Cognitive Computation manuscript No.
(will be inserted by the editor)

Detection of Sarcasm and Nastiness: New Resources for Spanish Language

Raquel Justo · José M. Alcaide · M. Inés Torres · Marilyn Walker

Received: date / Accepted: date

Abstract The main goal of this work is to provide the Cognitive Computing community with valuable resources to analyse and simulate the intentionality and/or emotions embedded in the language employed in social media. Specifically, it is focused on the Spanish language and online dialogues, leading to the creation of SOFOCO (Spanish Online Forums Corpus). It is the first Spanish corpus consisting of dialogic debates extracted from social media and it is annotated by means of crowdsourcing in order to carry out automatic analysis of subjective language forms, like sarcasm or nastiness. Furthermore, the annotators were also asked about the context need when taking a decision. In this way, the users' intentions

Raquel Justo
Universidad del País Vasco. (UPV/EHU). Sarriena s/n. Leioa 48940. Spain.
Tel.: +34-946013323
Fax: +34-946013071
E-mail: raquel.justo@ehu.eus

José M. Alcaide
Universidad del País Vasco. (UPV/EHU). Sarriena s/n. Leioa 48940. Spain.

M. Inés Torres
Universidad del País Vasco. (UPV/EHU). Sarriena s/n. Leioa 48940. Spain.

Marilyn Walker
University of California, Santa Cruz, 1156 N. High, SOE-3, Santa Cruz, CA 95064, USA

and their behaviour inside social networks can be better understood and more accurate text analysis is possible. An analysis of the annotation results is carried out and the reliability of the annotations is also explored. Additionally, sarcasm and nastiness detection results (around 0.76 F-Measure in both cases) are also reported. The obtained results show the presented corpus as a valuable resource that might be used in very diverse future work.

Keywords Online dialogues · Figurative language · Spanish resources · Sarcasm · Nastiness

1 Introduction

Cognitive linguistics views language as a form of communication that is embodied and situated in a specific environment [1]. According to Cognitive Linguistics human language is a cognitive capability that is inseparably intertwined with the way in which we interact with the environment [2]. This is one of the humans abilities that Cognitive Informatics investigates when creating numerical models to support the creation of artificial cognitive systems [1]. But language cannot be separated from the emotional status and intentionality during the cognitive process in human communication. Thus, language cues of intentionality and emotional status have to be automatically analyzed in order to develop artificial cognitive systems aimed to correctly interpret the human language. Moreover, the final goal of affective computing is to enable machines to perceive as well as to simulate human capabilities, like the ones inherent to language, during communication [1].

Automatic detection of peoples' emotions in what they say or write is possibly one of the most challenging problem that researchers on Artificial Cognitive Communication have to deal with. Primary emotions like fear, anger, sadness, disgust or joy, are the ones that we feel first as a reaction to external events. They have a direct impact in facial expressions, speech and biosignals [3] [4]. However it has been demonstrated that individuals from different cultural environments and languages, such as Italians and British, emphasize and prioritize different ways to

express emotions, such as facial expressions, voice inflections or body gestures [4]. Emotions also have a specific associated vocabulary, analyzed by both linguists and psychologists, that serves as a cue of how words express an individual's mood. Thus, emotional dictionaries are very useful for the automatic detection of primary emotions in social media. Secondary emotions, in contrast, come from reasoning about events. Since they are highly influenced by personal experiences, expectations, environment etc. the language, as a conceptual process, is a valuable cue in the automatic perception of these secondary emotions. However traditional Natural Language processing (NLP) barely copes with some aspects of language that require a cognitive and social perspective and suggest to shift from a word-based to a concept-based processing. This results in a multidisciplinary approach called *sentic computing*, which claims the importance of semantic features to study some aspects of cognitive communication trough the automatic analysis of the written language [5] [6] [7]. Moreover, different issues, like people's intentionality or sociocultural environment, can change the way in which they express themselves as well as the language people use to make an opinion explicit.

One of the current challenges facing NLP researchers is the extraction of subjective information from the huge amount of information residing in social networks, in online debate forums, opinion and discussion blogs, reviews of products and services, and microblogs. Most recent work aims to discover the contextual polarity of messages, information or reports, using new techniques in sentiment analysis and opinion mining. But current technology for automatic sentiment analysis performs poorly on ironic or sarcastic messages [8].

Subjective language forms, like sarcasm, irony or nastiness, that are commonly employed to emphasize or express intentionality, etc. are really difficult to detect and characterize. Actually, people seem to overestimate their ability to convey their intended tone — be it sarcastic, funny or serious — when they write an email as well. In fact they also overestimate their ability to correctly interpret the tone of the messages that others send to them. An experimental study found that only the

60% of participants were able to detect irony on emails they received whereas the 78% of the senders were confident about their ironic tone would be detected [9]. Therefore, subjective language can lead to misunderstanding in communication. Moreover irony, sarcasm or nastiness are very difficult to define as such and exhibit a high cultural dependence [4].

This work is focused on the automatic detection of subjective language forms like sarcasm and nastiness. However, it is not possible to define a set of words that are naturally associated with these specific language forms. In fact, it seems that more complex linguistic cues are needed for sarcasm [10] and nastiness [6]. All in all, the estimation of the percentage of the real use of subjective style in social media is an unachievable challenge. Moreover, such estimations necessarily show a great variety depending on the language form, idiom, media, topic, language, etc. Some machine learning experiments reported between 12% and 25% in sarcasm posts in dialogic forums or in twitter, for English language and specific corpora [11].

Ironic marks and indicators can also be defined from a pragmatic analysis of the language [12], as well as associated syntactic and semantic structures, changes of polarity or other statistical language features [13]; these have been demonstrated to be also useful for automatic detection of sarcasm [14, 15, 10]. Even though machine learning approaches have been extensively used to process these kinds of features. The most successful approaches have introduced some knowledge to the process in terms of syntactic templates [14], as well as part-of-speech and semantic, conceptual, and polarity information [15, 10, 13, 16, 8].

However, supervised methods in machine learning require annotated corpora, which entails an additional challenge of defining irony or sarcasm, even nastiness, sufficiently well to enable humans to judge it. Because humans themselves learn a model through social interaction over their course of their life, humans can understand and produce ironic sentences without a strict definition or a specific emotional vocabulary, specifying what is considered to be an ironic expression

[13]. A final challenge with developing corpus resources is that production and perception of social language is highly dependent on cultural norms. However the majority of research in disciplines like sentiment analysis addresses English. But the 48% of the internet contents are written in other languages, being Spanish language the second one among them. Thus there is a significant risk to miss essential information in texts written in other languages [17] for which there is an important lack of resources [7]. As a consequence an increasing interest arises to develop research and, consequently linguistic resources, for other languages [17].

In sum, human annotation of subjective language is needed for research, but it is difficult to provide robust definitions to judges and achieve high annotation reliability [18, 12, 19, 20]. To our knowledge the only annotated corpus of subjective language such as sarcasm or nastiness in dialog is the English Internet Argument Corpus (IAC) [21]. It consists of dialogic language processed from 4forums.net that has been annotated with figurative language tags such as sarcasm, irony and nastiness. Recently a self-annotated corpus for Sarcasm detection in English has been developed from Reddit[1] [11].

This paper is a first step towards the automatic analysis of subjective language in social media in Spanish. The main contributions of this work to the Cognitive Communication community are:

- the development of the first Spanish corpus consisting of dialogic debates extracted from social media. Our corpus SOFOCO (Spanish Online Forums Corpus) is focused on opinion mining and subjective language detection. It is a valuable new resource for the Cognitive Linguistic and Cognitive Informatics communities to investigate, analyze and simulate the intentionality and the emotions embedded in language. This resource is unique for Spanish language.
- an annotation procedure that involves the reader perception and judgement of sarcasm and nastiness of online dialogic debates.

[1] https://www.reddit.com

- statistics of the corpus that also include reliability for annotating sarcasm and nastiness in Spanish.
- experiments carried out by an artificial cognitive system that provides the automatic detection of the two language forms. Thus, a comparison between both, the natural and the artificial cognitive systems can be envisaged, to some extent.

Section 2 deals with sarcasm and nastiness language forms, which are the focus of the paper. Section 3 provides an overview of the related previous work, and Section 4 presents the details of how we obtained the SOFOCO corpus. Section 5 presents the annotation procedure and an analysis of the annotation results. Then Section 6 shows the baseline results of the automatic detection of sarcasm and nastiness. Finally the structure of the corpus and its annotated version are described in Section 7 and we conclude and discuss future work in Section 8.

2 Sarcasm and Nastiness

There are considerable differences in the degree of acceptance of ironic language in cultures as similar as those found in France, England, Germany or Spain is very different [22]. For the French, irony has a negative connotation because it is associated with ridicule; for the English, irony is an amusement that requires intelligence and imagination and to some extent some cunning and finesse; for German culture, the value of irony is predominantly negative because it is viewed as conveying hostility and open criticism, and finally in the Spanish culture, irony is viewed with some negative nuance, but like the English, it has a point of amusement, intelligence and even some cunning [22]. In the same way, the use of sarcasm has also a high dependence on cultural rules. Moreover, many ironic and sarcastic forms require sociocultural knowledge to detect e.g. “Ah, here comes the Spanish inquisition to save our souls”.

However, the boundaries between irony and sarcasm, or between sarcasm and satire, are not clear [19]. Some authors consider sarcasm more aggressive and of-

fensive than irony [23] or a class of irony attacking the image of the conversational partner [12], a negative irony [24], or an aggressive type of irony [19]. Sarcasm is sometimes defined as mocking, contemptuous, or ironic language intended to convey scorn or insult. In fact, while it may employ ambivalence, it is not necessarily ironic. For example, if a person requires a lot of time to obtain the solution of a very simple mathematical problem, one might ask "How many days does he need?". This is sarcastic but not ironic. However, "He is like Einstein" is a sarcastic sentence that uses irony. Ironic utterances, sometimes theorised as expressing the opposite of the actual situation, have been treated as polarity reversers [25,26]. The observations mean that it is typically difficult to achieve high levels of human agreement when labelling a turn as sarcastic in an online debate. Thus a major challenge is simply acquiring sufficient annotated data to use in order to develop automatic methods for detecting sarcasm in online dialogs [18]. As a result, there is relatively little prior work dealing with the automatic detection of sarcasm in social media in general, and even fewer studies applied to social media dialog [19].

Nastiness is another language form that provide valuable information related to the intentionality of the author. It is frequently associated to the use of insulting expressions or curse words, but it also depends on the sociocultural environment. In this way, the use of some kind of curse words is very extended in Spanish language and culture whereas the use of it is less frequent in English. For instance, in "I'm not angry. In fact, I'm laughing at you right now" a nasty tone can be perceived but no swear words are used. However, in "Actually, what they are really doing is ignoring silly irrelevant questions from clueless people with a religious agenda to promote" insulting words like "silly" and "clueless" are used to express nastiness. On the other hand, when referring to a third person, concept or idea, it is sometimes confusing whether a nasty tone is present or just a negative opinion is given ("People who complain most, have never played a videogame, they all are stuffed dinosaurs that live in another age"). Thus, although it seems easier to

identify at first, the presence of a nasty tone is also subjective and dependent on the cultural norms.

3 Related Work

The detection of subjective language forms is essential to the development of applications related to text analytics, because it helps in the identification of the intentionality of users. Text analysis typically makes use of resources such as annotated corpora, lexicons, and ontologies.

Knowledge-based approaches, relied mainly on the use of lexicons [27, 28] or other methods that do not need prior training [29]. There are many linguistic resources in English regarding affective and sentiment-based knowledge, like *WordNet-Affect* [30], *SenticNet* [31] (within the sentic computing framework), *SentiWordNet* [32] or *SentiSense* [33], that can be used to identify user's intentionality. Unfortunately, there are far fewer resources for the Spanish language. Some of them are listed below:

- A Spanish *WordNet* [34] was developed during the *EuroWordNet* project [35], but a version that includes affective knowledge is not available.
- The Spanish version of *SentiSense* was translated from English using Multilingual Central Repository (MRC) that integrates WordNet versions for five different languages: English, Spanish, Catalan, Basque and Galician.
- *EmoLib*[2] is a library that extracts the affect and emotions from an incoming text by tagging such text according to the feeling that is written or being conveyed. EmoLib is a flexible framework for building prototypes that allows studying the appropriateness of different strategies to label affect in text. It allows incorporating offline Linguistic knowledge derived from psychological studies as well as knowledge learnt from training examples.
- Díaz-Granjel et al. [36] developed a Spanish emotion lexicon, freely available, that presents 2036 words supplied with the Probability Factor of Affective use

[2] http://dtminredis.housing.salle.url.edu:8080/EmoLib/

(PFA) as the measure of their expression of basic emotions: joy, anger, fear, sadness, surprise, and disgust, on the scale of null, low, medium, or high.

- The lexicon developed by [37,38] for English, Spanish, Dutch and Italian was created to improve the performance of natural language analysis tools. It is a specialised lexicon to transform text from very informal language resources, like social networks, into more normalized forms. The use of it has proven to increase classification performance in an author profiling task.
- *MeSiento* Spanish corpus, available for the research community, was developed by [39,40]. They followed the same idea of *WeFeelFine* project [41] in English. This corpus was generated by collecting Spanish tweets containing the expression 'me siento"+ X, where X is a word supposed to be associated with a feeling.
- *Linguistic Inquiry and Word Count (LIWC)* dictionary [42] consists of a set of categories created to capture people's social and psychological states. Thus, they provide information related to the semantic meaning of the words, specifically the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech can be computed. In addition to the English version other dictionaries for different languages including Spanish are available.

Alternatively, machine-learning based approaches require the use of annotated corpora [43]. These annotations can be extracted from some kind of information provided by the author (specific tags in tweets, number of stars in product reviews, etc.), or by requiring the judgement of a natural cognitive system, that is, a human that provides a label without knowing the intentionality of the author a priori. This approach, although more costly, would be the most appropriate one when we want to design an artificial cognitive system that emulates the natural one.

Lots of annotated corpora can be found in English for different purposes. Corpora based on *product reviews* are often used in Sentiment Analysis [44,45]. They are also used in other applications related to subjective language forms detection

(like sarcasm) [46] although far fewer resources can be found in this case. But if we focus on Spanish the lack of resources becomes obvious. We can find some product review corpora, like the freely available SFU Review Corpus[3], for Spanish [47]. It consists of 400 reviews of cars, hotels, washing machines, books, cell phones, music, computers, and movies collected from the website Ciao.es, defined as positive or negative based on the number of stars given by the reviewer. There are other reviews corpora like CorpusCine [48,49] or HOpinion[4], that has also been used in different works [50,51] related to opinion mining. However, this kind of corpora should be judged or annotated by people who did not wrote the review, in order to be used for the detection of sarcasm or similar subjective language forms, in Spanish, and they do not exist, up to our knowledge.

The interest in carrying out text analytics on microblogs such as *Twitter* has greatly increased in the last years [52,53,54] due to the recent rise of social networks' use. When regarding sarcasm, irony or humor detection, hashtags like *#sarcasm*/*#sarcastic*, *#irony* or *#humor* are usually employed to get a labeled dataset [55,56]. However, a recent study found that only 45% of the utterances tagged as #sarcasm in a large corpus of Twitter utterances were judged by human annotators to be sarcastic without any prior context [25]. This suggests that data labeled through human perception experiments are still needed. Regarding Spanish language, one of the first corpus of tweets was built by [57] using the Twitter Search API[5]. It is composed of 34,634 tweets, of which 17,317 are considered as positive (those tweets that contain a positive emoticon such as *:-)*) and the other 17,317 are considered as negative (those tweets that contain a negative emoticon such as *:-(*). The use of hashtags to get a labeled dataset was also used for Spanish language in [58] for detecting ironic tweets (they consider sarcasm as a subclass of irony). Other kind of subjective language forms detection was carried out in [59, 60] where Twitter accounts of satiric newspapers like "El Mundo Today" or "El

[3] https:www.sfu.cam̃taboadaresearchSFU_Review_Corpus.html
[4] http://clic.ub.edu/corpus/hopinion
[5] http:dev.twitter.comdocgetsearch

Jueves" vs. not satirical ones like "El Mundo" or "El Pais" were used to build an annotated corpus devoted to the detection of satirical vs. not satirical Tweets. Another remarkable corpus is the one provided by TASS within the framework of SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) annual conference that is focused on sentiment and online reputation analysis. The last one was held in September 2016 [61]. The annotation was carried out semi-automatically, that is, a baseline machine learning model was first run and then all tags were checked by human experts.

Online forums provide another interactive form of online communication. There, users tend to express their opinions with highly subjective and often emotional language about different debate topics. Thus, accurate systems that could derive information from these resources could lead to interesting insights about people's opinions in a great variety of topics. This task differs from Twitter-based ones, because the discussions and each individual post in a discussion is usually longer, and context is used to a greater extent. Moreover, the dialogic nature of the utterances in online discussions also leads to more colloquial vocabulary and language style, especially when compared to product reviews. Internet Argument Corpus (IAC) [62] is a publicly available collection of 390,704 posts, written in English, in 11,800 discussions extracted from the online debate site 4forums.com. IAC is a very valuable resource due to the annotation process that was carried out using Amazon's Mechanical Turk. Turker's were asked about different emotions like Sarcasm or Nastiness. They were also asked about different aspects like agreement/disagreement of a post with respect to the prior one, or about the nature of the responses when regarding factual or emotional argument. Different works related to the detection of disagreement [63] or nastiness and sarcasm [14, 15] has been carried out using this corpus. Other applications like the detection of argument facets [64] or stance classification were also carried out with IAC [65]. On the other hand, an English self-annotated Reddit corpus (SARC) was recently presented in [11] for the detection of sarcastic posts. Reddit is a social media

site in which users communicate by commenting on submissions, which are titled posts consisting of embedded media, external links, and/or text, that are posted on topic-specific forums known as subreddits. Users on Reddit have adopted a common method for sarcasm annotation consisting of adding the marker "/s" to the end of sarcastic statements. However, as with Twitter hashtags, using these markers as indicators of sarcasm is noisy, specially since many users do not make use of the marker, do not know about it, or only use it where sarcastic intent is not otherwise obvious.

To the best of our knowledge, there is no annotated corpus of *online dialogues* in Spanish. Therefore, our work focuses on the development of such a corpus where subjective language will be annotated. The idea is to get a parallel Spanish corpus to the IAC, that would be useful to train a system capable of detecting subjective language forms like sarcasm or nastiness in Spanish, in the same way as people do. An additional benefit is that the dialogic nature of this data will also allow us to explore the impact of interactions between the participants.

4 Features of the Spanish Online Forum Corpus (Sofoco)

Here we describe the source of the corpus data and the way in which we curated it, and obtained annotations.

4.1 Overview

Sofoco draws on the website *Menéame*[6]. Launched in December 2005, Menéame is a social news aggregation and discussion site, initially modelled after Digg and quite similar to Reddit. Menéame is the most popular Spanish social news aggregator: about 25,000 users posted 1,200,000 posts to 17,500 published stories during 2014. More than 80% of visitors come from Spain; therefore, it is highly Spain-centric about discussion topics and cultural references, and also in language

[6] www.meneame.net

style. As of april 2017, Menéame is ranked at 130th place in Spain by web analytics company Alexa among all kinds of web sites (search engines, social networks, newspapers, etc.). We selected this website because the authors' profile tends to be quite specific in terms of education and cultural background. In addition, social news aggregators are best suited for our goals: news about hot and controversial topics usually earn more votes, gaining visibility and, in turn, a higher number of comments.

It should be noted that debating skills are not highly regarded in Spanish culture and education. As a consequence, debate forums (particularly those devoted to political discussions) are very uncommon and unpopular. In fact, the most popular discussion-oriented site in Spain is ForoCoches.com (ranked 45th by Alexa as of april 2017). This site contains a huge, unrated and uncategorized variety of discussions, often about vulgar topics, with many posts showing a rude and insulting tone. We ruled out this site as a source for SOFOCO in despite of its popularity because we are mainly interested in sarcasm (in addition to nastiness). Sarcasm is usually related to a higher cultural level, and more commonly found in discussions about specific topics.

Menéame works in a similar fashion to other social news aggregators such as Reddit. Registered users can submit content in the form of *stories*: links to news published on other web sites accompanied by a short comment. Other users – registered or not– can vote the stories ("menear" – "shake" in the site's jargon) in order to promote (*publish*) them to the main page. Registered users can post their own *comments* on each story, resulting in a comment thread. Published stories usually have longer comment threads containing more discussion and debate. (We employ the term "post" in place of "comment" in SOFOCO and, accordingly, in this paper we will use "post" when discussing Corpus details.)

Stories submitted by users of Menéame are categorized into different topics such as politics, sports, technology, etc. The number of topics grew over time up to a point where a shallow tree of categories was introduced. However, topics

are imprecisely defined, some being too broad in scope while others being very limited. Users also label the stories they submit with a number of freely chosen *tags*. An analysis of the tags borne by published stories clearly shows that those related to Spanish politics are the most frequent by far; however, tags themselves do not provide a precise topic categorization. Therefore we decided neither rely on Menéame's topic categorization nor stories' tags, defining instead our own topics using search queries (subsection 4.2).

Both users and comments are given *karma* values. A comment's karma is derived from the positive and negative votes obtained by that comment; therefore it is useful as a measure of the polarity and intensity of user reaction to any comment. Comments with high karma are highlighted in the story's comment thread, while those with very low karma are hidden. On the other hand, a user's karma is calculated from the votes given to all comments posted by that user, so it works as a measure of user reputation. A minimum user karma is required in order to submit stories, send comments and vote on comments (each vote is weighted based on the voter's karma). An example of the comments related to a new story are shown in Figure 1.

Menéame's source code is licensed under GNU GPL and fully available. In addition, the managers kindly provided us with a partial copy of the site's database. Database tables store full information about stories, posts, topic categories, and users (excluding personal data in order to comply with privacy regulations). These include IDs, timestamps, karma values, votes, keywords, etc.

We did not carry out a conventional scraping of Menéame website; in contrast, we devised a procedure for finding, retrieving and processing stories and posts. With this aim, a set of software tools were developed for performing online accurate topic searches, selecting stories from the search results, and then retrieving the contents of the selected stories and their post trees. These tools take advantage of many implementation details of Menéame's source code: search queries are built specifically for, and passed to the Sphinx search engine which Menéame employs;

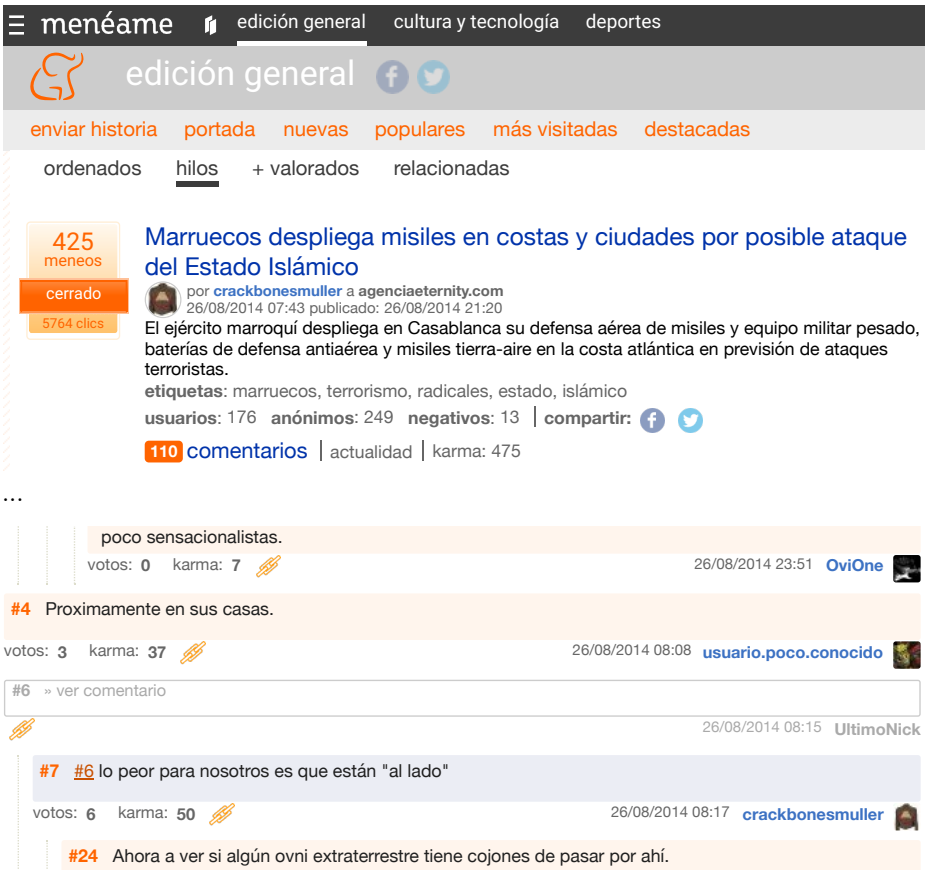

Fig. 1 Example of a new story from Menéame and some posts related to it. STORY: El ejército marroquí despliega en Casablanca su defensa aérea de misiles y equipo militar pesado, baterías de defensa antiaérea y misiles tierra-aire en la costa atlántica en previsión de ataques terroristas., translated into English: The Moroccan army deploys in Casablanca air defence missiles and heavy military equipment, anti-aircraft missile systems and surface-to-air missiles on the atlantic coast in anticipation of terrorist attacks. POSTS RELATED TO IT: #4 Próximamente en sus casas., translated into English: #4 Coming soon to your home., #7 #6 lo peor para nosotros es que están "al lado". translated into English: #7 #6 the worst thing for us is that they are next door., #24 Ahora a ver si algún ovni extraterrestre tiene cojones de pasar por ahí., translated into English: #24 Now, let's see if any alien UFO has the guts to go through.

full texts of the selected stories and their post threads are downloaded via remote execution of the appropriate scripts. The Menéame database is used for building and retrieving the post tree structure for each selected story, and getting all kind of story and post metadata (author, karma, etc.).

Table 1 Posts extracted from *Menéame* that comprise the SOFOCO corpus. Some posts appeared in different topics but are only considered once in the overall count.

Topic	Stories	Posts	
		total	average
Terrorism	280	25,109	90
Abortion	146	14,522	100
Indep. of Catalonia	84	17,408	207
Gay Marriage	185	14,191	77
Creationism	40	3,559	89
Overall	728	73,985	102

4.2 Topic selection

For the design of the corpus we selected a set of five controversial topics: *Terrorism*, *Independence of Catalonia*, *Abortion*, *Gay Marriage* and *Creationism*. The Terrorism topic was limited to that related to jihadist origin or related to ETA (a terrorist organization active for decades in Spain). Three of the chosen topics (Abortion, Gay Marriage and Creationism) are shared with the IAC [18]. Other topics included in the IAC do not exhibit the same level of contentiousness in Spain. However, we must point out that Creationism is not as controversial in Europe as it is in United States; we chose it because some posts from Menéame show humorous reactions, which could display a higher usage of sarcasm.

As previously stated, we employ Menéame's Sphinx search engine for finding stories related to a topic in the period between September 2009 and March 2015. Sphinx's extended query language is quite powerful and versatile, allowing accurate search queries.[7] Matches are performed on the stories' title, keywords and text, and results are ranked by phrase proximity and term frequency. A computer program has been devised for submitting search queries and retrieving the stories found, automatically completing them with relevant associated data and metadata from our local copy of Menéame's database.

A search query was constructed for each topic, saving the stories found along with their associated data into the Corpus database. By virtue of being performed

[7] http://sphinxsearch.com/docs/latest/extended-syntax.html

online, this procedure can be repeated every time a new topic needs to be added to the corpus. In Table 1 the number of stories found for each topic and their respective total and average number of posts (i.e. thread length) are shown. Using only the five topics, a total of 728 stories matching the search criteria were found, adding up to a total of roughly 74,000 posts. These posts represent 1.3% of all posts related to the stories published in the searched period.

5 Annotations

Once the selected set of posts was retrieved from the *Menéame* website we used a custom crowdsourcing platform[8] to carry out the annotation tasks. Amazon's Mechanical Turk is commonly used for this kind of task, but it is not available in Spain or, to our knowledge, in European countries either [66]. Other platforms like CrowdFlowder can provide Spanish annotators, but only around 19% of contributors who speaks Spanish are from Spain[9]. We were interested in having as much different judges from Spain as possible, because the differences with regard to american Spanish are very noticeable, even further when considering subjective language forms that are highly culture dependent. Additionally, we wanted to have a controlled set of annotators at first, so we decided to develop our own platform.

Since our work currently focuses on the annotation of nastiness and sarcasm, the annotators were asked to evaluate the sarcastic and nasty tone in the presented post. They were not provided with previous definitions for these language forms; we decided to allow them to use their own interpretation of the terms instead. They had to evaluate the presence or absence of sarcastic tone (binary answer) while a three grade scale was provided for nastiness. This was because of the notion that the annotation score can reflect how nasty a comment can be according to the employed language.

[8] http://cz.efaber.net
[9] https://www.crowdflower.com/crowdflower-now-offering-twelve-language-skill-groups/

When collecting annotations for sarcasm, contextual information might be needed to make annotation decisions as to whether a post is sarcastic. For online dialogues, there are several different kinds of contextual information that could possibly be taken into account: 1) the context given by the previous post or posts, e.g. "you are early" will be interpreted as sarcastic in a context where the addressee is clearly late, 2) the context associated to sociocultural knowledge, e.g. the sarcastic form in "The Nazis really brought good to the world", 3) the context associated to a specific author (some authors tend to write in a more sarcastic or ironic mode). In our work, we asked the annotators about the need for context, but we explicitly state that the context is that provided by the previous post. Let us note that the dialogic nature of the data allows us to study whether there are turns in a dialog (a post in this case) that can only be interpreted as sarcastic when considering the information of the previous turn (or post).

Thus when an annotator begins the task, they are provided with both the post to be annotated and the context i.e. the previous post in the tree for which the current post is providing an answer (it includes the source story for comments in the first level of the hierarchy). This previous post is initially hidden but it can be displayed by clicking on it when necessary. The annotator then answers the following questions for each post:

- Evaluate the nastiness tone (0-2) of the post:
  - 0 - not nasty at all
  - 1 - a bit nasty
  - 2 - very nasty
- Indicate whether a sarcastic tone is present in the post:
  - yes
  - no
- Indicate whether the context given by the previous post has been decisive to answer the previous question:
  - yes
  - no

Using the aforementioned platform, we defined an annotation task consisting of 13,717 posts that were selected to have a balanced number of topics as shown

Table 2 Features of the annotated set per topic: Number of Stories, Posts, Authors and Posts per Author.

Topic	Stories	Posts	Authors	P./A.
Terrorism	35	2.251	1.076	2,09
Abortion	44	2.816	1.497	1,88
Indep. of Catalonia	36	4.182	1.475	2,84
Gay Marriage	38	2.247	1.177	1,91
Creation.	35	2.384	1.222	1,95
Overall	185	13.717	4.338	3,16

in Table 2. From this set we first took a subset of 250 posts to carry out a pilot task. Then, the annotation task was extended to the crowd. Henceforth we will refer to these data sets as *Pilot-Phase* (250 posts) and *Crowd-Phase* (13,717 posts) respectively.

5.1 *Pilot-Phase* Annotations

In the *Pilot-Phase* two sets of annotators were involved: *Trial* and *Ling* sets. *Trial* set was designed to have a reliable set of annotators. The idea was to receive the feedback of these annotators to be able to improve the potential problems of the annotation process. It is comprised by 16 annotators that were selected among colleagues of our own department. Thus, although the majority of them are not familiar with annotation and machine learning techniques, all of them have scientific skills and are familiar with research processes related to electronics, robotics, and computation. Since they are close colleagues they were motivated and could be expected to do their best to provide correct annotations.

Ling set consists of 17 annotators with linguistic training. The idea was to have a set of annotators with a deeper knowledge of the way in which people express their intentions (irony, sarcasm, nastiness,...) in Spanish dialogic language. All of the annotators in this set are graduates in any field related to linguistics and some of them are also postgraduate students in this field.

The number of annotators required for each post in order to get reliable annotations is a significant issue. A compromise is needed between reliability and effort, in terms of time and expense. According to the work carried out with IAC [18], in most cases only 3 annotations are needed and this number can increase up to 7 when the initial judgments are highly ambiguous. Thus, in this phase we considered 3 annotations per post.

At this point the reliability of the annotations wanted to be anlayzed in terms of annotators, task and items.

First, the following measures were collected for each annotator in a set: the number of annotated posts, the employed time (average, median and standard deviation), the nastiness average value given to the annotated posts, the percentage of posts annotated as sarcastic by each annotator and the percentage of posts for which context was needed in order to decide whether the post was sarcastic or not. Table 3 shows the average values of these measures when considering all the annotators of each set.

These results show that annotators from the *Trial* set used more time than the *Ling* annotators to supply their annotations. This could be due to the specific profile of the annotators. *Ling* ones are more used to this kind of annotation task, and also in general, more accustomed to analysing language. *Trial* annotators are familiar with scientific processes and they carried out the task carefully, additionally they are not used to doing language analysis. When considering the percentage of sarcasm a high value is obtained with both sets of annotators (a bit higher for *Ling* ones). It seems that these results might be slightly biased by the question carried out, that is, when asking about sarcasm, annotators put special care in detecting it and they found sarcastic some posts that would not be perceived in that way in another scenario. Focusing on nastiness average value, although *Ling* annotators provided higher values, a similar result is obtained in the two sets.

Then, an additional set of measures was also collected, with the two sets of annotators, for each post: average time (of 3 annotations), no. words, nastiness

Table 3 Average values of the statistics per annotator when considering *Pilot-Phase*.

	Statistics per annotator						
	no. posts	time			Nast. avg.	Sarc. (%)	Cont. (%)
		avg.	median	dev.			
Trial	46.87	62.35	51.38	38.95	0.54	44.05	42.44
Ling	44.12	38.31	32.83	22.35	0.64	48.18	38.18

Table 4 Average values of the statistics per post when considering *Pilot-Phase*.

	Statistics per post					
	no. posts	time avg.	no words	nastiness avg.	Sarc. (%)	Sarc. Cont. Need (%)
Trial	250	57.58	60.21	0.53	22.00	18.18
Ling	250	41.62	60.21	0.60	18.00	6.67

average value (3 annotations), percentage of sarcastic posts (considering only sarcastic, those posts labeled as sarcastic by all the 3 annotators) and the percentage of sarcastic posts for which the context was needed (the agreement of all the 3 annotators was also required). Note that in this case the context need is a percentage of those posts labeled as sarcastic. The analysis per post (see Table 4) shows that the nastiness values given by the two sets are very similar, thus, it seems that the overall posts might have a slight nuance of nastiness but are not really nasty (results in the 0-1 interval). Note that the corpus was selected to avoid an abusive use of nasty tone. Regarding the sarcasm, the results in Table 4 show that if we require that all 3 annotators agree, then the percentage of sarcastic posts would not be very high. Thus, although each annotator labels as sarcastic almost the 50% of the posts, the biased result is corrected by requiring the agreement of different annotators.

The agreement percentages among the users, when regarding sarcasm, was also computed. Table 5 shows the number and percentages of posts for which 3, 2, 1 or 0 annotators said it was sarcastic when considering *Trial* and *Ling* sets. The sets where agreement for all the annotators (3/3 and 0/3) was achieved are highlighted in grey. The results in this table show that there is not a total agreement between the different annotators even when considering the *Ling* set (only 44.0%

was achieved). Furthermore, surprisingly, higher percentages of agreement are obtained with annotators from the *Trial* set (51.2%). The two sets of annotators were also used to measure the inter-annotator agreement in terms of Krippendorff's α. We obtained a value of α =0.34 for the *Trial* set and a value of α =0.24 for the *Ling* set. It is quite clear that the agreement of the *Trial* set is higher as mentioned above. This fact suggests that the *Trial* was a reliable set of annotators. Besides, it seems that linguistic knowledge can lead to a higher disagreement when annotating subjective language for which precise definitions do not exist. Thus, the task is susceptible to being extended to the crowd without specific linguistic knowledge requirement.

Table 6 shows the number of posts that were annotated in the same way by the annotators in the two sets. These measures show that there are more posts annotated in the same way within the total agreement sets (3/3 and 0/3). Thus, the

Table 5 Statistics of the agreement per post for sarcasm issue. Total agreement is the sum of 0/3 + 3/3.

	Trial		*Ling*	
Sarc. Agr.	no. posts	% posts	no. posts	% posts
3/3	55	22.0	45	18.0
2/3	49	19.6	52	20.8
1/3	73	29.3	88	35.2
0/3	73	29.2	65	26.0
Tot. Agr.	128	51.2	110	44.0

Table 6 Statistics for the posts labeled in the same way by *Trial* and *Ling* sets. Total agreement is the sum of 0/6 + 6/6.

	Common	
Sarc. Agr.	no. posts	% posts
6/6	25	10.0
4/6	10	4.0
2/6	26	10.4
0/6	31	12.4
Total Agr.	56	22.4

Table 7 Different features of the Crowd set: Sex, Education Level (Primary (P), Secundary (S), Superior (sp)), Age and University Student (Yes/No).

Sex		Education			Age			Stud.	
M	F	P	S	Sp	>45	25-45	<25	Y	N
103	104	7	85	115	50	60	97	99	108

posts in these sets seem to be easier to annotate. Some examples of the posts in 6/6 sets are: *"no sé si has leído que pone foto ilustrativa"* translated into English: *"I don't know if you read: illustrative photo"* or *"También sabemos que Dios perdonó a algunos de los dinosaurios durante el diluvio. Hilarante!"* translated into English: *We also know that God forgave some dinosaurs during the Flood. Hilarious!.*. In the same way 0/6 set have clear not sarcastic examples like: *"En el punto en el que estar en contra del matrimonio homosexual le quite más votos de los que le da. Entonces ceder como ha hecho la derecha toda la vida, primero el divorcio, luego los anticonceptivos, es cuestión de tiempo"* translated into English: *When to be against gay marriage takes less votes than the contrary. Then, they will give up, as the right wing has always done, first with the divorce, then with contraceptive methods, it's a matter of time.*

5.2 *Crowd-Phase* Annotations

The *Crowd-Phase* was carried out by a heterogeneous set consisting of 207 annotators fluent in European Spanish of which 176 annotated more than 5 posts. The specific features of the annotators in the set are summarised in Table 7. Regarding the number of annotations per post, given the ambiguity of the task, the influence of the agreement observed in *Pilot-Phase* (see Section 5.1) and the not controlled annotators in the *Crowd* set, 5 annotations per post were required in this case.

Statistics per annotator, described in Sec. 5.1, were also computed here as shown in Table 8. Comparing the results in Table 8 with those obtained in Table 3 it can be concluded that *Crowd* annotators needed more time than *Ling* ones but less than *Pilots*. This might be due to the low expertise and high diversity of this

set. Regarding the average nastiness, a similar value was given by the *Crowd* set, comparing it to the *Trial* set, but it was lower than the one given by *Linguistics*. Finally, the percentage of sarcastic posts labeled by each annotator is a high and similar value for the three sets, although the *Crowd* provided the lowest one. Thus, it seems that the *Crowd* tends to moderate the border effects that might appear in very specific sets of annotators. In this case, Krippendorf's α was also measured and a value of 0.24 was achieved, that is the same value obtained for *Ling* set. This means that *Crowd* annotators are still a reliable set of annotators. This measure is also employed to evaluate the reliability of the task. Although the values achieved in this work are low, they are similar to those obtained with IAC corpus (0.22) in [18], where they claim that it is unclear what this means for subjective tasks such as sarcasm annotations. An in depth analysis related to the quality of the annotations carried out in this work was presented in [67]. Furthermore, similar values were also achieved when considering different tasks for emotionally annotating synthesised speech [68] where subjectivity is also present.

When regarding statistics per post, some measures like average time (of 5 annotations), no. words and nastiness average value (5 annotations) were also collected and are given in Table 9. These results show that *Crowd* annotators needed lower periods of time than the previous ones to carry out the annotations. This might be due to the lower lengths of the posts that are also reflected in the table. However, the reduction in time seems much more noticeable, thus, we hypothesize that in the *Crowd* set there are annotators that try to carry out the annotations as fast as they can instead of doing the work carefully.

Table 8 Average values of the statistics per annotator when considering *Crowd-Phase*.

	Statistics per annotator						
	no.	time			Nast.	Sarc.	Cont.
	posts	avg.	med.	dev.	avg.	(%)	(%)
Crowd	66,26	47.41	37.31	36.95	0.53	42.77	36.17

Table 9 Average values of the statistics per post, related to time and post lengths, when considering *Crowd-Phase*.

	Statistics per post		
	no. posts	time avg.	no words
Crowd	13,717	33.26	58.52

Table 10 Average values of the statistics per post, related to percentage of sarcastic and nasty posts, when considering *Crowd-Phase* and different agreement requirements (3, 4 or 5 annotator for sarcasm and average values higher than 0.6 or 1 for nastiness).

	Statistics per post for different agr. requirements				
Sarcasm	>= 3	28.93 %	Context	>= 3	19.70 %
				>= 4	4.31 %
				>= 5	0.58 %
	>= 4	13.46 %	Context	>= 3	28.98%
				>= 4	9.26 %
				>= 5	1.25 %
	>= 5	4.47 %	Context	>= 3	36.37%
				>= 4	14.85 %
				>= 5	3.75 %

Then, an additional set of measures was also collected, nastiness average value, percentage of nasty posts (considering nasty, those posts with an average nasty value higher than 0.6 or higher than 1) percentage of sarcastic posts (considering sarcastic, those posts labeled as sarcastic by at least 3, 4 or 5 annotators) and the percentage of sarcastic posts for which the context was needed (the agreement of 3, 4 or 5 annotators was also required). The results are shown in Table 10. Note that in this case, since 5 annotations were gathered for each posts, we cannot do fair comparisons with the previous sets. According to the results, it seems that when a higher agreement is required the percentage of sarcastic or nasty posts decreases as expected. The requirement of 4 annotators agreement for sarcasm provide the most similar result if we compare it with IAC (12%). When the context need is analysed it can be concluded that if the same agreement criterion (>= 3, >= 4 or = 5), employed for sarcasm, is used, the percentage of sarcastic posts that needed context decreases drastically from 19.70% to 9.26% and 3.75%. Thus, it seems that the agreement among annotators is more difficult to obtain when considering context

Table 11 Statistics of the agreement per post when regarding sarcasm issue for *Crowd* set.

Agree. ppost (Sarcasm)	posts (%)	no. words	Avg. nastiness
5/5	4.47	26.74	0.40
4/5	9.00	30.39	0.36
3/5	15.48	42.44	0.39
2/5	18.05	52.68	0.31
1/5	26.16	66.97	0.25
0/5	26.85	78.20	0.16

need than when considering sarcasm. Additionally, when keeping the context need agreement criterion fixed, >= 3 for instance, the percentage of sarcastic posts that needed context increases when changing agreement for sarcasm from >= 3 to = 5, from 19.70% to 28.98% and 36.37%. Thus, it seems that the posts that are clearly sarcastic (the easiest ones to detect) needed more context because are presumably shorter and more dependent on the previous post.

The agreement percentages among the users, when regarding sarcasm, was also computed for the *Crowd* set. Table 11 shows the number and percentages of posts for which 5, 4, 3, 2, 1 or 0 annotators said it was sarcastic. When focusing on the number of words in each post an interesting issue is revealed by the obtained results. It is quite clear that not sarcastic posts (0/5) are longer than sarcastic ones (5/5), indicating that when authors wanted to be sarcastic they tend to employ fewer words, while not sarcastic posts are more argumentative and longer in Spanish. For instance, an example of a post labeled as not sarcastic by all the 5 annotators where the poster explains a point of view would be *"Porque desde la transición la derecha necesitó de la extrema derecha cuando la izquierda era aplastante mayoría en Valencia. Se inventaron que la izquierda era una aliada a los catalanes, que querían invadir Valencia y eso fue calando en el imaginario colectivo. Algo totalmente ridículo."*, translated into English: *"Because from transition times the right wing needed extreme right wing when left wing had a majority in Valencia. They invented that the left wing was an ally of catalonians that wanted to invade Valencia and that was permeating the collective imagination'. Something*

Table 12 Annotation results related to nastiness average value and percentage of sarcastic posts, when considering different agreement requirements, for different topics.

Annotated Corpus	Nastiness	Sarcasm (%)		
		>= 3	>= 4	= 5
Terrorism	0.32	33.14	15.55	5.24
Abortion	0.30	31.25	14.88	4.97
Indep. of Catalonia	0.22	18.53	7.27	2.05
Gay Marriage	0.32	33.69	15.84	5.21
Creationism	0.24	35.53	18.12	6.50
Overall	0.28	28.93	13.49	4.46

completely ridiculous.' . This contrasts with a much shorter sarcastic one: *"Digamos que es la versión bíblica de - lo hizo un mago -"*, translated into English: *"Let's say that it is the biblical version of - a magician did it -"*

Additionally, the relationship between nastiness and sarcasm was studied. Thus, the average nastiness values of the different sarcasm agreement sets (5/5, 4/5, 3/5, 2/5, 1/5, and 0/5) were compared to each other in Table 11. The obtained results show a quite clear correlation between the two language forms, since sarcastic posts (5/5) show the highest nastiness average value and the clearly not sarcastic posts (0/5) the lowest one.

Then, the nastiness value and the percentage of sarcastic posts were computed for the different topics that were involved in the corpus. The results are given in Table 12. These results show that people tend to be nastier when talking about Abortion, Gay Marriage and Terrorism and nicer when talking about Creationism and Independence of Catalonia. This might be due to the sociocultural issues and the fact that there are not so many controversial laws in Spain related to Creationism as they are related to Terrorism or Abortion. Thus, people talking about it do not got so involved, keep calm and do not employ a nasty language. Looking at sarcasm issue, this topic shows the highest percentage instead. This might be due to the same reason, that is, the humour associated to the use of sarcasm is often present when talking about Creationism. However, authors seem to be very sarcastic also regarding Terrorism, Gay Marriage and Abortion. This

might be related to the cruelty and the intention to offend that is also present in sarcasm. It is also interesting to notice that the less sarcastic posts with lower nastiness values are related to Creationism and Catalonian Independence. This effect could be explained again by the sociocultural reasons and by the origin of some annotators that in the first stage were most of them from the Basque Country. Let us note that in the Basque Country also exists a pro-independence movement and this might be reflected in the annotator set.

6 Detection of Sarcasm and Nastiness

The annotations described in the previous section can be used to train a system devoted to the detection of sarcastic and nasty posts. Thus, balanced sets of sarcastic vs. not sarcastic and nice vs nasty posts have to be defined for the training procedure. When focusing on sarcasm we built two training sets, in *SSet1* we considered sarcastic those posts labeled in that way by at least 3 annotators, whereas in *SSet2* we considered sarcastic those posts labeled in that way by at least 4 annotators. The agreement of all the annotators (5) was considered too restrictive and we did not build a set with this requirement.

Tables 13 and 10 revealed that when requiring the agreement of at least 4 annotators the percentage of sarcastic posts drops significantly from 28.93 to 13.49. Thus, we have a smaller *SSet2* with higher reliability that might be used as a seed to provide automatically annotated bigger training sets, for instance. Additionally, a bigger *SSet1* might be used when more annotated data is needed although the reliability is lower. The same procedure was carried out in order to obtain a balanced set of nice vs. nasty posts. In such case we considered as nasty those posts for which the average value of nastiness given by the different annotators is higher than 0.6 (*NSet1*) or higher than 1 (*NSet2*). Let us note that the average nastiness value per post was 0.27, as Table 10 shows, thus a value of 0.6 is quite high.

A baseline experiment was carried out to detect sarcasm and nastiness using different sets for comparison purposes. The n-grams inside the posts were selected

Table 13 Percentage of sarcastic and nasty posts and the number of posts in balanced sets for different agreement requirements (>= 3, >= 4, = 5), when considering sarcasm, and different threshold values (0.6 or 1) for average nastiness value.

	Sarcasm		Nastiness	
	SSet1	*SSet1*	*NSet1*	*NSet2*
Sarc. Agree./Nast. Avg.	>= 3	>= 4	>= 0.6	>= 1
Sarc./Nast. (%)	28.93	13.46	19.49	6.75
No. post balanced set	7938	3692	5346	1852

as features with $n = 1, 2, 3$. A Multinomial Naive Bayes classifier was built and the 10-cross validation results are shown in Table 14. It can be concluded that *SSet2* and *NSet2* provide significantly better results than *SSet1* and *NSet1* in terms of all the measures (F-measure, Accuracy, Precision and Recall). This fact reveals the impact that the annotation procedure (the selected interannotator agreement in this case) has in the obtained results, that might be even more relevant than the classification procedure or the selected feature set. When considering IAC corpus and the same feature set and classification procedure [15] similar but slightly lower results were achieved for sarcasm detection (F=0.69) and slightly higher for nastiness detection (F=0.79). This means that our annotated corpus is appropriate for sarcasm and nastiness detection and it is similar to the English IAC one as expected. However, it seems that nastiness is less frequent in SOFOCO than it was in IAC. In fact, the vocabulary employed is not as nasty as expected considering that the Spanish language has a lot of forms of nastiness that are frequently employed. Thus, the nastiness perception might be more ambiguous and the system performance a bit lower than it was with IAC. This might be because the corpus was specifically chosen not to be specially insulting as the majority online forums and in order to have a higher sarcasm presence.

It can be concluded that an appropriate selection of the corpus within the corresponding sociocultural environment is essential to get valuable results, when an interesting percentage of the language form to be detected (sarcasm and nastiness in this case) can be found. Additionally, the annotated corpus with the appropriate interannotator agreement is a very valuable resource to carry out data-driven

Table 14 Classification results for Sarcasm and Nastiness with MNB classifier.

	Sarcasm		Nastiness	
	SSet1	*SSet2*	*NSet1*	*NSet2*
F-measure	0.69	0.73	0.69	0.74
Accuracy (%)	71.15	74.91	67.84	71.22
Precision (%)	74.4	78.26	66.20	67.63
Recall (%)	64.5	71.2	73.51	81.74

approaches. In fact, the description of the procedure itself might be essential for the development of similar resources in different languages and cultures.

7 Corpus Structure and Distribution

Although both the corpus and the annotations, in their present state, are not yet ready to be granted open access for the scientific community, we are working toward this goal. In the meantime, we might provide a copy of them upon request[10], including full information about its organization and structure.

The key piece of SOFOCO is a file which contains a list of pairs composed by each one of the topic search strings and the list of stories (their link IDs) found for that topic, also including the number of posts associated to each story.

The core of the Corpus consists of a directory containing a file for each one of the stories selected for annotation. These files store information about each particular post associated to the story (its text, author, karma, place in the thread, etc.), and also contains a list of pairs [previous post, current post] which represents the tree of comments.

In order to track each post along the annotation tasks, a unique post tracking label is built in the form "link ID-previous post-current post": for example, `2074471-15-42` identifies the post number 42 in the comment thread of story with link id 2074471, being itself one of the answers to the post number 15 in the same thread. In the annotation task, the current post provides the text to annotate, while the previous post provides the context.

[10] raquel.justo@ehu.eus

In another directory, a file for each post stores the post's text segmented into sentences, which are in turn tokenized into words. Text parsing was carried out using the FreeLing toolkit [69]. These files could potentially store additional data resulting from further text analysis.

Finally, the results of each annotation task are stored into its own file. These files contain a list of all forms completed by the annotators, each one identified by its post tracking label, and including the annotator ID, completion time, date and time of submission, and the answers to the three questions about nastiness, sarcastic tone and context need. An example of an annotation stored in such a file is:

```
time: 63.059819,
finished_at: "2015-09-21T14:43:22.960+02:00",
pots text: "Esto no va a quedar asi",
worker id: 24,
post tracking label: "2074471-15-42",
answers: [ { "value": "1 - a bit nasty" },
           { "value": "No"},
           { "value": "No"} ]
```

Since each post was annotated by 5 different annotators there would be other 4 annotations with the same post id, and post text that might have different answers, time, worker id, and finish date. be other 4 annotations with the same post id, and post text that might have different answers, time, worker id, and finish date.

8 Concluding Remarks and Future Work

The analysis of social network is a difficult task that can take benefit from the development of different resources like annotated corpora. In this work, the Spanish Online Forum Corpus (SOFOCO) was developed. To our knowledge SOFOCO is the first Spanish corpus consisting in debate turns extracted from the internet devoted to opinion mining and subjective language identification.

Nastiness and sarcasm were annotated trough human perception experiments using a custom crowdsourcing platform. Annotators also provided information about the context needed to take their decisions, during the defined crowdtask. Thus, SOFOCO is also the first corpus with Sarcasm and Nastiness annotations in Spanish Language. The analysis carried out show that crowdsourcing is an appropriate annotation strategy and that the annotation task is reliable enough. Using the aforementioned annotations different balanced training sets for the detection of sarcasm and nastiness were built. The classification results show that the presented corpus is a valuable resource for data driven approaches.

SOFOCO might be used in different research areas related to social network analysis like subjective language detection, (time-evolving) opinion mining, political forecasting... The development procedure itself, can be helpful for the design of additional resources when ambiguity and subjectivity is present along with the influence of cultural norms. Furthermore, since we have developed a very valuable set of tools to extract and process the information from the Internet along with the the crowdsourcing platform, it might get bigger and bigger, and also include a great variety of topics and different kind of annotations for different purposes.

9 Compliance with Ethical Standards

Conflict of Interest Raquel Justo, José M. Alcaide, M. Inés Torres and Marilyn Walker declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

10 Funding

This study was funded by the Spanish Government (TIN2014-54288-C4-4-R) and by the National Science Foundation of USA (NSF CISE R1 #1202668)

References

1. Baranyi P, Csapó A. Definition and Synergies of Cognitive Infocommunications. Acta Pytechnica Hungarica. 2012;9(1):67–83.
2. Croft W, Cruse DA. Cognitive Linguistics. Cambridge Textbooks in Linguistics. Cambridge University Press; 2004.
3. Becker-Asano C, Wachsmuth I. Affective computing with primary and secondary emotions in a virtual human. Autonomous Agents and Multi-Agent Systems. 2010;20(1):32–49.
4. Esposito A. The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. Cognitive Computation. 2009;1:268–278.
5. Recupero DR, Presutti V, Consoli S, Gangemi A, Nuzzolese AG. Sentilo: Frame-Based Sentiment Analysis. Cognitive Computation. 2015;7:211–225.
6. Vogel C. Denoting Offence. Cognitive Computation. 2014;6:628–639.
7. Hawalah A. A Framework for Building an Arabic Multi-disciplinary Ontology from Multiple Resources. Cognitive Computation. 2017;.
8. Maynard D, Greenwood MA. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al., editors. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. European Language Resources Association (ELRA); 2014. p. 4238–4243.
9. Kruger J, Epley N, Parker J, Ng ZW. Egocentrism Over E-Mail: Can We Communicate as Well as We Think? Journal of Personality and Social Psychology. 2005;89(6):925–936.
10. Alcaide JM, Justo R, Torres MI. Combining Statistical and Semantic Knowledge for Sarcasm Detection in Online Dialogues. In: Paredes R, Cardoso JS, Pardo XM, editors. Pattern Recognition and Image Analysis. vol. 9117 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 662–671.
11. Khodak M, Saunshi N, Vodrahalli K. A Large Self-Annotated Corpus for Sarcasm. CoRR. 2017;arXiv:1704.05579.
12. Ruiz Gurillo L, Padilla García XA, editors. Dime cómo ironizas y te diré quién eres. Una aproximación pragmática a la ironía. vol. 45 of Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt am Main: Peter Lang Internationaler Verlag der Wissenschften; 2009.
13. Hernández-Farías I, Benedí J, Rosso P. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In: Paredes R, Cardoso JS, Pardo XM, editors. Pattern Recognition and Image Analysis. vol. 9117 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 337–344.
14. Lukin S, Walker M. 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. Atlanta, Georgia: Association for Computational Linguistics; 2013. p. 30–40.

15. Justo R, Corcoran T, Lukin SM, Walker M, Torres MI. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems. 2014;69:124–133.
16. Hernández Farías DI, Sulis E, Patti V, Ruffo G, Bosco C. ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics; 2015. p. 694–698.
17. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AYA, Gelbukh A, et al. Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. Cognitive Computation. 2016;8:757–771.
18. Swanson R, Lukin S, Eisenberg L, Corcoran T, Walker M. Getting Reliable Annotations for Sarcasm in Online Dialogues. In: Chair) NCC, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al., editors. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA); 2014. p. 4250–425–7.
19. Reyes A, Rosso P. On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowledge and Information Systems. 2014;40(3):595–614.
20. Filatova E. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In: Language Resources and Evaluation Conference, LREC2012; 2012. p. 392–398.
21. Walker MA, An P, Tree JEF, Abbott R, King J. 2012. A corpus for research on deliberation and debate. In: In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012; 2012. p. 23–25.
22. Martí JC, Casanova I. La traducció cultural: el concepte d'ironia en francés, anglés, espanyol i catalá. In: Martos JL, editor. La traducció del discurs. Universitat d'Alacant; 2009. p. 120–152.
23. Wang PYA. # Irony or# SarcasmA Quantitative and Qualitative Study Based on Twitter. In: Proceedings of the PACLIC: the 27th Pacific Asia Conference on Language, Information, and Computation. Taipei, Taiwan; 2013. p. 349–356.
24. Alvarado Ortega MB. Los indicadores lingüísticos de la ironía en corpus escritos. Interlingüística. 2009;(18):91–97.
25. Riloff E, Qadir A, Surve P, De Silva L, Gilbert N, Huang R. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics; 2013. p. 704–714.

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

27. Ding X, Liu B, Yu PS. A Holistic Lexicon-based Approach to Opinion Mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. WSDM '08. New York, NY, USA: ACM; 2008. p. 231–240.

28. Thelwall M, Buckley K, Paltoglou G. Sentiment Strength Detection for the Social Web. J Am Soc Inf Sci Technol. 2012 Jan;63(1):163–173. Available from: http://dx.doi.org/10.1002/asi.21662.

29. Turney PD. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002. p. 417–424.

30. Valitutti R. WordNet-Affect: an Affective Extension of WordNet. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation; 2004. p. 1083–1086.

31. Cambria E, Olsher D, Rajagopal D. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In: Brodley CE, Stone P, editors. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada. AAAI Press; 2014. p. 1515–1521.

32. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, et al., editors. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA); 2010. p. 2200–2204.

33. de Albornoz JC, Plaza L, Gervs P. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In: Chair) NCC, Choukri K, Declerck T, Doan MU, Maegaard B, Mariani J, et al., editors. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA); 2012. p. 3562–3567.

34. Atserias J, Villarejo L, Rigau G. Spanish WordNet 1.6: Porting the Spanish Wordnet Across Princeton Versions. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. European Language Resources Association; 2004. p. 161–164.

35. Vossen P, editor. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Norwell, MA, USA: Kluwer Academic Publishers; 1998.

36. Díaz-Granjel I, Sidorov G, Suárez-Guerra S. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomázein. 2014;29:31–46.

37. Gómez-Adorno H, Markov I, Sidorov G, Posadas-Durán JP, Arias CF. Compilación de un lexicón de redes sociales para la identificación de perfiles de autor. Research in Computing Science. 2016;115:19–27.
38. Gómez-Adorno H, Markov I, Sidorov G, Posadas-Durán JP, Sanchez-Perez MA, Chanona-Hernández L. Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts. Comp Int and Neurosc. 2016;2016:1638936:1–1638936:13.
39. Ráez AM, Díaz-Galiano MC, Ortega JMP, López LAU. Spanish knowledge base generation for polarity classification from masses. In: Carr L, Laender AHF, Lóscio BF, King I, Fontoura M, Vrandecic D, et al., editors. 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume. International World Wide Web Conferences Steering Committee / ACM; 2013. p. 571–578.
40. Montejo-Rez A, Daz-Galiano MC, Martnez-Santiago F, Urea-Lpez LA. Crowd explicit sentiment analysis. Knowledge-Based Systems. 2014;69:134 – 139.
41. Kamvar SD, Harris J. We Feel Fine and Searching the Emotional Web. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11. New York, NY, USA: ACM; 2011. p. 117–126. Available from: http://doi.acm.org/10.1145/1935826.1935854.
42. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007. Austin, TX, LIWC Net. 2007;.
43. Pang B, Lee L, Vaithyanathan S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002. p. 79–86.
44. Dave K, Lawrence S, Pennock DM. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: WWW2003; 2003. p. 519–528.
45. Mcdonald R, Hannan K, Neylon T, Wells M, Reynar J. Structured Models for Fine-to-Coarse Sentiment Analysis. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics; 2007. p. 432–439.
46. Tsur O, Davidov D, Rappoport A. ICWSM–A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: Proceedings of the fourth international AAAI conference on weblogs and social media; 2010. p. 162–169.
47. Taboada M, Grieve J. Analyzing appraisal automatically. In: In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications; 2004. p. 158–161.
48. F C, A TJ, amd Ortega J EF. Experiments in sentiment classification of movie reviews in Spanish. Procesamiento del lenguaje Natural (Sociedad Española para el Procesamiento del Lenguaje Natural). 2012;41.

49. Martín-Valdivia MT, Martínez-Cámara E, Perea-Ortega JM, na López LAU. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications. 2013;40(10):3934 – 3942.
50. Vilares D, Alonso MA, Gómez-Rodríguez C. A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering. 2015;21(1):139–163.
51. Vicente IS, Agerri R, Rigau G. Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages. In: Bouma G, Parmentier Y, editors. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden. The Association for Computer Linguistics; 2014. p. 88–97.
52. Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 241–249.
53. Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent Twitter Sentiment Classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 151–160.
54. Montejo-Ráez A, Martínez-Cámara E, Martín-Valdivia MT, Ureña López LA. Ranked WordNet Graph for Sentiment Polarity Classification in Twitter. Comput Speech Lang. 2014 Jan;28(1):93–107.
55. González-Ibáñez R, Muresan S, Wacholder N. Identifying sarcasm in Twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. vol. 2. Citeseer; 2011. p. 581–586.
56. Reyes A, Rosso P, Buscaldi D. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. Data Knowl Eng. 2012 Apr;74:1–12.
57. Martínez-Cámara E, García-Cumbreras MA, Martín-Valdivia MT, Ureña López LA. Detecting Polarity in Spanish Tweets. Procesamiento del Lenguaje Natural (SEPLN). 2011;47.
58. Jasso G, Meza-Ruíz IV. Character and Word Baselines Systems for Irony Detection in Spanish Short Texts. Procesamiento del Lenguaje Natural. 2016;56:41–48. Available from: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285.
59. Barbieri F, Ronzano F, Saggion H. Is this Tweet Satirical? A Computational Approach for Satire Detection in Spanish. Procesamiento del Lenguaje Natural. 2015;55:135–142. Available from: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5225.

60. Barbieri F, Ronzano F, Saggion H. Do We Criticise (and Laugh) in the Same Way? Automatic Detection of Multi-Lingual Satirical News in Twitter. In: Yang Q, Wooldridge M, editors. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. AAAI Press; 2015. p. 1215–1221. Available from: http://ijcai.org/Abstract/15/175.

61. Cumbreras MÁG, Villena-Román J, Cámara EM, Díaz-Galiano MC, Martín-Valdivia MT, López LAU. Overview of TASS 2016. In: Villena-Román J, Cumbreras MÁG, Cámara EM, Díaz-Galiano MC, Martín-Valdivia MT, López LAU, editors. Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September 13th, 2016.. vol. 1702 of CEUR Workshop Proceedings. CEUR-WS.org; 2016. p. 13–21. Available from: http://ceur-ws.org/Vol-1702/tass2016_proceedings_v24.pdf.

62. Walker M, Tree JF, Anand P, Abbott R, King J. A Corpus for Research on Deliberation and Debate. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA); 2012. p. 23–25.

63. Misra A, Walker M. Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue. In: Proceedings of the SIGDIAL 2013 Conference. Association for Computational Linguistics. Metz, France: Association for Computational Linguistics; 2013. p. 41–50.

64. Misra A, Anand P, Fox Tree JE, Walker M. Using Summarization to Discover Argument Facets in Online Idealogical Dialog. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics; 2015. p. 430–440.

65. Walker MA, Anand P, Abbott R, Grant R. Stance Classification Using Dialogic Properties of Persuasion. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 592–596.

66. De Winter JCF, Kyriakidis M, Dodou D, Happee R. Using CrowdFlower to Study the Relationship between Self-reported Violations and Traffic Accidents. In: 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, {AHFE} 2015. vol. 3; 2015. p. 2518 – 2525.

67. Justo R, Torres MI, M AJ. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. In: Proceedings of 8th Iberian Conference on Pattern Recognition and Im-

age Analysis (in press). International Association for Pattern Recognition (IAPR); 2017. .

68. Buchholz S, Latorre J, Yanagisawa K. In: Crowdsourced Assessment of Speech Synthesis. John Wiley & Sons, Ltd; 2013. p. 173–216.
69. Padró L, Stanilovsky E. FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey: ELRA; 2012. p. 2473–2479.