ATLANTIS
Journal of the Spanish Association of Anglo-American Studies
44.1 (June 2022): 186-207
e-ISSN 1989-6840
DOI: http://doi.org/10.28914/Atlantis-2022-44.1.11

Corpus Annotation and Analysis of Sarcasm on Twitter: #CatsMovie vs. #TheRiseOfSkywalker

Antonio Moreno-Ortiz and María García-Gámez
Universidad de Málaga
amo@uma.es, mggamez@uma.es

Sentiment analysis is a natural language processing task that has received increased attention in the last decade due to the vast amount of opinionated data on social media platforms such as Twitter. Although the methodologies employed have grown in number and sophistication, analysing irony and sarcasm still poses a severe problem. From the linguistic perspective, sarcasm has been studied in discourse analysis from several perspectives, but little attention has been given to specific metrics that measure its relevance. In this paper we describe the creation of a manually-annotated dataset where detailed text markers are included. This dataset is a sample from a larger corpus of tweets (n= 76,764) on two highly controversial films: *Cats* and *Star Wars: The Rise of Skywalker*. We took two different samples for each film, one before and one after their release, to compare reception and presence of sarcasm. We then used a sentiment analysis tool to measure the impact of sarcasm in polarity detection and then manually classified the mechanisms of sarcasm generation. The resulting corpus will be useful for machine learning approaches to sarcasm detection as well as discourse analysis studies on irony and sarcasm.

Keywords: sarcasm detection; sentiment analysis; corpus annotation; social networks

. . .

Anotación de corpus y análisis del sarcasmo en Twitter: #CatsMovie vs. #TheRiseOfSkywalker

El análisis de sentimiento es una de las aplicaciones del procesamiento del lenguaje natural que más atención ha recibido en la última década, principalmente debido a la cantidad de

opiniones vertidas en redes sociales como Twitter. Pese a que las metodologías empleadas son cada vez más sofisticadas, el sarcasmo sigue siendo un gran problema. Aunque el sarcasmo ha sido estudiado desde varias perspectivas en el análisis del discurso, no se ha prestado mucha atención a su presencia y relevancia real, aportando métricas concretas. En este trabajo se describe la creación de un *dataset* anotado manualmente en el que se incluyen marcadores textuales. Dicho *dataset* es la muestra de un corpus de tweets (n= 76.764) sobre dos películas controvertidas: *Cats* y *Star Wars*. *El Ascenso de Skywalker*. Tomamos dos muestras para cada película, antes y después de su estreno, para comparar su acogida. Empleamos una herramienta de análisis de sentimiento para medir el impacto del sarcasmo en la detección de la polaridad, y posteriormente identificamos y clasificamos los mecanismos de generación de sarcasmo. Este corpus puede ser de gran utilidad para la detección del sarcasmo mediante aprendizaje automático, así como para estudios de análisis del discurso sobre la expresión del sarcasmo.

Palabras clave: detección de sarcasmo; análisis de sentimiento; anotación de corpus; redes sociales

1. Introduction

Detecting the mechanisms of sarcasm is essential for successful communication and understanding: when Donald Trump, former president of the US, suggested injecting disinfectant as a useful method to fight COVID-19 (Feltman 2020), he did not produce any formal markers that could hint at the presence of a sarcastic undertone; his voice tone remained the same throughout the speech, his gestures were not unusual and he did not even resort to figures of speech such as metaphors or hyperboles. For this reason, when he later affirmed that he was actually being sarcastic with these proposals, this took the general public by surprise: "Trump was not being 'sarcastic' on Thursday when he raised the possibility of injecting disinfectant. There was simply no indication that he was being anything less than serious" (Dale 2020). This example shows that sarcasm must provide a trail of markers, so that the receiver of the message can understand it as such.

Broadly speaking, *irony* is defined as a form of figurative language through which the literal meaning of a sentence is substituted by the opposite, and *sarcasm* is considered a subtype of verbal irony that conveys a negative attitude in a more aggressive and bitter way than irony, and which aims to ridicule something or someone (Sperber and Wilson 1981; Kreuz and Roberts 1993; Attardo 2000; Wilson 2013). It must, therefore, be emphasized that although these rhetorical devices are very similar, they are not the same. Studies, nonetheless, seem to have struggled to establish a clearer distinction between these terms, since the differences often depend on factors such as tone or usage, which are sometimes difficult to identify (Reyes and Rosso 2011). As a result, ironic expressions are often misunderstood as sarcastic and vice versa. For this reason, the two terms have been used interchangeably in the literature. In fact, as reported by Van Hee (2017), most computational approaches to the subject do not distinguish between these two concepts either. According to the definitions above, however, sarcasm is more relevant for our present research, as the compiled corpus is mainly based on tweets that criticize *Cats* (Hooper 2019) and *The Rise of Skywalker* (Abrams 2019).

Sentiment Analysis (SA) is a field of study that attempts to automatically process people's opinions and sentiments towards certain entities—e.g., products, services, individuals—and their attributes (Liu 2011, 459). Its basic tasks are polarity detection and emotion recognition, which are done through the use of a classification task. As a Natural Language Processing (NLP) task, it thus attempts to automatically classify the semantic orientation of a sentence or, more often, a document, a practical application that can be of great value for decision making in organizations, institutions or companies, particularly given the ever-increasing amount of opinionated data that is generated on the Web and social networks (Cambria et al. 2017a). One such social network where a large amount of evaluative language can be found is Twitter.

The presence of sarcasm in opinionated texts has long been known to pose a serious challenge to accurate sentiment analysis (Barbieri et al. 2014), hence the increasing attention that the NLP research community has given to the topic. Like so many other NLP tasks, approaches to sarcasm detection fall into one of two groups: a) rule-based or b) machine learning (ML), the latter including deep learning. Rule-based approaches

aim to identify sarcasm through specific evidence, while ML approaches use sets of features such as bag-of-words, to train prediction models. Deep learning approaches make use of neural networks, which have gained popularity in NLP (Amir et al. 2016; Gosh and Veale 2016; Joshi et al. 2016).

Sarcasm detection is highly dependent on the availability of high-quality annotated datasets, but most existing ones have been created in a semi-automatic manner, for example using hashtag-based labelling like #sarcasm. As a consequence, their quality is often questionable and, as reported by Joshi et al. (2017), this is especially true of the analysis of the hashtag #not as a form of indicating sarcasm. But even the value of manually annotated datasets might be deemed dubious, as sarcasm is a subjective phenomenon that might not be perceived equally by everyone. It is for this reason that the annotation schema must be specifically designed, and the annotation process has to be carefully controlled with precise instructions as well as clear guidelines and traceable results, thus requiring the input of highly trained language specialists.

The present work, therefore, aims to design an effective annotation schema and to create an annotated corpus for automatic sarcasm detection that includes detailed text markers of sarcasm mechanisms and their formal markers in written text. Thus, our work is relevant from two perspectives. First, it will further linguistic inquiry concerning the typology and frequency of sarcasm-generating language devices; second, the resulting annotated value can be used by ML algorithms, perhaps in combination with unsupervised ML and deep learning, in sarcasm detection, sentiment analysis, emotion detection and classification and other complex NLP tasks.

2. Theoretical Background

Sarcasm invariably has a negative implied sentiment, but it may also carry a negative surface sentiment, positive surface sentiment or no surface sentiment at all (Joshi et al. 2017). Bouazizi and Ohtsuki (2015) consider that sarcasm can be classified into three categories depending on the purpose of the message: a) sarcasm as wit, used to be funny; b) sarcasm as whimper, used to express annoyance or opposition; and c) sarcasm as avoidance, used when the speaker wants to avoid providing a direct answer. Furthermore, they define four sets of features that characterize it, these being a) sentiment-related, b) punctuation-related, c) lexical and syntactic or d) pattern-related. In the case of the first set of features, they pay attention to any sort of inconsistency that may appear between the sentiment conveyed by words and other components within a message, for example the concurrent presence of opposite pairs such as "happy" and "sad." Camp (2012) adds that sarcasm can be classified according to its linguistic context, so that it can be propositional, embedded, *like*-prefixed and illocutionary.

Sarcasm has also been deemed easier to detect in prosodic aspects, such as nasality, a lower pitch level, slower tempo and greater intensity (Attardo 2000; Rockwell 2000). Campbell and Katz (2012) state that one of the main characteristics of sarcasm is that it occurs along several dimensions, i.e., failed expectation, negative tension, pragmatic

insincerity and the presence of a victim. Eisterhold et al. (2006), nonetheless, affirm that sarcasm is commonly followed by the responses that it can provoke: laughter, smiling, a change of topic, a literal reply, a non-verbal reaction, sarcasm in retort or no response at all.

Previous social media studies that have considered the presence of sarcasm in text have relied on the use of corpora based on tweets that included hashtags such as #sarcasm or #notreally, employing various methodologies (Davidov et al. 2010; González-Ibáñez et al. 2011; Liebrecht et al. 2013; Reyes et al. 2013). Likewise, social media platforms other than Twitter have also been explored, especially comments on Reddit (Wallace et al. 2014). Supplementary annotation, which goes beyond the annotation of the absence or presence of sarcasm, has also been carried out (Mishra et al. 2016) and Amazon reviews have also been the object of many studies attempting to identify sarcasm automatically (Tsur et al. 2010; Buschmeier et al. 2014; Liu et al. 2014).

Nevertheless, the problem remains the same: high-quality annotated datasets are still scarce, and their unavailability is becoming a significant hindrance to achieving accurate sentiment analysis systems and automatic sarcasm detection, thus our practical motivation to create this annotated corpus.

3. Research Design

3.1. Objectives

The general objective of this work is to create an annotated corpus for automatic sarcasm detection. This involves a number of epistemic and operational prerequisites which determine the specific objectives, as detailed below:

- Specific objective one: to design an effective annotation schema for sarcasm detection and typology.
- Specific objective two: to annotate an opinion corpus of tweets from the perspective of sarcasm.
- Specific objective three: to explore sarcasm mechanisms and their formal markers in written text.
- Specific objective four: to check the real impact that sarcasm has on sentiment analysis systems.

3.2. Sample and Annotation Process

The dataset used in our research is a sample from a larger corpus of tweets (n= 76,764) on two highly controversial films released at roughly the same time: *Cats*[1] and *Star*

[1] *Cats* is a musical fantasy film directed by Tom Hooper based on the stage musical of the same name by Andrew Lloyd Webber, which was in turn based on the 1939 poetry collection *Old Possum's Book of Practical Cats* written by T. S. Eliot. Critic reviews of the film were devastating, and the film was named one of the worst films ever made (Rotten Tomatoes 2021). The film had a budget of nearly $100 million but only grossed $75 million.

Wars: The Rise of Skywalker.[2] The tweets cover two distinct time spans: before and after the premiere of these films, so tweets produced between December 11, 2019 and December 19, 2019 were downloaded to represent the first time frame, and tweets created between December 30, 2019 and January 8, 2020 were chosen for the second one. This was done to compare audience expectations prior to the films' premieres with their responses and reactions after they had watched them. The total number of tweets and words in each sample is shown in table 1.

TABLE 1. Sample size by film and time frame

Film	Time frame	Tweets	Words
Cats	Pre-release	17,578	266,043
	Post-release	8,701	157,709
The Rise of Skywalker	Pre-release	38,909	822,517
	Post-release	11,576	282,661
Total		76,764	1,528,930

The corpus was then pre-processed to remove hyperlinks, line breaks and other characters. In addition, the tweets were randomized to prevent accumulation on the same date, and the first 500 from each time period and film were subsequently selected for annotation. As a result, 2,000 tweets were manually annotated in total.

The annotation was carried out using Prodigy (Montani et al. 2021), a tool specifically developed for the annotation of corpora to be used in ML tasks. The annotation schema for sarcasm identification was performed in two steps. In the first, an initial sample of 400 tweets was annotated—100 instances from each time frame and film. Three coders, trained in linguistic annotation, were asked to identify the presence/absence of sarcasm in each tweet and keep a record of problematic cases. Afterwards, a feedback session was carried out to compare results, check problematic cases and unify criteria for sarcasm identification. Finally, a second round of 1,600 tweets (400 tweets from both pre-and post-release dates of each film) was annotated, providing a total of 2,000 sarcasm annotated tweets.

Most problematic cases reported by the coders were related to the difficulty in discriminating sarcasm from other rhetorical devices or figures of speech, as in the following examples, which include wordplays (1), hyperbole (2) and humour (3):

[2] *Star Wars: The Rise of Skywalker* (also known as *Episode IX*) is an epic space opera directed by J. J. Abrams from a script by Abrams and Chris Terrio. The film received mixed reviews from critics (Rotten Tomatoes reports an average score of 6.1/10), it is one of the most expensive films ever made, and grossed over $1 billion worldwide, being one of the fifty highest-grossing films of all times.

(1) Please enjoy this sneak peak of my trip to Cat School! □ Careful not to take myself too seriously... Gotta let go trying to be purrrrrrrrfect.@catsmovie

(2) If you buy tickets to see @catsmovie in theatres, our relationship is over!

(3) A long time ago, in a podcast far, far away.... #StarWars #TheRiseOfSkywalker #podcast

In the case of such uncertain examples, it was decided to follow the strict definition of sarcasm as being negative in nature, even if not overtly negative. Therefore, the message had to contain some sort of harsh criticism and, following Joshi et al. (2017), "an implied negative sentiment because it intends to express contempt" (2) for it to be classified as sarcastic: for example, although wordplay and hyperbole can be markers of sarcasm, in those uncertain cases they had to portray a harsh criticism for them to be considered sarcastic instances.

3.3. Inter-Annotator Agreement

Corpus annotation by multiple human annotators needs to be checked for consistency and reliability (Artstein and Poesio 2008). This is true of all types of annotation, but especially of corpora annotated for subjective features, such as ours. One annotator may deviate, sometimes considerably, from the rest due to a different perspective or understanding of the annotation guidelines. An inconsistently annotated corpus ultimately defeats its purpose, since the ML algorithms will be unable to extract useful patterns to be used in the prediction process. Checking the reliability of the annotation is thus absolutely crucial and must be performed using very specific metrics.

In order to validate our annotation procedure, we used the stats module Python included in the SciPy package to calculate several metrics of Inter-Annotator Agreement among the three coders. Specifically we obtained Krippendorf's *alpha* (α) coefficient (Krippendorff 2004) and Fleiss's *kappa* (κ) coefficient (Fleiss 1981). These metrics measure the reliability of the agreement between the annotators, as compared to that which would be expected by mere chance, and the index of both ranges from 0—chance agreement—to 1—total agreement.

Reliability is a property of data achieved by repeating the same tests under different conditions—in our case, by using different coders—and is measured by the level of agreement or disagreement across those repetitions (Krippendorff 2004). It guarantees that a research test can be replicated and obtain similar results, and it is an essential metric to ensure the validity of an annotation schema and, by extension, of an annotation procedure (Moreno-Ortiz et al. 2019). Both Krippendorf's *alpha* and Fleiss's *kappa* can be used to measure agreement among more than two annotators. In our dataset both returned the same results, shown in table 2.

Table 2. Inter-annotator agreement metrics

Metric	Result	Interpretation
Krippendorf's *alpha* (α) coefficient	0.71	substantial agreement
Fleiss's *kappa* (κ) coefficient	0.71	substantial agreement

To check for possible differences between pairs of annotators, we also calculated pair-wise agreement using Cohen's *kappa* statistic (Cohen 1960) and other similar agreement coefficients, which showed that there was a higher agreement between annotators A-B than between annotators A-C and B-C, even though all three results were strongly positive. The results of these metrics are summarized in table 3 below:

Table 3. Pair-wise inter-annotator agreement metrics

	A-B	B-C	A-C
Cohen's kappa	0.87	0.65	0.61
Pearson's correlation	r(499)=0.81, p<0.001	r(499)=0.65, p<0.001	r(499)=0.61, p<0.001
Spearman's correlation	r(499)=0.87, p<0.001	r(499)=0.65, p<0.001	r(499)=0.61, p<0.001
Kendall's correlation	r(499)=0.87, p<0.001	r(499)=0.65, p<0.001	r(499)=0.61, p<0.001
Joint probability	0.97	0.93	0.92

Our results confirmed that the annotation criteria were correctly understood, despite the complexity of the task and the inherent difficulties of problematic cases spotted by the annotators.

3.4. Annotation of Sarcasm Mechanisms

Once the tweets were classified as containing sarcasm or not, a further classification task was carried out following an essentially data-driven protocol. We initially proposed a very basic set of sarcasm mechanisms that are frequently used in the literature and instructed the coders to identify sarcasm patterns, which mainly referred to lexical-grammatical, typographic or formal patterns. Additionally, we included a series of semantic and contextual mechanisms also used to create a sarcastic effect, such as wordplay and hyperbole. Finally, we added more specific categories as required when none of the initial ones were appropriate to describe the examples found in the corpus. This is the case for different types of phrasal irony—which was included following the definition put forward by Partington (2011) as "the reversal of customary collocational patterns of use of certain

lexical items" (1786)—as well as those instances identified as *embedded incongruity*, a more subtle type of sarcasm in which there is a mismatch between one element and the rest of the sentence (Camp 2012) and situational sarcasm, through which Twitter users tried to call attention to an incongruence between two situations: what is expected to happen, and what actually happens (Kreuz and Roberts 1993, 99; Van Hee 2017, 10). Incongruity, in addition, is a term that has also been widely employed in humour detection, and what distinguishes its sarcastic tone from its humoristic one is, according to Stieger et al. (2011), the element of mockery and aggressiveness that characterizes the former. In section 4 we will describe and exemplify some of these patterns and mechanisms, which are either semantic or contextual in nature, or more easily tractable formal patterns. Table 4 below summarizes the sarcasm patterns in the corpus.

Table 4. Sarcasm patterns in the corpus

Semantic & contextual patterns	Formal patterns
Wordplay	Comparison
Hyperbole	Rhetorical question
Intertextual sarcasm	Typographic effect
Phrasal irony	Sarcastic locution
Semantic clash	Situational sarcasm
Embedded incongruity	Other

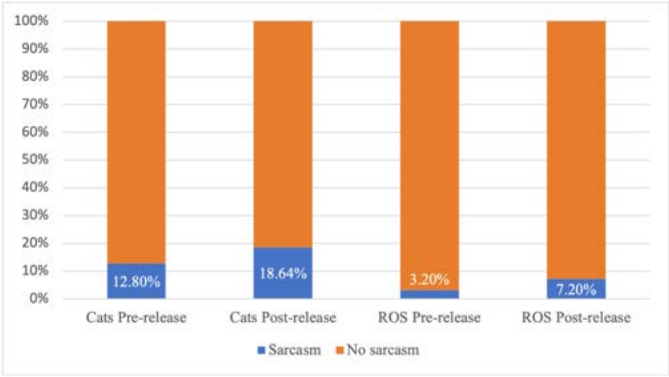
4. Analysis of Results

In this section we discuss, from a quantitative and qualitative point of view, the presence of sarcasm in the corpus, as well as the sarcasm mechanisms employed, the main themes found and the predominating polarity switches.

4.1. Presence of Sarcasm

Regarding the results of our first classification task, the total number of sarcastic tweets in the complete dataset was 211 (10.55%). In the case of the film *Cats*, sarcasm is more prominent in the dataset of tweets produced in the post-release period, with over 18% of sarcastic instances, whereas 12.8% of the tweets were identified as sarcastic in the pre-release period. This may be because people preferred commenting more critically on the film once they had watched it; the post-release dataset, therefore, contained more evaluative language and more of that evaluation was sarcastic in nature. These results are summarized in figure 1.

FIGURE 1. Percentages of sarcastic tweets found in tweets (n= 2,000)



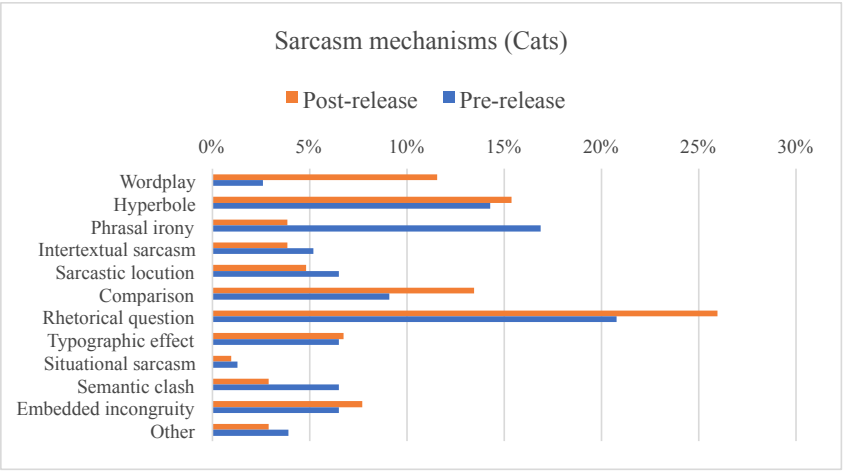
Two observations are immediately apparent from these data. First, the number of sarcastic comments on *The Rise of Skywalker* (ROS) is significantly lower than in the case of *Cats*, comprising 7.2% of the sample in post-release tweets and only 3.2% in pre-release tweets, making an average of 5.2% of sarcastic tweets, versus 15.72% for *Cats*. Second, in both cases the proportion of sarcasm is higher in the post-release tweets. The first observation suggests that the presence of sarcasm seems to be highly dependent on the topic and the user community. In this particular case, it may be due to the fact that the *Star Wars* saga has a larger fanbase that has expanded over the years and across generations, with followers who were counting the days until the release of the film and were more inclined to criticize the film directly and harshly if it did not meet their expectations. *Cats*, on the other hand, does not have any sequels or prequels and its target audience was much more varied, mainly composed of families: in fact, the premiere was scheduled near the Christmas holiday and the film was advertised on the Universal Pictures official website as "the most joyful event of the holiday season." However, its lack of a pre-established fanbase audience may be why users more generally resorted to sarcastic devices to talk about the movie when they felt disappointed or did not like it.

These results also have implications for the polarity of the tweets, a concept that correlates with the presence or absence of sarcasm. However, as we will describe in detail in section 4.4., the impact of sarcasm on polarity is not as great as might be expected, as many of the sarcastic tweets contain other lexical markers that determine the negative polarity.

4.2. Sarcasm Mechanisms

As for the sarcasm mechanisms and patterns used in the four different datasets, the most frequent ones found in *Cats* are rhetorical questions, hyperbole and phrasal irony. Figure 2 shows the percentages of all categories.

FIGURE 2. Sarcasm mechanisms in *Cats*



Here are some examples that illustrate the sarcasm mechanisms we classified: a) rhetorical question as sarcasm mechanism (example 4), b) hyperbole as sarcasm mechanism (5) and c) phrasal irony as sarcasm mechanism (as in example 6).

(4) Remember when @UniversalPics thought that #CatsMovie was Oscar worthy?

(5) The cat dragged in James Corden. Redemption is nigh. He is dancing in the crosswalk.... WORLDS ARE COLLIDING! Brought to you by @catsmovie

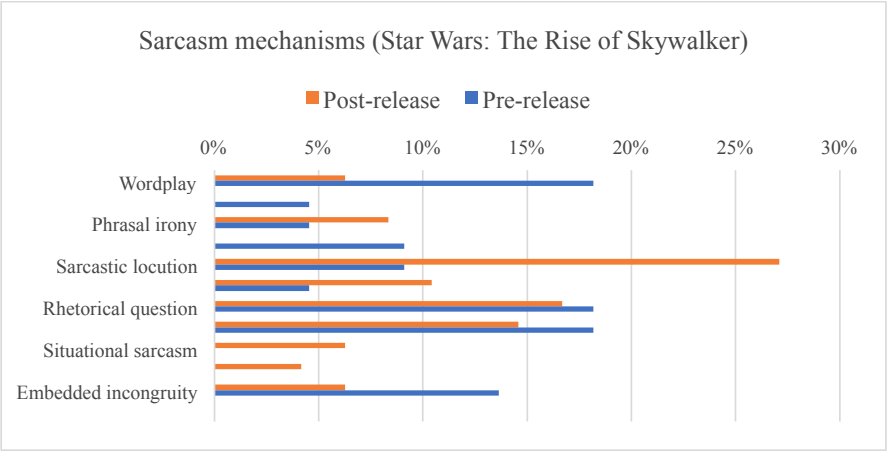
(6) Before entering the theater there were children, eagerly anticipating Cats. After witnessing this cinematic masterpiece I can only assume those children are no longer with us and have been sent to the Heaviside Layer. To be honest I'm still not sure.

In example (4) the user resorts to a rhetorical question, an utterance that has the structure of a question, but which does not expect an answer (Rohde 2006), in order to ridicule the film: they call attention to the fact that Universal Pictures might have submitted *Cats* as worthy of an Oscar nomination and the implication is that the film was not Oscar worthy at all. Example (5) represents the usage of exaggerated or hyperbolic concepts as a form of sarcasm, as the speaker suggests that the Earth might collide because of the scenes included in the film. In addition, the combination of these hyperbolic elements with formal markers, such as the use of upper case and exclamation marks, implies an intensification that increases the sarcastic effect of the tweet, as pointed out by Kunneman et al. (2015). In turn, example (6) illustrates the relevance of phrasal irony, where the semantic clash appears in the collocational pattern or collocational prosody of a word or

phrase (Partington 2011); in this case, the use of the verb "witness," which may have a negative semantic prosody,[3] and is frequently associated with crimes, violence or murder, with the object "cinematic masterpiece" achieves the sarcastic effect of a humorous, but harsh, critique, reinforced by the use of elevated language.

As for the sarcasm mechanisms in *#TheRiseOfSkywalker*, rhetorical questions and wordplay were also found, together with sarcastic locutions—by far the most frequent in this set, especially in post-release tweets—and typographic effects, as shown in figure 3.

FIGURE 3. Sarcasm mechanisms in *The Rise of Skywalker*



The examples below provide instances of the most relevant sarcasm mechanisms, including sarcasm locutions (example 7), a concatenation of rhetorical questions (as in example 8) or in the case of typographic effects, the use of negative locutions in upper case in combination with emojis (example 9).

(7) Seriously, honey? You could've made him live again like you did with Palps, or like they did with Darth Maul, etc. There were SO many things you could've done differently. You also could've found better excuses, yet here you are, being a 12 years (*sic*) old. #TheRiseOfSkywalker

(8) Why did Kylo reforge his mask and why are there red bits? Is it really hard to just make more of them? Is the technology to properly repair helmets without cosmetic damage just not there yet? Did he just think a red inlay would look cool? #TheRiseOfSkywalker

[3] A search in the English Web 2015 corpus provided by *Sketch Engine* shows that the words "violence," "murder," "incident" and "decline" are among the most frequent collocates of the verb *witness*.

(9) There's NO WAY this trilogy is about Rey and Kylo. Absolutely NO WAY. The marketing is NOT telling us that it's about them. □ #reyandKylo #KyloandRey #TheRiseOfSkywalker

This leads us to another interesting aspect that we found repeatedly in our dataset: very frequently, users achieved the sarcastic effect by resorting to a combination of semantic and pattern-based mechanisms. We see combinations of rhetorical questions and typographic effects in example (10), situational sarcasm plus sarcastic locutions in example (11) and sarcastic locutions such as "it's not been discussed enough" together with the typographic emphasis of capital letters in example (12). Finally, example (13) employs a combination of upper case, reduplication and two types of phrasal irony. In this case, the phrasal irony comes from the set phrase "it was everything I wanted it to be and…" in which the final item that normally appears at the end has been reversed from the original "more" to the negative "less."

(10) @sten1225 @catsmovie You're telling me that you DON'T want to see Ian McKellen as a humanoid cat!?

(11) Rian Johnson: *kills off Snoke* JJ Abrams: Okay, now we need a new villain! Hux: Am I a joke to you? #TheRiseOfSkywalker #hux #HuxDeservesBetter

(12) Dame Judi Dench LIFTING HER LEG in praise of Ian McKellen, shortly after he licked a bowl of water, is not a (sic) being discussed enough #CatsMovie

(13) Afterwards, we could not stop saying, "BUT THEIR FEET AND HANDS... Whhhhhyyyy?!" It was everything I wanted it to be and less. 5 stars for accidental entertainment! □ #CatsMovie #CatsReview 7/8

It is worth noting that in example (13) we also find what Partington (2011, 1789) calls *evaluative oxymoron*, "a type of oxymoron where the two constituent elements are of opposing evaluative polarity." As he explains, the overall evaluation is usually unfavourable, which is what we find in this example: although "5 stars" and "entertainment" are positive, the negativity of "accidental," together with the cumulative effect of the other mechanisms, transforms the tweet into a sarcastic message.

4.3. Sarcasm Themes

Another interesting aspect of our analysis is that, in relation to their content, many of the sarcastic tweets revolved around similar themes, which seems to indicate that the way sarcasm is constructed, at least on Twitter, is somehow dependent on the shared knowledge of the community of users and the topic they talk about. In the case of *Cats*, most sarcastic

tweets were built on three recurring themes: a very high percentage compared *Cats* with a terror movie or talked about it as if it were a horror film, whereas others achieved their sarcastic effect mentioning the possibility or obligation of taking drugs to be able to watch the movie. A third sarcastic theme in the *Cats* dataset was related to *hate-watching*—in the sense that the film was so bad that it could be considered quite good—and compares this film to others that have become famous over the years because of this hate-watch phenomenon.[4] Table 5 contains a breakdown of the sarcasm themes in *Cats*.

TABLE 5. Sarcasm themes in *Cats*

Themes	%
Cats as a terror movie	73.33%
Taking drugs or alcohol to endure watching it	22.22%
Hate-watching	4.45%

These three themes are recurrent in sarcastic tweets in the case of *Cats* and they seem to be somehow thematically developed by the user community, which of course makes those tweets easier to identify as sarcastic.

In the case of *The Rise of Skywalker*, nevertheless, sarcasm themes are not readily identifiable, and the sarcastic content of the dataset was more frequently related to technical or narrative aspects of the film or even of the complete saga. A small percentage of tweets (8.33%) included the hashtag #BenSoloDeservesBetter, which represented the viewers' diverging opinions about the film's plot. It is not easy to derive conclusions as to how the content of sarcastic tweets is constructed in the two films, but from a sociological perspective, the community of users and the target audience of these two films are by no means comparable. As we already mentioned in section 4.1. *Star Wars* is a kind of pop-culture phenomenon,[5] with a saga of 9 episodes that started back in 1977 and fans all over the world. For *Star Wars* fans, using social media is yet another way to satisfy their "need to belong" to the community (Pallavicini et al. 2017), together with using other specialized media such as blogs, YouTube channels, Reddit, Instagram or other fan organizations, where they establish strong ties in their interpersonal relationships.[6]

[4] The term *hate-watching* was coined by *New Yorker* critic Emily Nussbaum. Critical and commercial failure films such as *Showgirls* (directed by Paul Verhoeven in 1995) and *The Room* (directed by Tommy Wiseau in 2003) have achieved cult status years after their disastrous initial releases and are now shown with great success at late-night screenings. *The Rocky Horror Picture Show* (directed by Jim Sharman in 1975) is the epitome of a film critically panned on initial release that has become a cultural phenomenon.

[5] In fact, the *Star Wars* franchise includes an endless variety of merchandising and memorabilia related to each of the films in the saga.

[6] Morris (2019) estimates that *Star Wars* fans make up one in four Internet users and two in three *Star Wars* fans agree the Internet makes them feel closer to people.

It is likely that such a community, when using Twitter, feels more aggrieved than amused while criticizing the film, so the content of their sarcastic tweets tends to be much more negative in nature, expressing the fans' disenchantment with the film. This also has a significant impact on the polarity distribution of our dataset, which we describe in the following section.

4.4. Sentiment Polarity and Sarcasm

Sentiment analysis is an NLP task based on the classification of messages according to their overall semantic orientation. Rhetorical devices such as sarcasm, however, can have an impact on the accuracy of the performance of sentiment analysis systems, since messages that should be understood as negative are identified by the systems as positive, and vice versa (Hernández Farias and Rosso 2017). For this reason, research workshops such as SemEval and Fig-Lang have paid increasing attention to automatic sarcasm detection, especially in contexts such as Twitter: in 2018, SemEval put forward a shared task on irony detection in tweets (Van Hee et al. 2018), while in 2020, Fig-Lang held a workshop on sarcasm detection (Ghosh et al. 2020). Shared tasks in general, and those that focus on sarcasm detection in particular, require high-quality annotated datasets, such as the one we generated in this research study.

On the other hand, although irony and sarcasm are frequently mentioned as hurdles to sentiment analysis, few studies provide quantitative evidence of the extent to which this is actually true, that is, what proportion of evaluative texts contain such rhetorical devices and how often they pose a problem for the successful classification of the sentiment expressed in them. Thus, we set out to measure the number of cases in which sarcasm did pose a problem for the calculation of polarity in tweets using a lexicon-based sentiment analysis system, Lingmotif 2 (Moreno-Ortiz 2021). Lingmotif determines the semantic orientation of a text through the identification of sentiment-laden linguistic expressions. In addition, it offers quantitative data and graphic visualizations of the sentiment and other content metrics, such as topics and entities. Its lexical resources constitute the core of the sentiment engine: *Lingmotif-lex* (Moreno-Ortiz and Pérez-Hernández 2018), a manually-curated English sentiment lexicon of wide coverage. This core lexicon contains over 28,000 single-word forms and more than 38,000 multi-word expressions. Furthermore, the system includes a set of sentiment shifters that account for context-dependent valence modifications.

Our method consisted in using Lingmotif to automatically classify, from the sentiment perspective, the tweets that we had previously manually identified as sarcastic, the underlying assumption being that, since sarcasm implies negativity, semantic orientation should be negative. Table 6 summarizes the results.

TABLE 6. Sentiment classification of sarcastic tweets

		Tweets	Presence of sarcasm	Classified as POS	Classified as NEU	Classified as NEG	Classif. error
Cats	Count	1000	159	52	40	67	92
	Percent		15.9%	5.2%	4%	6.7%	9.2%
ROS	Count	1000	52	17	10	25	27
	Percent		5.2%	1.7%	1%	2.5%	2.7%
Total	Count	2000	211	69	50	92	119
	Percent		10.55%	3.45%	2.5%	4.6%	5.95%

These data suggest that an advanced lexicon-based sentiment classification system would fail to classify correctly about 6% of tweets due to the presence of sarcasm. From an NLP perspective, this is probably the most stand-out finding of our research because, to our knowledge, no other formal study has produced a specific figure that attempts to measure the impact of sarcasm on lexicon-based sentiment analysis. Obviously, this percentage may vary with other datasets and other systems. Due to the nature of the tweets in our corpus, which comment on highly controversial films, the proportion of sarcastic tweets and therefore the proportion of those that pose a problem for sentiment analysis is probably higher than average on Twitter, so other less "intense" Twitter datasets will likely show a lower proportion of sarcasm.

To better understand why a lexicon-based sentiment analysis system would misclassify sarcastic content, we now provide examples from both movies where the identified sentiment words are shown in bold in table 7—no sentiment items were identified in neutral tweets, hence this classification.

TABLE 7. Examples of misclassified tweets

	Cats
Neutral	
(14)	This is what I imagine when I run up the stairs after turning off the lights
(15)	@catsmovie @UniversalPics Nobody Literally nobody is counting for this movie No (*sic*) even cats
Positive	
(16	Most anticipated **horror** film of 2019
(17)	This is the **greatest** campaign for the **most terrifying horror** movie of the last century.
	The Rise of Skywalker
Neutral	
(18	What is this, a crossover episode?!? #TheRiseOfSkywalker
(19)	I wanna read @rianjohnson's treatment of #TheRiseOfSkywalker just so I can pretend I have any emotional catharsis from what we got

Positive	
(20)	Don't even try to understand .. #TheRiseOfSkywalker scenario is unintentionally **hilarious**. I wonder if @jjabrams remembered he directed #TheForceAwakens.
(21)	As a movie, #TheRiseOfSkywalker was **confusing**. BUT ... As a 2 1/2 hour compilation of video-game cut scenes with StarWars quotes, locations, and plot-points that we all know and **love**?! **FANTASTIC**!!!

As for the classification results of all tweets in the dataset, both sarcastic and non-sarcastic, the results show that the tweets that refer to *The Rise of Skywalker* contain more polarized messages than those of *Cats*: neutral messages do not predominate, while positive and negative tweets constitute 25% and 58.33%, respectively. In this sense, the fact that there are more negative tweets in *The Rise of Skywalker* than in *Cats* is in accordance with our results for the most significant sarcasm themes found in the corpus. This illustrates that sarcasm is not always associated with a surface positive statement that has to be understood as negative, but rather on many occasions, sarcasm is simply identified as directly negative and sarcastic.

5. Conclusions and Further Research

The schema design and annotation tasks described in sections 3.2. and 3.3. complete specific objectives one and two.[7] We believe that this corpus will be useful to train ML prediction models, which can contribute to the problem of automatic sarcasm detection, but it can also be a very useful resource for discourse analysis.

We carried out the analysis of sarcasm mechanisms (specific objective three) by comparing the two films, since it was apparent that they presented distinct features. The results thus suggest that the general impact of sarcasm seems to be highly dependent on the topic and the user community. In terms of the mechanisms employed for the expression of sarcasm, those found in *Cats* were mostly based on semantic patterns, although other devices were also identified, such as rhetorical questions, hyperbole and phrasal irony, often in combination with semantic incongruity. *The Rise of Skywalker*, on the other hand, presented devices that generally relied on formal patterns, such as sarcastic locutions, rhetorical questions and typographic effects. Additionally, very specific sarcasm themes were identified for the film *Cats*, but not for *The Rise of Skywalker*. These themes spark off spontaneously on the social networks, are often embodied and condensed into memes and then are fuelled by the community of users.

Finally, specific objective four, which aimed to measure the impact of sarcasm on sentiment analysis, also produced considerable differences between the two datasets, with *Cats* having a significantly higher proportion of sarcastic tweets. On average we found that over 10% of the tweets in the corpus contained some type of sarcasm, but

[7] The annotated corpus is available in JSON format on https://github.com/amouma/SarCats

only 6% actually posed a problem for lexicon-based sentiment analysis since the rest were in fact classified as negative.

Although the sample of the annotated corpus we generated—2,000 tweets—is enough for the scope of this research, a more sizeable, and especially more varied, sample should be annotated for this last finding to be generalizable to Twitter as a whole, since, in light of our own findings, the discussion topic clearly determines the proportion of sarcasm. This is an avenue of research that we intend to tackle in the future.[8]

Works Cited

Abrams, Jeffrey Jacob. 2019. *Star Wars: The Rise of Skywalker*. Walt Disney Studios Motion Pictures.

Amir, Silvio et al. 2016. "Modelling Context with User Embeddings for Sarcasm Detection in Social Media." In Riezler and Goldberg 2016, 167-77.

Apidianaki, Marianna et al., eds. 2018. *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans: Association for Computational Linguistics.

Armstrong, Jennifer Keishin. 2017. "The Joy of Hate-Watching." *BBC*. June 26. [Accessed July 20, 2021].

Artstein, Ron and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555-96.

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

Balahur, Alexandra et al., eds. 2011. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Portland: Association for Computational Linguistics.

—, eds. 2014. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Baltimore: Associacion for Computational Linguistics.

—, eds. 2016. Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. San Diego: Associacion for Computational Linguistics.

Balahur, Alexandra, Erik van der Goot and Andres Montoyo, eds. 2013. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity and Social Media Analysis*. Atlanta: Association for Computational Linguistics.

Barbieri, Francesco, Horacio Saggion and Francesco Ronzano. 2014. "Modelling Sarcasm in Twitter, a Novel Approach." In Balahur et al. 2014, 50-58.

[8] The research reported in this article was funded by the *Consejería de Economía, Conocimiento, Empresas y Universidad de la Junta de Andalucía* through the research projects *SentiTur: Sistema de monitorización de opinión de usuarios de recursos turísticos andaluces basado en análisis de sentimiento y análisis visual* (UMA18-FEDERJA-158) and *EAVITur: Extracción, análisis y visualización de inteligencia turística. Ecosistema innovador con inteligencia artificial para Andalucía 2025* (CEI A-Tech).

Bouazizi, Mondher and Tomoaki Ohtsuki. 2015. "Sarcasm Detection in Twitter: 'All Your Products Are Incredibly Amazing!!!' - Are They Really?" In Tiedemann 2015, 1-6.

Buschmeier, Konstantin, Philipp Cimiano and Roman Klinger. 2014. "An Impact Analysis of Features in a Classification Approach to Irony Detection in Product Reviews." In Balahur et al. 2014, 42-49.

Calzolari, Nicoletta et al., eds. 2018. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Language Resources Association.

Cambria, Erik et al., 2017a. "Affective Computing and Sentiment Analysis." In Cambria et al. 2017b, 1-10.

—, eds. 2017b. *A Practical Guide to Sentiment Analysis*. Berlin: Springer.

Camp, Elisabeth. 2012. "Sarcasm, Pretense and the Semantics/Pragmatics Distinction." *Nous* 46 (4): 587-634.

Campbell, John D. and Albert N. Katz. 2012. "Are There Necessary Conditions for Inducing a Sense of Sarcastic Irony?" *Discourse Processes* 49 (6): 459-80.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* XX (1): 37-46.

Cohen, William et al., eds. 2010. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. Menlo Park: The AAAI Press.

Cole, Peter, ed. 1981. *Radical Pragmatics*. New York: Academic Press.

Cui, Bin et al., eds. 2016. *Web-Age Information Management*. Berlin: Springer.

Dale, Daniel. 2020. "Fact Check: Trump Lies That He Was Being 'Sarcastic' When He Talked about Injecting Disinfectant." *CNN*, April 24. [Accessed July 20, 2021].

Davidov, Dmitry, Oren Tsur and Ari Rappoport. 2010. "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon." In Farkas et al. 2010, 107-16.

Dekang, Lin, Yuji Matsumoto and Rada Mihalcea, ed. 2011. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Oregon: Association for Computational Linguistics.

Eisterhold, Jodi, Salvatore Attardo and Diana Boxer. 2006. "Reactions to Irony in Discourse: Evidence for the Least Disruption Principle." *Journal of Pragmatics* 38 (8): 1239-56.

Erk, Katrin and Noah A. Smith, eds. 2016. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics.

Farkas, Richárd et al., eds. 2010. *Proceedings of the 14th Conference on Computational Natural Language Learning*. Uppsala: Association for Computational Linguistics.

Feltman, Rachel. 2020. "Drinking, Bathing in or Injecting Yourself with Bleach Can Be Deadly—and It Won't Cure COVID-19." *Popular Science*, April 24. [Accessed July 20, 2021].

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. New York: John Wiley.

Ghosh, Debanjan, Avijit Vajpayee and Smaranda Muresan. 2020. "A Report on the 2020 Sarcasm Detection Shared Task." In Klebanov et al. 2020, 1-6.

González-Ibáñez, Roberto, Smaranda Muresan and Nina Wacholder. 2011. "Identifying Sarcasm in Twitter: A Closer Look." In Dekang, Matsumoto and Mihalcea 2011, 581-86.

Gosh, Aniruddha and Tony Veale. 2016. "Fracking Sarcasm Using Neural Network." In Balahur et al. 2016, 161-69.

Hernández Farias, Delia Irazú and Paolo Rosso. 2017. "Irony, Sarcasm and Sentiment Analysis." In Pozzi et al. 2017, 113-28.

Hooper, Tom. 2019. *CATS*. Universal Pictures.

Joshi, Aditya, Pushpak Bhattacharyya and Mark J. Carman. 2017. "Automatic Sarcasm Detection: A Survey." *ACM Computing Surveys* 50 (5): 1-22.

Joshi, Aditya et al. 2016. "Are Word Embedding-Based Features for Sarcasm Detection?" In Su, Duh and Carreras 2016, 1006-11.

Klebanov, Beata B. et al., eds. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association for Computational Linguistics.

Kreuz, Roger J. and Richard M. Roberts. 1993. "On Satire and Parody: The Importance of Being Ironic." *Metaphor and Symbolic Activity* 8 (2): 97-109.

Krippendorff, K. 2004. *Content Analysis: An Introduction to its Methodology*. Thousand Oaks: SAGE Publications.

Kunneman, Florian et al. 2015. "Signaling Sarcasm: From Hyperbole to Hashtag." *Information Processing & Management* 51 (4): 500-09.

Liebrecht, Christine, Florian Kunneman and Antal van den Bosch. 2013. "The Perfect Solution for Detecting Sarcasm in Tweets #not." In Balahur, van der Goot and Montoyo 2013, 29-37.

Liu, Bing. 2011. *Web Data Mining: Exploring Hyperlinks*, *Contents and Usage Data*. Berlin: Springer.

Liu, Peng et al. 2014. "Sarcasm Detection in Social Media Based on Imbalanced Classification." In Cui et al. 2016, 459-71.

Mishra, Abhijit et al. 2016. "Harnessing Cognitive Features for Sarcasm Detection." In Erk and Smith 2016, 1095-104.

Montani, Ines et al. 2021. *Prodigy v1.11.4* (version v3.1.0). Explosion.

Moreno-Ortiz, Antonio. 2021. *Lingmotif 2* (version 2.05). Python, Angular.

Moreno-Ortiz, Antonio and Chantal Pérez-Hernández. 2018. "Lingmotif-Lex: A Wide-Coverage, State-of-the-Art Lexicon for Sentiment Analysis." In Calzolari et al. 2018, 2653-59.

Moreno-Ortiz, Antonio, Soluna Salles-Bernal and Aroa Orrequia-Barea. 2019. "Design and Validation of Annotation Schemas for Aspect-Based Sentiment Analysis in the Tourism Sector." *Information Technology & Tourism* 21 (4): 535-57.

Morris, Tom. 2019. "The Fandom Menace: Profiling Star Wars' Influential Fanbase." *GWI*. December 3. [Accessed July 20, 2021].

Pallavicini, Federica, Pietro Cipresso and Fabrizia Mantovani. 2017. "Beyond Sentiment: How Social Network Analytics Can Enhance Opinion Mining and Sentiment Analysis." In Pozzi et al. 2017, 13-30.

Partington, Alan. 2011. "Phrasal Irony: Its Form, Function and Exploitation." *Journal of Pragmatics* 43 (6): 1786-800.

Pozzi, Federico et al., eds. 2017. *Sentiment Analysis in Social Networks*. Milan: Elsevier.

Reyes, Antonio and Paolo Rosso. 2011. "Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection." In Balahur et al. 2011, 118-24.

Reyes, Antonio, Paolo Rosso and Tony Veale. 2013. "A Multidimensional Approach for Detecting Irony in Twitter." *Language Resources & Evaluation* 47 (1): 239-68.

Riezler, Stefan and Yoav Goldberg, eds. 2016. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin: Association for Computational Linguistics.

Rockwell, Patricia. 2000. "Lower, Slower, Louder: Vocal Cues of Sarcasm." *Journal of Psycholinguistic Research* 29 (5): 483-95.

Rohde, Hannah. 2006. "Rhetorical Questions as Redundant Interrogatives." *San Diego Linguistics Papers* 2: 134-68.

Rotten Tomatoes. 2021. "Cats (2019)." [Accessed March 17, 2022].

Sestero, Greg and Tom Bissell. 2013. *The Disaster Artist: My Life inside the Room, the Greatest Bad Movie Ever Made*. Simon & Schuster.

Sperber, Dan and Deirdre Wilson. 1981. "Irony and the Use-Mention Distinction." In Cole 1981, 295-318.

Stieger, Stefan, Anton K. Formann and Christoph Burger. 2011. "Humor Styles and Their Relationship to Explicit and Implicit Self-Esteem." *Personality and Individual Differences* 50 (5): 747-50.

Su, Jian, Kevin Duh and Xavier Carreras, eds. 2016. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin: Association for Computational Linguistics.

Tiedemann, Ed, ed. 2015. *Proceedings of the 58th Global Communications Conference (IEEE GLOBECOM 2015)*. San Diego: Institute of Electrical and Electronic Engineers.

Toutanova, Kristina and Hua Wu, eds. 2014. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Baltimore: Association for Computational Linguistics.

Tsur, Oren, Dmitry Davidov and Ari Rappoport. 2010. "ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews." In Cohen et al. 2010, 162-9.

Van Hee, Cynthia. 2017. "Can Machines Sense Irony?: Exploring Automatic Irony Detection on Social Media." PhD diss., Ghent University.

Van Hee, Cynthia, Els Lefever and Veronique Hoste. 2018. "SemEval-2018 Task 3: Irony Detection in English Tweets." In Apidianaki et al. 2018, 39-50.

Wallace, Byron C. et al. 2014. "Humans Require Context to Infer Ironic Intent (so Computers Probably Do, Too)." In Toutanova and Wu 2014, 512-6.

Wilson, Deirdre. 2013. "Irony Comprehension: A Developmental Perspective." *Journal of Pragmatics* 59: 40-56.

Received 23 July 2021 Revised version accepted 20 September 2021

Antonio Moreno-Ortiz is a Senior Lecturer and researcher at the University of Málaga, where he has worked for more than 20 years. His research interests include computational linguistics, corpus linguistics and language technologies, which has led him to develop multiple linguistic resources for natural language processing, such as BNC Indexer, OntoTerm, Sentitext and Lingmotif.

María García-Gámez is a research assistant at the University of Málaga, where she is currently working on her PhD thesis as a fully funded candidate. She holds a BA degree in English Studies and an MA in English Studies and Multilingual and Intercultural Communication. Her research interests involve corpus linguistics, sentiment analysis and the use of sarcasm in social media.