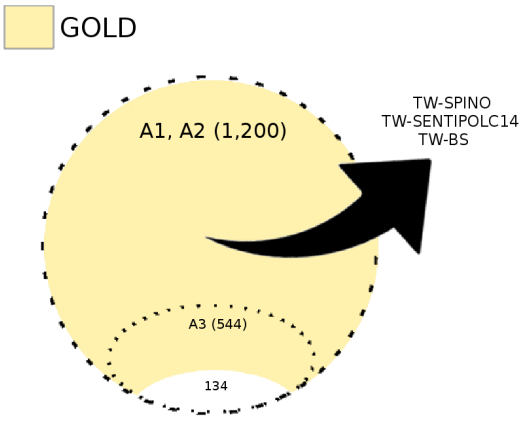

Figure 1: Portions of different agreement level and annotator's contribution on TWITTIRÒ data

[3] https://github.com/IronyAndTweets/Scheme

[4] Considering that the complexity of tasks related to the detection of pragmatic phenomena does not only depend on the inner structure of irony, but also on unbalanced data distribution, we built the novel resource by maintaining the same proportions considered in the first collection of 400 Italian ironic tweets described in (Karoui et al., 2017).

Figure 1 shows the portions of TWITTIRÒ that were annotated, and how many annotators give their contribute in the annotation for achieving the agreement (or not achieve it). As it can be seen, both annotators A1 and A2 annotated all the corpus, but, after their work, the disagreement analysis shows that they gave different annotations on 544 tweets. Therefore, annotator A3 provided a further annotation for these tweets, allowing the achievement of the agreement for 410 additional tweets. The last 134 tweets remain in disagreement and they were, then, discarded.

3.1. Annotation Scheme and Examples

The multi-layer scheme described in (Karoui et al., 2017) includes 4 different levels. Provided that Level 1 concerns the classification of tweets into **ironic** or **not ironic** which has been previously applied on the data we collected for our corpus (that are all ironic), we don't discuss its application in this paper. The three human annotators A1, A2, and A3, indeed, worked on the application of tags concerning Level 2 and Level 3.

LEVEL 2: IRONY ACTIVATION. As far as Level 2 is concerned, for each ironic tweet the annotators decided whether the type of contradiction that activates irony was EXPLICIT or IMPLICIT, that is determined by a contradiction between two items directly cited within the message, or by a contradiction between a directly cited item and some other things in the external context.

LEVEL 3: IRONY CATEGORIES. Both explicit and implicit activation types can be expressed in different ways. At this level, annotators were asked to classify the linguistic device triggering irony by applying one category tag from the following list: ANALOGY, EUPHEMISM, HYPERBOLE/EXAGGERATION, CONTEXT SHIFT (explicit only), OXYMORON/PARADOX (explicit only), FALSE ASSERTION (implicit only), RHETORICAL QUESTION or OTHER (humor or situational irony). In Table 2 we provide a brief description for each category, while here below we also discuss some examples, in order to better clarify the application of the scheme to our Italian social media texts.

1. ANALOGY
   Leo Messi: "Firmo quello che mi dice papà". Pure la Boschi.
   → *Leo Messi: "I sign what daddy tells me". Also Minister Boschi.*

The analogy here links two figures: the footballer Lionel Messi and the Italian Minister Boschi. The figure of speech is referred to the fact that the world-known footballer once affirmed that, if his father tells him to sign something, he would do it without hesitation. In particular, the athlete has signed (apparently without knowing) some contracts regarding the rights on his public image, and money has been transferred in fiscal paradises such as Uruguay and Belize. The second element of the tweet is Maria Elena Boschi, whose dad's shady affairs and implications with the bankruptcy of *Banca Etruria* are still nowadays elements of tension and discussion.

2. HYPERBOLE
   #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 from now there will not be 170,000 teachers more, I will bring them all to Palazzo_Chigi #labuonascuola.*

While reforming Italian School, Prime Minister Renzi promised the opening of 170,000 new job places for teachers. The exaggeration in the tweet is referred to the fact that, the user states that if this should not happen, he will drag all those unoccupied workers to Palazzo Chigi in Rome, where Italian Prime Ministers normally live.

3. EUPHEMISM
   Nel 2006 Charlie Hebdo aveva pubblicato delle vignette satiriche su Maometto. Ci hanno messo un po' a capirle. [nicodio]
   → *In 2006 Charlie Hebdo had published some satirical images on Mohamad. It took them a while to understand them.*

CATEGORY	DESCRIPTION
ANALOGY	In this category are summoned *analogy*, and also other figures of speech that comprehend mechanisms of comparison, such as *simile* and *metaphor*.
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.
CONTEXT SHIFT	It occurs by the sudden change of the topic/frame in the tweet.
FALSE ASSERTION	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).
HYPERBOLE	It is a figure of speech which consists in expressing an idea or a feeling with an exaggerated way.
OXYMORON / PARADOX	This category is equivalent to the category FALSE ASSERTION except that the contradiction is explicit.
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.
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.

Table 2: Description of categories

The tweet is a reference to the terrorist attack that took place at the offices of the French satirical weekly newspaper *Charlie Hebdo* on January 7, 2015. The second part of the tweet states: "It took them a while to understand them", it is an euphemism implying that the attack of 2015 is a consequence of a satirical comic strip published in 2006 about Mohamed, from the French newspaper.

4. CONTEXT SHIFT
   L'auto di Salvini assalita al campo rom. Rovinato il safari. [@paniruro]
   → *The car of Salvini assaulted at the Roma camp. The Safari is ruined. [@paniruro]*

The tweet points at news from November 8, 2014 in which Matteo Salvini, Secretary of the Italian right party *Lega Nord*, visited a Roma camp in Bologna, and his car has been assaulted, punched and kicked by youths of left-wing associations. The implicit connection of the user is that the young rebels behaved as aggressively as fierce animals normally do during a safari trip, the context is therefore, shifted.

5. FALSE ASSERTION
   Brunetta sostiene di tornare a fare l'economista, Mario Monti terrorizzato progetta di mollare tutto ed aprire un negozio di pescheria.
   → *Brunetta affirms that he will go back to be an economist, Mario Monti plans of leaving everything and opening a fish monger's.*

The false assertion is referred to many affirmations of the politician Renato Brunetta within the past years "*I am more rigorous then Tremonti [...] I know well all this topics, because I AM an economist, Tremonti isn't.*". Several times Brunetta publicly discredited other economists and colleagues such as Tremonti or Monti, who are instead his peers.

6. OXYMORON / PARADOX
   "Potrei non opporre veti a un presidente del Pd", ha detto Berlusconi iscrivendosi al Pd.
   → *"I could not deny rights to a President of the PD", said Berlusconi while subscribing to the party.*

A tweet in which Silvio Berlusconi (center-right politician) declares that he will subscribe to the PD (center-left party) is clearly a paradox, but the user is subtly making a reference to the blurry ideology of the leftist party, which, since a couple of years seems more a right-centered party. Hence, not so different from the berlusconian party: FORZA ITALIA.

7. RHETORICAL QUESTION
   Mario Monti? non era il nome di un antipasto? #FullMonti #laresadeiconti #elezioni #308.
   → *Mario Monti? Wasn't it the the name of an entree? #FullMonti #finalcountdown #elections #308*

The tweet contains a rhetorical question based on a pun that associates the name of the premier Mario Monti and a common pizza flavor named *Mari e Monti* (seas and mountains), in which normally you can find seafood combined with mushrooms. Other typical elements of social media texts are present, such as the humorous hashtag *#FullMonty*[5], and the hashtag *#308* which is referred to the number of deputies who voted "yes" to the new harsh financial law proposed by Monti.

8. OTHER
   Sicilia, arriva barcone di migranti e a bordo c'è anche un gatto. Vengono a rubarci i nostri like.
   → *Sicily, a big boat full of migrants arrives and there's also a kitty on board. They come here and steal our likes.*

The tweet regards a news fact[6] and it has been labeled as OTHER because of the presence of an overlapping of more than one category. Firstly, the commonplace on immigrants "They come here and steal our jobs" has been mutated in the era of social networks, in "They come here and steal our likes". The joke is based on the implicit knowledge of the Internet-world that videos and pictures containing kitties receive tons of likes from users. All this adds up to a paradox, because thinking that a tragic situation as the arrival of migrants on boats is compared at the pursuit of likes on the net, is just dramatic.

Finally, let us notice that the annotation scheme provided in (Karoui et al., 2017) also includes **LEVEL 4**, which is referred to an even finer-grained annotation of irony and takes into account the presence of several **clues** such as punctuation negation words, emoticons, punctuation marks, interjections, named entities (and mentions). Nevertheless, since the extraction of the information about this level can be done, to a great extent, by automatic tools, this specific task is not addressed by manual annotation and it is not discussed in this paper.

4. Annotation Analysis

In this section we will analyze the distribution of the annotated labels within the corpus and the inter-annotator agreement.

4.1. Label Distribution

The annotation process described in the last section allowed the achievement of an agreement for 1,066 tweets out of the 1,200 tweets collected and annotated. Together with the 400 tweets analyzed in (Karoui et al., 2015) and described at the beginning of section 2. they can be considered as a novel gold standard for Italian, that is TWITTIRÒ, consisting of almost 1,500 tweets. In the rest of this section we describe the distribution of the annotated labels on the TWITTIRÒ corpus.

Figure 2 shows the distribution among the three subcorpora from where the tweets were extracted (TW-SPINO,

[5] http://it.wikipedia.org/wiki/Full_Monty_-_Squattrinati_organizzati

[6] http://www.ansa.it/web/notizie/rubriche/associata/2014/01/04/Immigrazione-soccorso-anche-gatto-migrante-barcone_9851581.html

TW-SENTIPOLC-14, and TW-BS), labeled with either EXPLICIT or IMPLICIT tag, concerning the type of activation of irony.

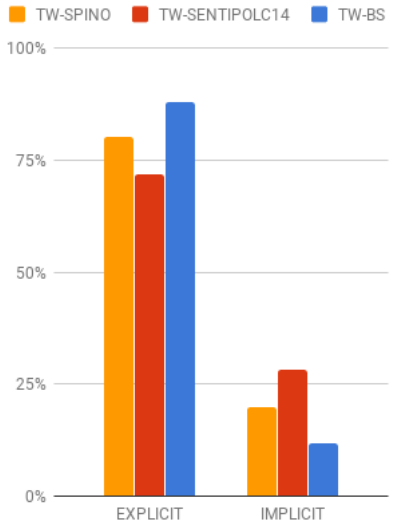

Figure 2: Distribution of types

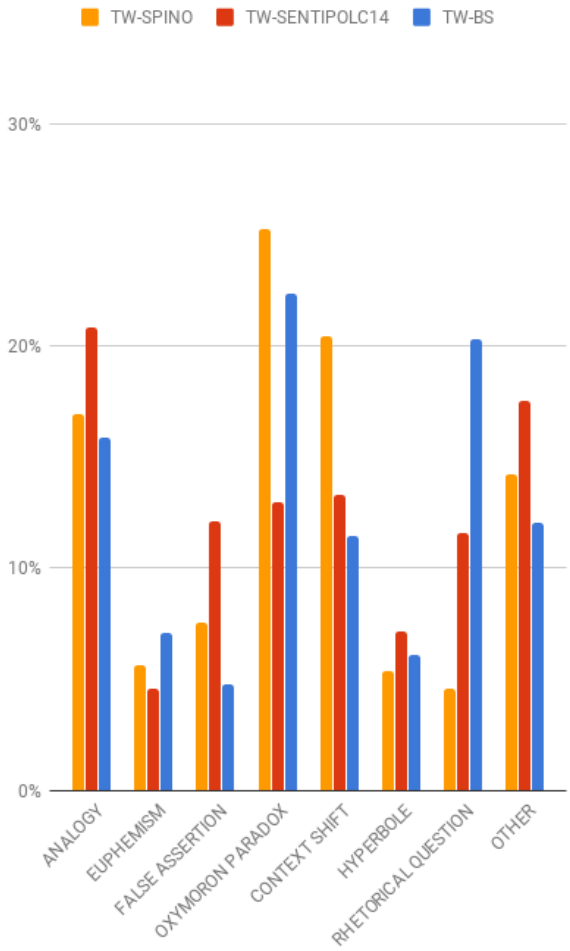

Figure 3: Distribution of categories

As it can be seen in Figure 2, in the majority of tweets, irony is triggered by an explicit contradiction in all the three datasets (80% in TW-SPINO, 72% in TW-SENTIPOLC14, and 88% in TW-BS). This confirms the findings in (Karoui et al., 2017) on the first section of the corpus, which highlight that Italian 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 3 shows the distribution of labels for Level 3 of the annotation scheme, describing how devices that trigger irony are distributed in the three different sub-corpora. The sub-corpus TW-SPINO is characterized by a strong use of the devices of OXYMORON/PARADOX and CONTEXT SHIFT. This suggests that the contributors of the satirical blog *Spinoza.it* often recur to this explicit kind of devices in order to create a sense of surprise, as in the following example:

- **TW-SPINO** - Marino si è dimesso. Ora la metro sembra nuova. [pirata_21]
  Marino resigned. Now the underground looks new. [pirata_21]
  LEVEL 2: EXPLICIT
  LEVEL 3: OXYMORON/PARADOX

Another observation that can be discussed about Figure 3, is the fact that TW-BS, a corpus composed by tweets about the reformation of the Italian School: *"La Buona Scuola"*, contains a high number of RHETORICAL QUESTION tags, which we can link to the dissatisfaction of the people, or at least, their perplexity on the matter.

- **TW-BS** - Ma i punti de #labuonascuola sono riferiti anche a quella pubblica? #perdire
  Are the bullet points of #labuonascuola also referred to public school? #justsaying
  LEVEL 2: EXPLICIT
  LEVEL 3: RHETORICAL QUESTION

4.2. Inter-annotator Agreement

As cited in the previous section, we compared the annotation of A1 and A2 on all the 1,200 tweets, calculating that the two annotators achieved the agreement for 656 tweets, but they disagree on the other 544. To solve the disagreement we applied a further independent annotation of a third human expert (A3), and we achieved the agreement on further 410 tweets.
We compared the annotation of A1 and A2 on all the 1,200 tweets. Their agreement, calculated through Cohen's kappa coefficient shows some interesting results. In fact, the inter-annotator agreement (IAA) between A1 and A2 concerning the choice between IMPLICIT and EXPLICIT (Level 2), calculated with Cohen's coefficient, is $\kappa = 0.41$, that is a low agreement (Landis and Koch, 1977). On the other hand, the IAA regarding category tags (Level 3) is $\kappa = 0.46$, i.e. a moderate agreement.
Figure 4 shows how the labels of category tags chosen by A1 is distributed for each category chosen by A2. This type of illustration is helpful to understand how the choices of one annotator agree (or disagree) with the choices of the

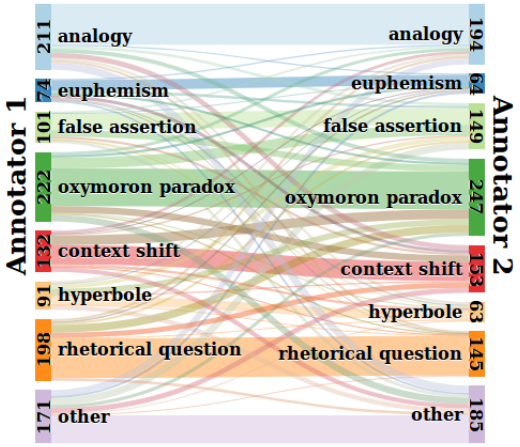

Figure 4: Detail over the agreement between A1 and A2

other, and at the same time what happens in case of disagreement on the tagging of a certain tweet. On the left the choices of A1 are displayed, whereas on the right we see the choices of A2. All the lines, or "fluxes", connecting left and right, describe the sparsity of the annotation work, and are coherent with a certain rate of disagreement. In fact, the more the annotators agree, the more the lines are straight and the less we would see divergent fluxes.

We noticed that 55 times (about 10% of disagreement cases), OXYMORON/PARADOX was chosen by an annotator and FALSE ASSERTION was chosen by the other. This statistics corroborate the intuitions of previous work (Cignarella et al., 2017), stating that there might be a problem on the comprehension of the true value of these two specific category tags. 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 (Category Type) is triggered from the previous decision of Level 2 (Contradiction Type). 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.

Observing the tag distribution between A1 and A2 in Figure 4, the tag OXYMORON/PARADOX is the more frequently exploited, followed by ANALOGY.

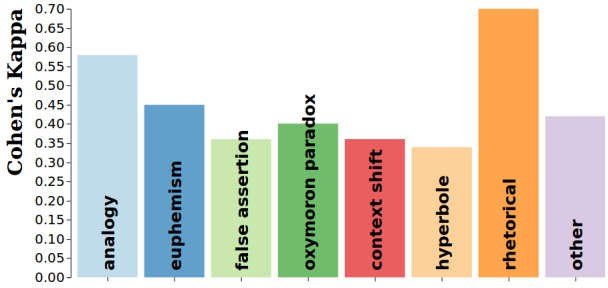

Figure 5: IAA between A1 and A2 on each category tag

To further validate our intuitions, we calculated the agreement of A1 and A2 on each category tag, and as it can be seen in Figure 5 the category tags OXYMORON/PARADOX (κ = 0.40) and FALSE ASSERTION (κ = 0.36) are among the three worst categories in agreement, preceded only by HYPERBOLE (κ = 0.34). This means that even though annotators exploit the OXYMORON/PARADOX tag several times, they rarely agree on its correct application. The categories with the highest inter-annotator agreement are instead RHETORICAL QUESTION (κ = 0.70) and ANALOGY (κ = 0.56). In general, as it is summarized in Table 3, annotators A1 and A2 reach a moderate agreement, κ = 0.46, on category tags. Moreover, it is interesting to notice that the value of the average kappa, on all 1,200 is increased by the pretty good value of the annotation on the sub-corpus of SPINOZA (κ = 0.57) and lowered by the poor results on TW-SENTIPOLC14 (κ = 0.34).

n# of tweets	sub-corpus	IAA
1,200	ALL	κ = 0.46
300	TW-SPINO	κ = 0.57
312	SENTIPOLC	κ = 0.34
588	TW-BS	κ = 0.47

Table 3: Cohen's kappa (A1 and A2) on each sub-corpus

Our intuition is that the use of correct grammar, good writing style and punctuation, revised by the authors of the satirical blog, 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. As we already mentioned, also after the application of a third human independent annotation, we didn't reach an agreement for classifying 134 tweets according to our scheme. Thinking that a deeper observation of tweets in disagreement can give interesting information about the different nuances through which irony is produced and about the complexity of the task, we further discuss such hard cases in the next section.

5. Discussion on Difficult Cases

Of the last remaining 134 tweets, which still are in a condition of disagreement (referred as *HardCases* henceforth), 20% pertains to TW-SPINO, 23% pertains to TW-SENTIPOLC14 and 57% to TW-BS.

In the following, we provide linguistic examples and description of one tweet of *HardCases* extracted from each sub-corpus.

- **TW-SPINO** - Alla stazione di Budapest nasce una bimba chiamata Speranza. In Italia l'avrebbero chiamata "Ci scusiamo per il disagio". [guli1979]
  At the Budapest train station a child named Hope was born. If had been born in Italy she'd be called "We're sorry for the inconvenience". [guli1979]
  A1: ANALOGY,
  A2: OXYMORON,
  A3: CONTEXT SHIFT.

The tweet makes an implicit reference to the primary train operator in Italy, *Trenitalia*, which is well known to offer

unsatisfying service and to often accumulate hours of delay on its convoys. The sentence "We're sorry for the inconvenience" is what travelers hear from a registered voice aired through the loud speakers in the train stations when such delay happen.

- **TW-SENTIPOLC14** - Tutti i ministri del governo Monti sono più vecchi di me. Sono soddisfazioni.
  All the ministers in the Monti government are older than me. This is satisfying!
  A1: CONTEXT SHIFT,
  A2: OTHER,
  A3: EUPHEMISM.

The ironic tweet refers to the fact that Mario Monti, the Italian Prime Minister between 2011 and 2013 in the wake of the Italian debt crisis, is an old man and, as he was appointed, he composed his commission with politicians even older than him.

- **TW-BS** - @Corriereit "non si consegna mai il compito senza rileggere": regola n.1 della "buona scuola". Glielo diciamo?
  @Corriereit "You never hand in the paper without correcting first": first rule of "la buonascuola". Should we tell them?
  A1: ANALOGY,
  A2: CONTEXT SHIFT,
  A3: RHETORICAL QUESTION.

The tweet makes an implicit reference to the event in which the first draft of the document presenting the school reform *"La Buona Scuola"* contained six orthographic errors and, has therefore been mocked ever since as a non-serious school reform. Furthermore the sentence "You never hand in the paper without correcting first" is an echoic mention based on common knowledge among Italian students, because it is what each teacher would repeat to pupils before they hand in a test.

Below each tweet we reported the annotation tags, assigned by each annotator. It can be observed that, when irony is activated, there can be the co-occurrence of one or more categories. None of the choices made from the annotators could be depicted a priori as *right* or *wrong*. Therefore, a good improvement in the application of the scheme would be that of accepting the labeling of two or more categories from each annotator on a single tweet.

6. Conclusions

The present research aimed at seeking more answers on the applicability of the multi-layered annotation scheme for irony (Karoui et al., 2017) on Italian texts extracted from Twitter. In doing so, we expanded the TWITTIRÒ corpus in order to create a gold standard which can be exploited in the cross- and multi-lingual perspective embraced in the research that inspired it.

In the present work, we conducted both a quantitative and qualitative study on the annotated data, with a focus on the annotation phase, its outcomes and an analysis of annotation disagreement. We deeply discussed the inter-annotator agreement (IAA), commenting on the difficult cases, providing several tweets of example. 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.

While performing a linguistic analysis of the TWITTIRÒ corpus, we observed that syntax plays a significant role in the activation of irony in Italian, especially in social media short messages. Therefore, 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. The choice of this type of format comes firstly, because of its popularity within Computational Linguistics, as it has been already exploited in other evaluation campaigns (*EVALITA* and *CoNLL*). Secondly, its application on social media text has already been proven useful (Sanguinetti et al., 2017; Sanguinetti et al., in press 2018).

Together with the application of new morphological and syntactic labels, we are also planning to explore the activation of different types of irony (Sulis et al., 2016; Van Hee et al., 2016), such as the presence of situational irony, sarcasm and puns based on stereotypes.

Acknowledgments

C. Bosco, A. T. Cignarella and V. Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) and by Fondazione CRT (*Hate Speech and Social Media*, 2016.0688).

Bibliographical References

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016). Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of 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), Napoli, Italy, December 5-7, 2016.*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Basile, V., Bolioli, A., Nissim, M., Patti, V., and Rosso, P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 50–57, Pisa, Italy. Pisa University Press.

Farah Benamara, et al., editors. (2017). *Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017*. TALN 2017, Orléans.

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

Cignarella, A. T., Bosco, C., and Patti, V. (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), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Semeval-2015 task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of SemEval 2015, Co-located with NAACL*, page 470–478. ACL.

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

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

Hernández Farías, D. I., Patti, V., and Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technologies*, 16(3):19:1–19:24.

Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., and Belguith, L. H. (2015). Towards a contextual pragmatic model to detect irony in tweets. In *53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Karoui, J., Benamara, F., Moriceau, V., Patti, V., and Bosco, C. (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*.

Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the ICWSM: International AAAI Conference on Web and Social Media*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Maynard, D. and Funk, A. (2011). Automatic Detection of Political Opinions in Tweets. In *Proceedings of the ESWC: Extended Semantic Web Conference*.

Mihalcea, R. and Pulman, S. (2007). Characterizing Humour: An Exploration of Features in Humorous Texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*.

Reyes, A., Rosso, P., and Buscaldi, D. (2010). Finding Humour in the Blogosphere: the Role of Wordnet Resources. In *Proceedings of the 5th Global WordNet Conference*.

Reyes, A., Rosso, P., and Buscaldi, D. (2012). From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74:1–12.

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., and Tamburini, F. (2017). Annotating Italian social media texts in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 229–239.

Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., and Tamburini, F. (in press, 2018). PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of LREC18*.

Sperber, D. and Wilson, D. (1981). Irony and the Use-Mention Distinction. *Philosophy*, 3:143–184.

Stranisci, M., Bosco, C., Farías, D. I. H., and Patti, V. (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)*.

Sulis, E., Hernández Farías, D. I., Rosso, P., Patti, V., and Ruffo, G. (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.

Van Hee, C., Lefever, E., and Hoste, V. (2016). Exploring the realization of irony in Twitter data. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Van Hee, C., Lefever, E., and Hoste, V. (in press, 2018). SemEval-2018 Task 3: Irony detection in English Tweets. In *In Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA, June 2018.

Van Hee, C. (2017). *Can machines sense irony?: exploring automatic irony detection on social media*. Ph.D. thesis, Ghent University.

Wilson, D. and Sperber, D. (2007). On verbal irony. *Irony in language and thought*, pages 35–56.