Irony Detection in Hebrew Documents: A Novel Dataset and an Evaluation of Neural Classification Methods

Avi Shmidman, Elda Weizman, Avishay Gerczuk
Bar-Ilan University, Ramat Gan, Israel
{avi.shmidman,elda.weizman,avishay.gerczuk}@biu.ac.il

Abstract

This paper focuses on the use of single words in quotation marks in Hebrew, which may or may not be an indication of irony. Because no annotated dataset yet exists for such cases, we annotate a new dataset consisting of over 4000 cases of words within quotation marks from Hebrew newspapers. On the basis of this dataset, we train and evaluate a series of seven BERT-based classifiers for irony detection, identifying the features and configurations that most effectively contribute the irony detection task. We release this novel dataset to the NLP community to promote future research and benchmarking regarding irony detection in Hebrew.

1 Introduction

Irony understanding involves a complex interpretation process. Although irony is inherently indirect, its interpretation may be enhanced by textual markers. This paper focuses on the use of one of the most prevalent irony markers – quotation marks enclosing single words. The analysis combines a theory-based, pragmatically oriented textual analysis of the pattern under study with experiments aiming to train a neural network to automatically identify ironic quotation marks and differentiate them from similar non-ironic quotes, used for naming and marking peculiar lexical choices. Whereas ironic quotation marks received some theoretical and experimental attention in pragmatics and in computational linguistics, we are not aware of studies which compare systematically ironic quotes with their non-ironic counterparts.

The paper is structured as follows: following a concise overview of related studies (section 2), we illustrate the three aforementioned categories (section 3), and report on experiments conducted to train neural networks to classify any given instance of a word in quotation marks as one of them. The assumption underlying these experiments is the following: if the distinctions that we have identified are in fact sufficiently indicated within the text, then we would expect a neural network to be able to pick up on these indications and correctly classify these categories. We present the classifier's pitfalls in (7), while the results and their implications for irony detection are discussed in the closing section (8).

2 Related Work

Within the large body of research on verbal irony in pragmatics, it is widely accepted that verbal irony has two defining features: it is inherently indirect, and it necessarily conveys the speaker's attitude, mostly some degree of dissociation and criticism.

Most researchers agree that the interpretation of indirectness in general and irony in particular involves several levels of meaning and context-based identification of some incongruity between them. We rely on Grice's three-level distinction (Grice, 1968; Dascal, 1983) between sentence meaning, utterance meaning and speaker's meaning, whereby speaker's meaning is what the speaker means to convey by uttering a given utterance in a given situation. In irony interpretation, contextual information is exploited for two different purposes: as a *cue*, when it indicates that the utterance meaning is not a plausible candidate for being the speaker's meaning, and as a *clue*, when it is used to compute an alternative, ironic speaker's meaning which, under the circumstances, may be intended by the speaker (Weizman and Dascal, 1991). Full interpretation of the speaker's meaning includes the detection of the ironic criticism, as well as the identification of the victim of irony (towards whom the criticism is addressed) and its locus (towards what it is directed) (Weizman, 2001, 2008). In everyday discourse, indirect speaker's meanings in general and ironic criticism in particular may be missed or reconstructed partially.

Competing pragmatic accounts provide us with

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 91–101
May 3-4, 2025 ©2025 Association for Computational Linguistics

insights into the nature of cues which trigger an ironic interpretation, the major ones being: a blatant flouting of the maxim of quality, ("Try to make your contribution on that is true") (Grice 1975:46, 1978), related to the classic, Aristotelian view of irony as conveying the opposite meaning; a blatant flouting of other Gricean maxims, i.e., the expectations underlying cooperative communication (Colston, 2000; Attardo, 2000); the reversal of evaluation (Partington, 2007; Burgers et al., 2012; Zappavigna, 2022); as a pretense (Clark and Gerrig, 1984; Currie, 2006); and irony as a non-attributive, echoic metarepresentation (Sperber and Wilson, 1981; Wilson and Sperber, 1992, 2012; Wilson, 2012). The latter is specifically related to the use of ironic quotation marks.

In this view, a necessary condition for irony comprehension is the recognition that the speaker implicitly mentions, or echoically metarepresents, a true or imagined proposition, thought, belief, opinion, norm or an interpretation thereof, without explicitly attributing it to its source, be it real or imagined. By so doing, she expresses a derogatory attitude towards the echoed utterance, thought, opinion or their interpretation and implicitly criticizes its source (Sperber and Wilson, 1981, 1986; Wilson and Sperber, 1992, 2012; Wilson, 2012). Accordingly, in ironic utterances the literal meaning is not substituted for by an indirect, opposite meaning. Rather, “the speaker mentions a proposition in such a way as to make clear that she rejects it as ludicrously false, inappropriate, or irrelevant” (Sperber and Wilson, 1981, 308).

Viewing the pattern under discussion as a case of non-attributive metarepresentation explains why the use of quotation marks is non-arbitrary: it might be considered as "borrowed" from typically attributive metarepresentations such as direct speech. Studies indicate that quotation marks are associated with irony (Partington et al., 2013) and play a beneficial role in its recognition and processing (Schlechtweg and Härtl, 2023). The more partial the quotation is vis-à-vis its presumed source, the more likely it is to convey irony (Weizman, 1984). We examine single words in quotation marks since they are manifestly partial in this respect (Weizman, 2020). Longer units in quotes will be explored at a later stage.

From a *pragmatic viewpoint*, the indirect nature of irony presupposes that textual markers are non-obligatory. However, When they do exist they are mostly equivocal, as they may be used for other purposes as well. Studies of irony markers in written discourse shed light on phonological, morphological and non-verbal patterns, including exclamation marks, emoticons and quotation marks in written discourse (Attardo, 2000; Attardo et al., 2003; Partington, 2011; Yus, 2023).

An interesting marker is "multiple uses of irony" (Burgers et al., 2013) or "redundancy" (Hirsch, 2011; Livnat, 2011; Weizman, 2011), whereby various cues for irony or multiple occurrences of irony markers in a given co-text support each other and enhance the identification of irony. This may also apply to numerous uses of quotation marks in the same text (Weizman, 2011).

In pragmatics, the interplay between ironic quotation marks and their co-textual environments in mediated political discourse have received special attention (Gruber, 1993, 2015a,b, 2017; Weizman, 1984, 2001, 2011, 2020, 2022) highlighting their evaluative and attitudinal functions. Weizman (2020) considers *John has been “successful" these last years* as a case of non-attributive echoic metarepresentation, whereby an ironic reading relies on the identification of quotation marks as a marker of echoic mention, which, in turn, is a cue for the detection of a mismatch between the proposition in quotes and contextual information.

Over the past decades, relevant studies in *computational linguistics* have evolved significantly in their approach to irony detection. Initially, researchers focused on lexical and syntactic features, punctuation marks, and positive/negative polarity. In addition to these linguistic features, scholars have particularly emphasized the role of non-verbal elements in social media contexts, such as emoticons and hashtags (e.g., Wallace 2013; Joshi et al. 2017; Golazizian et al. 2020; Veale 2021; Wiślicki 2023; Chen et al. 2024).

In terms of computational modeling, early approaches primarily relied on statistical methods, specifically utilizing features like bag-of-words (Wallace et al., 2015) and pattern-based analysis (Davidov et al., 2010a,b). Building upon these foundations, researchers then developed rule-based approaches, examining elements such as sentiment disparity between hashtags and text content on Twitter (Van Hee et al., 2018). Despite their contributions to the field, these methods proved to be time- and labor-intensive (Chen et al., 2024). Consequently, the field has witnessed a shift toward more sophisticated approaches, particularly deep-learning techniques. For instance, the use of similarity between word embeddings as features for sarcasm detection (Joshi et al., 2017).

Throughout this evolution, quotation marks have consistently been included with other markers of irony in several multi-variant studies of irony detection in computational linguistics and neighboring approaches (e.g., Carvalho et al. 2009, 2011; Davidov et al. 2010a,b; Buschmeier et al. 2014; Karoui et al. 2015, 2017). Furthermore, while these various approaches have advanced our understanding, most models continue to treat irony as a rhetorical device or figure of speech rather than a pragmatic phenomenon, often employing binary classifications (ironic vs. non-ironic). Moreover, their cotextual environments and various functions have not received specific consideration.

In our data, quotation marks enclosing single words are used for three purposes – conveying *irony*, *naming* and marking the journalist's awareness of a peculiar lexical choice (henceforth *lexical peculiarity*). We proceed to illustrate the distinction between them.

3 Analysis

The textual realizations of all three functions are identical: each consists of a single word in quotation marks. Furthermore, since in Hebrew there are no capital letters, the category of naming is not formally differentiated from the two other categories in any way. The following utterances represent the three categories:

1. They are very particular about saying **"halel"** every day.
2. This is the time of "**how**".
3. People all over the world are murdered because they do not belong to the **"right"** religion.

In example (1), the quotes encolsing *halel* mark a proper name – the name of a Jewish prayer. In (2), the quotes indicate that the journalist is aware of the non-normative use of an interrogative adverb as a noun. In (3), the word in quotes, *right*, echoically metarepresents the belief that religions may be perceived as either right or wrong, and convey the journalist's ironic criticism of this simplistic and harmful perception. Hence, whereas in example (3) the quotation marks are metarepresentational and typically judgmental, in example (1) they are referential and in example (2) they are meta-linguistic since they convey the speaker's linguistics awareness. Additionally, whereas in (1) and (2) the quotes are local, in the sense that they pertain to the meaning or the form of the word they enclose, in (3) the conveyed stance touches upon a larger co-textual environment since the ironic criticism is directed also at the belief that prescriptive judgments of religion may justify murders on its behalf.

3.1 Category 1: Naming

In our data, naming quotes usually indicate the title of a book, journal, institution, party, prayer, or a widely accepted concept.

Typically, the identification of this function is based on the reader's acquaintance with its extralinguistic specific context. This is the case in example (1), where "halel" designates the name of a prayer, as well as in example (4) below, where "gesher" is the name of a political party:

4. It is difficult to understand how an experienced politician like Peretz can believe even for a moment that the alliance with **"gesher"** could change the basic formula of Israeli politics. (Ze'ev Sternhell, *Ha'aretz,* 23.8.2019)

Naming may be utterly context-dependent (Ex. 1,4) or supported by the contextual enviornment (Weizman 2020; 2022), for example through the construction of a semantic field (Ex. 5) (explicitations underlined):

5. On July 28, Vygotsky's coffin was placed on the stage of the theatre where he was supposed to play the Danish prince in **"hamlet"**. (Dimitry Shumsky, *Ha'retz,* 23.7.2020)

3.2 Category 2: Lexical peculiarity

The quotes falling under this category convey the speaker's meta-linguistic awareness of and distanciation from the lexical peculiarity of the word or phrase enclosed in them. Typical uses include live metaphors, slang, connotations, register shift and code-switching. In a way, the speaker implicitly admits that his or her linguistic choice may be viewed as unacceptable for some reason, or is being "apologetic" (Predelli, 2003, 2), but insists on using it. This category partly overlaps with scare quotes (Predelli, 2003; Schlechtweg and Härtl, 2023).

The following examples illustrate quotes marking register shift from formal language to slang ("blanked on", Hebrew *fisfes*, 6), a live metaphor ("fat", Hebrew *shamen*, designating the public sector considered as avid consumer, 7) and euphemism ("the illness", Hebrew *hamaxala*, avoiding specific reference to its nature, 8):

6. However, in the ruling it was determined that the first examination was indeed negligent, and the doctor **"blanked on"** [missed, Hebrew *fisfes*] the defect in the fetus. Had the defect been discovered then, the pregnancy could have been terminated. (Assaf Posner, *Ha'aretz*, 16.7.2019)

7. Despite the image he [PM Netanyahu] built for himself, he failed miserably in the domain of economics. [. . .] He did not take care of the **"fat" [***shamen***]** (the public sector), which he made even fatter [*shamen yoter*]. (Nehemia Shtrasler, *Ha'aretz*, 22.9.2020)

8. "I am still within the thirty-day mourning period of my partner's passing from **"the illness"** [Hebrew *maxala*]. (No Name, *Ha'aretz*, 8.8.2019)

3.3 Category 3: Irony

As explained above (section 2), the use of quotation marks, which typically mark *attributivee* metarepresentations (e.g. in reported speech) supports the view of ironic quotation marks as conveying an echoing, *non-attributive* metarepresentation of a previous utterance, thought, concept, norm or their interpretation, and the criticism they convey may be directed at the wording of the echoed source, its content or both (Sperber and Wilson, 1981; Weizman, 1984; Wilson and Sperber, 1992, 2012; Wilson, 2012). This is the case in the following examples.

9. Yes, as long as Arab men in Arab society continue to sanctify and protect their **"honor"** and their **"pride"**, Arab women will be murdered. (Shirin Fallah Saab, *Ha'aretz*, 24.11.2020)

Through the use of ironic quotes, the journalist mentions cultural keywords characterizing traditional perceptions and beliefs, without explicitly attributing them to specific sources. By so doing, she conveys harsh criticism addressed at the society who practices them.

10. The Knesset committee, which was established last week specifically in order to discuss Prime Minister Benjamin Netanyahu's request for immunity, found itself on Thursday discussing **"only"** the request for immunity submitted by MP Katz (Likud), after the Prime Minister had withdrawn his request at the last minute. (Editorial, *Ha'aretz*, 2.2.2020)

This unsigned editorial of *Ha'aretz* has been published against the background of two requests for immunity, submitted to a special Knesset [Israel parliament] committee by Israel PM Benjamin Netanyahu and by MP Israel Katz, both accused of fraud and breach of confidence. At the end of the editorial, the writer calls upon the special committee to reject MP Katz's request. In the utterance under consideration, the word in quotes ("only") echoically metarepresents the arguments of those who underestimate the severity of the MP's conduct. The ironic criticism seems to be addressed at the committee in particular and possibly at public agents in general, for not taking seriously legal accusations.

11. In order to win the elections and bring the [center-left] bloc under one roof, [the party] *kaxol-lavan* [="Blue and White"] must include Yoaz Handel in its list. [. . .] One fact stands out: a center party that aspires to succeed should display in its showcase a handsome, talented young man, considered a **"moderate"** right-wing person. Why? because [the party's] leaders believe that striving for a peace settlement, opposing the annexation of territories and demanding to abolish the nationality law will not earn it the status of a leading power. (Uzzi Bar'am, *Ha'aretz*, 20.1.20).

In this extract, the journalist criticizes the center-party *Kaxol Lavan* for attending to populist strategies (such as calling upon a handsome politician to join it) at the expense of ideological principles. By enclosing *"moderate"* in quotation marks, he echoically mentions the party's presumed evaluation of Hendel's political orientation and challenges the belief that Hendel is indeed moderate. The irony is further directed at the belief that a right-wing politician can indeed be considered moderate.

So far, we presented a pragmatic analysis of single words in quotation marks and illustrated the different functions they fulfill in context – naming, awareness of lexical peculiarity and ironic criticism, foregrounding the role of co-text in solving some of the complexities involved in their interpretation. If the distinctions that we have identified are in fact sufficiently indicated within the text, then we would expect a neural network to be able to pick up on these indications and correctly classify these specimens.

Thus, we proceed to present our annotated dataset for Hebrew irony, followed by our neural-network experiments upon the dataset.

4 Annotated Dataset for Hebrew Irony

Our dataset is annotated to distinguish ironic uses of quotation marks from other uses. It is the first of its kind in Hebrew, since we are not aware of any other datasets comprised to address the phenomenon of ironic quotes. The dataset consists of op-eds from major and popular Israeli news platforms. We collected the data using two methods: (a) Automated crawling of op-ed articles from the opinion sections in the platforms in 2019-2020. This data was collected by a social media monitoring and analysis company. (b) Manual collection of op-ed articles published in 2020.

The data was annotated by three pragmatics experts, who annotated each instance of a single word enclosed in quotation marks in the context of the entire article, distinguishing between *naming/lexical peculiarity/irony*. In case of disagreement, two labels were assigned to the disputed word, such that the label assigned by two annotators preceded the label assigned by a single annotator. The classifier considered only the first label for the target word. On the whole, we have 59 cases (1.4%) of double annotation.The vast majority of these (56) are related to the distinction between *irony* and *lexical peculiarity*.

We are pleased to release this new annotated dataset to the NLP community.[1]

5 Experimental Setup

We train neural networks to classify any given instance of a word enclosed within quotation marks (henceforth: “target word”) as one of the aforementioned classes: "Naming", "Lexical Peculiarity" or "Irony". The foundational model underlying our experiments is DictaBERT, the current state-of-the-art BERT model for modern Hebrew (Shmidman et al., 2023).

Statistics	**Count**
Total Documents	2,700
Total Words	1,504,153
Category Distribution	
Naming	1,889 (45.1%)
Lexical Peculiarity	980 (23.4%)
Irony	1,321 (31.5%)
Total	4,190

Table 1: Statistics on the number of documents, words and category distribution in total in the data collection

We run each of our sentences through DictaBERT in order to produce a contextual embedding for each instance of a word within quotation marks. We then aim to train a multi-layer perceptron (MLP) to classify each instance of these contextual embeddings into one of our three categories. As we describe in detail below, we experiment with multiple trains of such an MLP, each time progressively providing the classifier with more information about the word, the sentence, and the surrounding context, in order to determine how much information is truly needed to correctly assess the presence or absence of irony within the quoted word. All MLPs are trained 10 epochs, with a learning rate of 0.0001, a hidden layer of size 100, with the Adam optimizer, and a batch size of 32. We evaluate the performance of each MLP using 10-fold cross validation; we calculate separate recall, precision, and F1 scores for each of the classes.

6 Experiments and Results

6.1 Masking the target word

In our initial experiment, we mask the target word; thus, the contextual embedding produced by DictaBERT is informed only by the word's prior and subsequent co-text. The point of this experiment is to see whether the information regarding the ironic usage is sufficiently encoded within the surrounding words, without regard for the target word itself. Results are displayed in Table 2.

	Precision	Recall	F1
Irony	71.0%	86.5%	.780
Naming	87.8%	82.4%	.850
Lexical Peculiarity	61.6%	40.8%	.491

Table 2: Results when masking the target word

This certainly leaves room for improvement; yet

[1] https://www.dropbox.com/scl/fo/3ssz89hlfqvhwjcfewnsy/ABb_z1GUjKYvVkeXp4ski1A?rlkey=a1gwwwbbw1ncgxzf57ndbefu3&st=4ve7im2t&dl=0

it is remarkable that the system was able to correctly identify so many cases of irony based on the sentence co-text alone (F1 score of 0.780 for the irony category).

6.2 Unmasking the target word

In our second experiment, we unmask the target word, to see whether knowledge of the specific word improves the system's ability to classify the cases. Indeed, this improves our success rates substantially in all three categories. Results are displayed in Table 3.

	Precision	Recall	F1
Irony	76.9%	87.2%	.818
Naming	92.5%	88.5%	.903
Lexical Peculiarity	63.2%	50.3%	.560

Table 3: Results when unmasking the target word

6.3 Adding the CLS embedding

In this experiment, we keep the unmasked embeddings of the target word as per the previous experiment, and we add in the “CLS" embedding produced by DictaBERT for the sentence overall. This embedding is concatenated to the embedding of the target word, and the result of the concatenation is provided as input to the MLP. Our theory is that this embedding could provide an overall characterization of the sentence supporting or discouraging an ironic reading of the target word. Indeed, adding the CLS embedding boosts our F1 score for all three categories. Results are in Table 4.

	Precision	Recall	F1
Irony	78.0%	87.6%	.825
Naming	93.3%	88.0%	.905
Lexical Peculiarity	64.8%	54.4%	.591

Table 4: Results when adding the CLS embedding

6.4 Adding more extensive co-text

In this experiment, we continue to build upon the successful setup of the previous experiment (unmasked embedding plus CLS token), and we attempt to further bolster the system's ability to classify the target word by providing it with more co-text. When generating the unmasked contextual embedding from DictaBERT, in addition to the sentence containing the target word, we also provide the preceding sentences within the paragraph (up to a maximum of five sentences). Thus, when DictaBERT calculates the embedding for any given target word, it does so with an eye toward the preceding sentences as well.

Results are displayed in Table 5. It turns out that the extra co-text does not improve our ability to recognize instances of irony. In fact, it caused the F1 score for the "irony" category to turn downwards. Overall, it seems that the extra co-text only added extra clutter, and did not provide helpful clues for identifying irony.

	Precision	Recall	F1
Irony	77.4%	87.7%	.822
Naming	93.2%	88.6%	.908
Lexical Peculiarity	63.1%	51.2%	.566

Table 5: Results when adding more extensive co-text

6.5 Adding extra redundancy information

In this experiment, we add three extra pieces of information to each training sample. The three pieces are as follows: (a) an embedding indicating how many pairs of quotation marks were used in the paragraph (0, 1-2, or 3+); (b) an embedding indicating the paragraph size (under 500 words, 500-1000 words, or more than 1000); (c) an embedding indicating how often the target word recurs within the paragraph (0, 1, or 2+). This information is aimed at testing the effect of redundancy (section 2.2) on the irony detection mechanism. We concatenate this extra information together with the unmasked embedding of the target word and the CLS token.

Results are displayed in Table 6. It turns out that these extra pieces of information do not improve the system's ability to identify irony; the F1 score for the irony category is lower than when we train with only unmasked embeddings and CLS, without the extra information. Regarding the other two categories, this method provides a slight boost in the F1 score of the lexical peculiarity category, but at the same time slightly lowers the F1 score of the naming category.

	Precision	Recall	F1
Irony	78.5%	86.6%	.823
Naming	92.7%	88.0%	.903
Lexical Peculiarity	64.3%	56.1%	.599

Table 6: Results when adding redundancy information

In summary, although the system achieves impressive accuracy in detecting irony based on the co-text of the sentence alone (in the masked scenario), knowledge of the target word does substantially improve our accuracy. Adding in the CLS token boosts the accuracy even higher. However, our other attempts to add extra information, whether via extra co-text, or via information regarding density and redundancy, did not advance the accuracy any further.

6.6 Binary Experiments

Having established our ideal approach – that is, using an unmasked target word and concatenating the CLS embedding – we proceed to utilize this approach in training three separate binary classifiers, in order to focus on the system's ability to recognize each category individually.

Irony vs. Other. In this experiment, we train a classifier to identify each specimen as either "Irony" or "Not Irony". Results are displayed in Table 7. The classifier's ability to identify irony remains about the same as with our most successful three-class experiment above.

	Precision	Recall	F1
Irony	79.7%	85.3%	.824
Non-Irony	87.2%	82.1%	.846

Table 7: Binary Classification (Irony vs. Other Categories)

Lexical Particularity vs. Other. As we saw above, identifying the category of lexical peculiarity is particularly difficult for our neural network; in the three-class classifiers, the precision and recall scores for this category were consistently low. Our binary classifier for this category also proved to be rather unsuccessful. The results in Table 8 demonstrate how much the system struggles with this category.

	Precision	Recall	F1
Lexical Peculiarity	67.7%	43.4%	.529
Not Lexical Peculiarity	84.4%	93.7%	.888

Table 8: Binary Classification (Lexical Peculiarity vs. Other Categories)

Naming vs. Other. The category of "Naming" is the easiest category to spot. As we saw above, the precision and recall numbers were consistently high for this category. Indeed, when we train a binary classifier to distinguish between Naming and Not Naming, we achieve F1 scores above 0.90 for both classes; results are displayed in Table 9:

	Precision	Recall	F1
Naming	95.1%	86.6%	.906
Not Naming	94.1%	98.0%	.960

Table 9: Binary Classification (Naming vs. Other Categories)

7 Where does the model fail?

Analyzing the model's failures may be beneficial for improving its performance. The following illustrate two errors specifically related to the distinction between *irony* and *lexical peculiarity*:

12. This is one of the reasons why our organization requested to join as a **"friend"** of the court in the case of J.

The expression *friend of the court* is the Hebrew legal term for *amicus curiae*. The live metaphor "friends" was annotated by the experts as *lexical peculiarity*. The neural network, on the other hand, classified it as *irony*, possibly due to its emotive value, which lends itself to a reversal of meaning.

13. Facebook has completely distorted clear concepts such as **"social"** or **"friends"**.

The three experts read both quotes as echoic mentions of misconceptions, further relying on the journalist's criticism implied by *distorted*, and therefore annotated them as irony. The neural network classified the target words as *lexical peculiarity*, possibly influenced by their qualification as 'clear concepts', which is textually closer to the target words than the verb *distorted*.

8 Discussion and conclusions

Starting with the premise that irony is necessarily indirect, this paper aims to delve into the nature of irony detection, by combining pragmatic analyses with experimentation purporting to train neural networks to identify ironic speaker's meaning. Through this experiments we can learn about the validity of our predictions and improve them where necessary. With this purpose in mind, we focused on single words enclosed in quotation marks, conceptualized as textual realizations of non-attributive echoic metarepresentation which, in turn, is a possible cue for the detection of a mismatch between the proposition in quotes and contextual information. The analysis of ironic quotation marks shows that a full interpretation of the speaker's ironic meaning requires the detection of echoic mention, somewhat facilitated by the quotation marks, and the identification of the victim of irony (who is being criticized) and its target (what is being criticized. Since the textual pattern under study fulfills two additional functions – naming and marking the speaker's awareness of a peculiar lexical choice, we proposed a distinction between these three polysemous patterns, foregrounding the pragmatic differences between them. To our knowledge, no such comparison has been made before.

Drawing on the pragmatic distinction, we proceeded to examine to what extent the three patterns are distinguished by a neural network, with the underlying assumption that if the distinctions identified through pragmatic analysis are sufficiently indicated in the text, then we would expect a neural network to be able to pick up on these indications and correctly classify these categories.

All in all, the experiments yielded good results concerning our primary goal, i.e. the classifier's ability to identify cases of irony (F1 score of .825, as per Table 4). However, we were surprised to find that this ability was not improved by the addition of extra co-text, nor with the addition of extra information regarding redundancy (the number of single words in quotation marks used in the paragraph, the paragraph size and how often the target word recurs within the paragraph). One possible explanation may be that since DictaBERT was mostly trained on single sentences, its familiarity with complex co-textual environments is limited. It is noteworthy, however, that in the majority of ironic quotation marks which were correctly classified based on the sentence alone, the information that was available within the target sentence yielded a good result. Still, the role of the co-text in ironic interpretation has been widely acknowledged in pragmatic research in a way that encourages us to delve in the textual analysis, further characterize the supportive co-text and conduct additional experiments to test this characterization.

As for the other two categories, we obtained very good results regarding its ability to distinguish 'Naming' from the two other categories (F1 score of .906, as per Table 9). The category 'Lexical Peculiarity', however, is more challenging: 67.7% precision and 43.4% recall in the binary experiment (Table 8). This is not very surprising if we consider that the category 'Lexical Peculiarity' has some resemblance to 'Irony' since both convey some degree of the speaker's negative attitude and involve meta-pragmatic awareness. The difference is that in our data, 'Irony' usually conveys the speaker's harsh criticism, its victim is mostly an echoed third party (self-irony is rare in journalistic op-eds) and its locus varies depending on the context, whereas 'Lexical peculiarity' conveys mild distanciation, its target is the speaker herself and its locus is invariably some linguistic choice she has made. The results indicate the need to refine the analysis of this category and the experimental design related to it. We intend to start by exploring the lexical specificity of the peculiar lexical choice enclosed in quotation marks. At this stage of the research, we believe that the classifier can indicate a "red flag" over specific words in the text, alerting the reader to the fact that they might convey ironic speaker's meaning. Nevertheless, the classifier is not yet perfect, and it would certainly be preferable to improve its accuracy before its deployment.

To conclude, we adopt Gibbs and Colston's (2023:9) view:

> We typically believe that irony is a completely human affair, but there have been interesting attempts to create computational models of irony use and understanding. [. . .] One of the beauties, and major challenges of computer modeling is that it forces researchers to make concrete decisions on how best to implement some linguistic observation or theoretical idea (e.g., how to create a workable model of echoic mention, pretense, or what is meant by incongruity).

This statement introduces Veale's (2023) discussion of computational models designed to detect irony and produce it. Veale compares various computational models and proposes his EPIC model, combining a theoretical approach with computational expertise, and concludes: "A computational approach to irony is no substitute for an actual theory of irony".

The two sides of the mirror are illuminated: Gibbs and Colston (2023) highlight the potential contribution of computational studies to pragmatics, whereas Veale (2023) manifestly foregrounds the indispensable contribution of theoretical thinking to a computational approach. The belief in this mutual contribution has been underlying the study we describe in this paper.

Acknowledgements

The work of the first author has been funded by the Israel Science Foundation (Grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829; Principal investigators: Nachum Dershowitz, Tel-Aviv University; Judith Olszowy-Schlanger, EPHE-PSL; Avi Shmidman, Bar-Ilan University, and Daniel Stoekl Ben Ezra, EPHE-PSL), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The data collection and annotation has been partly supported by a grant from the German Israeli Foundation for Scientific Research and Development (GIF Grant I-1475-104.4/2018), awarded to the second author in collaboration with Prof. Anita Fetzer.

References

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

Salvatore Attardo, Jodi Eisterhol, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260.

Christian Burgers, Margot van Mulken, and Peter Jan Schellens. 2012. Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics*, 44(3):231–242.

Christian Burgers, Margot van Mulken, and Peter Jan Schellens. 2013. The use of co-textual irony markers in written discourse. *Humor*, 26(1):45–68.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49.

Paula Carvalho, Luis Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, pages 53–56.

Paula Carvalho, Luís Sarmento, Jorge Teixeira, and Mário J. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568.

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:1–18.

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Herbert L. Colston. 2000. On necessary conditions for verbal irony comprehension. *Pragmatics and Cognition*, 8(2):277–324.

Gregory Currie. 2006. *Why Irony is Pretence*, page 111–134. Oxford University PressOxford.

Marcelo Dascal. 1983. *Pragmatics and the philosophy of mind*. Pragmatics & Beyond. John Benjamins Publishing, Amsterdam, Netherlands.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010a. Enhanced sentiment learning using Twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, Beijing, China. Coling 2010 Organizing Committee.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010b. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Raymond W. Gibbs and Herbert L. Colston. 2023. *Irony and Thought: The State of the Art*, pages 3–14. Cambridge Handbooks in Psychology. Cambridge University Press.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony detection in Persian language: A transfer learning approach using emoji prediction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2839–2845, Marseille, France. European Language Resources Association.

Herbert Paul Grice. 1968. *Utterer's Meaning, Sentence-Meaning, and Word-Meaning*, page 49–66. Springer Netherlands.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and Semantics: Speech Acts*, III:41–58.

Herbert Paul Grice. 1978. Further notes on logic and conversation. volume 9, pages 113–127. Pragmatics. Academic Press.

Helmut Gruber. 1993. Evaluation devices in newspaper reports. *Journal of Pragmatics*, 19(5):469–486.

Helmut Gruber. 2015a. *Intertextual references in Austrian parliamentary debates: Between evaluation and argumentation*, page 25–56. John Benjamins Publishing Company.

Helmut Gruber. 2015b. Policy-oriented argumentation or ironic evaluation: A study of verbal quoting and positioning in austrian politicians' parliamentary debate contributions. *Discourse Studies*, 17(6):682–702.

Helmut Gruber. 2017. Quoting and retweeting as communicative practices in computer mediated discourse. *Discourse, Context amp; Media*, 20:1–9.

Galia Hirsch. 2011. Redundancy, irony and humor. *Language Sciences*, 33(2):316–329.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 50(5):1–22.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Jihen Karoui, Farah Benamara Zitoune, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 644–650, Beijing, China. Association for Computational Linguistics.

Zohar Livnat. 2011. Quantity, truthfulness and ironic effect. *Language Sciences*, 33(2):305–315.

Alan Partington. 2007. Irony and reversal of evaluation. *Journal of Pragmatics*, 39(9):1547–1569.

Alan Partington. 2011. Phrasal irony: Its form, function and exploitation. *Journal of Pragmatics*, 43(6):1786–1800.

Alan Partington, Alison Duguid, and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins Publishing Company.

Stefano Predelli. 2003. Scare quotes and their relation to other semantic issues. *Linguistics and Philosophy*, 26(1):1–28.

Marcel Schlechtweg and Holden Härtl. 2023. Quotation marks and the processing of irony in english: evidence from a reading time study. *Linguistics*, 61(2):355–390.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *Preprint*, arXiv:2308.16687.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. In P. Cole, editor, *Radical pragmatics*, pages 295–318. Academic Press, New York.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell, Oxford.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. We usually don't like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Tony Veale. 2021. *Your Wit Is My Command: Building AIs with a Sense of Humor*. The MIT Press.

Tony Veale. 2023. *Great Expectations and EPIC Fails: A Computational Perspective on Irony and Sarcasm*, page 216–234. Cambridge Handbooks in Psychology. Cambridge University Press.

Byron C. Wallace. 2013. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 43(4):467–483.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Elda Weizman. 1984. Some register characteristics of journalistic language: Are they universals? *Applied Linguistics*, 5(1):39–50.

Elda Weizman. 2001. Addresser, addressee and target. In *Negotiation and Power in Dialogic Interaction*, pages 125–137. John Benjamins Publishing Company, Amsterdam.

Elda Weizman. 2008. *Positioning in Media Dialogue*. Dialogue Studies. John Benjamins Publishing, Amsterdam, Netherlands.

Elda Weizman. 2011. Conveying indirect reservations through discursive redundancy. *Language Sciences*, 33(2):295–304.

Elda Weizman. 2020. The discursive pattern 'claim+ indirect quotation in quotation marks': Strategic uses in french and hebrew online journalism. *Journal of Pragmatics*, 157:131–141.

Elda Weizman. 2022. Explicitating irony in a cross-cultural perspective: Discursive practices in online op-eds in french and in hebrew. *Contrastive Pragmatics*, 4(3):437–465.

Elda Weizman and Marcelo Dascal. 1991. On clues and cues: Strategies of text-understanding. *Journal of Literary Semantics*, 20(1):18–30.

Deirdre Wilson. 2012. *Metarepresentation in linguistic communication*, page 230–258. Cambridge University Press.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87(1–2):53–76.

Deirdre Wilson and Dan Sperber. 2012. *Meaning and Relevance*. Cambridge University Press.

Jan Wiślicki. 2023. Scare quotes as deontic modals. *Linguistics*, 61(2):417–457.

Francisco Yus. 2023. Inferring irony online. In *The Cambridge Handbook of Irony and Thought*, pages 160–180. Cambridge University Press.

Michele Zappavigna. 2022. Social media quotation practices and ambient affiliation: Weaponising ironic quotation for humorous ridicule in political discourse. *Journal of Pragmatics*, 191:98–112.