Guidelines for Writing and Evaluating Explanations for Irony in Tweets

LT3 Technical Report – LT3 25-02

Aaron Maladry, Cynthia Van Hee,
Els Lefever and Véronique Hoste

LT3, Faculty of Arts and Philosophy, Ghent University, Belgium.

May 14, 2025

Contents

1 Introduction 1

2 Annotating Irony Explanations and World Knowledge 3

3 Explanation Evaluation 6

4 Data Statement 10

Chapter 1

Introduction

Human language is rich, complex, and often ambiguous. The field of Natural Language Processing (NLP) combines insights from linguistics with the capabilities of modern machine learning to model human language computationally. From syntactic parsing to open-domain text generation, NLP seeks to bridge the gap between human expression and machine understanding.

A key challenge in this field is understanding **figurative language**; a creative use of language in which the intended meaning differs from the literal one. In this work, we focus on a specific type of figurative language, namely irony. Drawing on longstanding theoretical research, we propose the following **working definition of irony**, which encompasses two forms: verbal and situational irony.

> *verbal irony is a type of figurative language in which the intended meaning of a message contrasts with its literal interpretation* (Wilson and Sperber, 2012).

For example, the statement "I do enjoy getting sick when I'm on vacation!" should not be taken at face value; rather, it expresses frustration or annoyance about the situation. In general, (verbal) ironic statements violate Grice's maxims of communication (Grice, 1975), which outline the expectations that communication should be (1) truthful, (2) relevant (i.e. the information should serve a meaningful purpose in the context of the conversation), and (3) appropriately informative (neither too brief nor overly detailed). Within this framework, Grice (1975) defines irony, like other forms of figurative language, as a violation of the maxim of Quality, which centers on truthfulness. In many cases, verbal irony can rely on **sentiment clashes** (Riloff et al., 2013; Van Hee, Lefever, and Hoste, 2016a), where the opinion expressed (e.g., "I do enjoy") about a topic or situation ("getting sick when I'm on vacation") contradicts the expected sentiment in that context.

Irony can also manifest as a nonsensical, exaggerated or **inappropriate reaction** aimed at criticizing a previous statement. When the proud Italian cook Gino D'Acampo is confronted with the idea of a "British Carbonara", he jests "and if my grandmother had wheels, she would be a bike". With this absurd statement, the message is not about the content, but only expresses an opinion or draws attention to something, here, how ridiculous he finds the concept. This makes the example a so-called "echoic mention" (Sperber and Wilson, 1986). Common linguistic devices used to convey irony include hyperboles (Kreuz and Roberts, 1995), simile (Hao and Veale, 2009), and rhetorical questions (Kreuz et al., 1999). Our working definition also covers **sarcasm**, as a form of verbal irony that is specifically intended to ridicule or hurt.

Recognizing irony often requires **external background knowledge**, as the literal content of the statement may not clearly indicate its ironic intent. The goal of our research is to make this background knowledge, which is implied, explicit through human-generated explanations. The annotation guidelines presented here are designed to facilitate this process.

Besides verbal irony, we also include **situational irony** within our working definition. Like verbal irony, situational irony involves a contrast between expectations and reality. However, unlike verbal irony, it does not involve a shift in the meaning of an utterance. However, unlike verbal irony, it does not involve a shift in the meaning of an utterance. Instead, the irony arises from the situation itself, and the expression or acknowledgment of this contrast often serves to highlight its comical or critical effect.

Many NLP tasks, including irony detection, are often conceptualized as **classification problems**. As a result, linguistic phenomena are often reduced to categorical judgments. For example, determining whether a text expresses positive or negative sentiment, whether it conveys emotions such as sadness, happiness, or anger, or, in our case, whether a text is ironic or not. As stated earlier, figurative language -particularly irony- remains a significant challenge for NLP. In response, several shared tasks have been organized specifically for irony detection (Van Hee, Lefever, and Hoste, 2018; Cignarella et al., 2018; Farha et al., 2022), advancing model development and achieving relative success. The rise of large language models (LLMs) such as GPT, Llama (Grattafiori, Dubey, and et al., 2024), and DeepSeek (DeepSeek-AI et al., 2025), has significantly transformed the field. These decoder-only architectures are not only capable of classifying text but also of generating coherent, open-ended responses and natural language explanations. This means that, rather than simply identifying whether a statement is ironic, we can now prompt models to articulate *why* it is ironic - shifting the task **from detection to interpretation**. However, this shift introduces a new set of challenges: What constitutes a good explanation of irony? And how can we evaluate the quality of such explanations, especially when multiple interpretations may be valid? These guidelines are designed to produce a high-quality corpus that serves two goals: (1) enabling linguistic analysis of irony in natural discourse, and (2) offering structured training data for generative models tasked with explanation writing.

Before automatic systems can be reliably trusted to explain irony in a meaningful way, we must first investigate to what extent their system-generated explanations align with human reasoning. Assessing open-text explanations presents unique challenges. Unlike classification tasks, where model outputs can be directly compared to gold-standard labels, explanations are inherently open-ended. Multiple phrasings may convey the same underlying reasoning, and in some cases, more than one explanation may be valid. Consequently, **human judgment** remains an essential component of the evaluation process.

This report is structured as follows. Chapter 2 introduces our annotation framework for producing human-written explanations of ironic content and presents some annotation examples. In Chapter 3, we outline our evaluation framework, tailored to assess the quality of irony explanations while accounting for their inherent subjectivity and variability.

These annotation guidelines were guided by existing research on irony while intentionally allowing room for individual interpretation. This approach reflects the inherently subjective nature of irony and acknowledges the interpretive flexibility required when dealing with figurative language. These guidelines were applied to the iRONNIE dataset, published as part of xxxxx (paper under review), for which we provide a data statement in the style of Bender and Friedman (2018) in Chapter 4.

Chapter 2

Annotating Irony Explanations and World Knowledge

As a starting point for the explanation guidelines, annotators are introduced to our conceptual definition of irony (Chapter 1), which highlights the contrast between literal meaning and intended meaning. Understanding this contrast, and by extension, the intended meaning of a statement, requires more than just linguistic cues; it depends on shared background knowledge between writer and reader. This is commonly referred to as common knowledge (Lewis, 1969) or mutual knowledge (Schiffer, 1972). However, as noted by Sperber and Wilson (1986), any utterance can be connected to a large collection of shared knowledge, but not all knowledge is relevant to understand the intention. Based on this insight, explaining irony comes down to identifying a contradiction between an expression in the text and relevant world knowledge.

Therefore, **irony explanations** should contain three components: (1) a statement that should not be understood literally, (2) the relevant world knowledge that helps them understand that the statement should not be understood literally and (3) the reasoning that connects these two types of information and that makes the contradiction clear. In some cases (see Example 1), the contradiction between a evaluation and a situation can seem obvious, because the situations are common and the association with a negative sentiment is so obvious. However, the underlying knowledge can exhibit varying degrees of implicitness and require an additional reasoning steps. In Example 5, the symptom of a stuffed nose first has to be connected to having a cold and the fact that it causes trouble breathing, which becomes annoying when you are trying to sleep.

Example 1 *Explicit contradiction:*

Tweet: cleaning up spew is a definite highlight of my job!!

Explanation: Cleaning up spew is dirty and unpleasant work. Therefore, it would not be the highlight or positive side of any job.

Example 2 *Additional reasoning step:*

Tweet: Nose is stuffed. Awesome.

Explanation: A stuffed nose is a common symptom of a cold. A stuffed nose can be annoying because it makes it hard to breathe through your nose. This is especially annoying when you try to sleep. Therefore, a positive evaluation is likely ironic.

To optimize the coverage and to capture as much relevant background knowledge as possible, from varying degrees of implicitness, we instruct annotators to think about all relevant knowledge they use that is not explicitly mentioned in the tweet. In making this knowledge explicit, they should be as thorough as possible and avoid assuming that any background knowledge is obvious. To help illustrate this, we suggested scenarios such as the "explain it to me like I'm five" approach, or imagining the explanation being given to someone unfamiliar with the culture or language, like an alien with no knowledge of this world. These scenarios were designed to emphasize that the required knowledge, even though it might seem obvious, should be made explicit. To avoid producing too much irrelevant information, we first ask the annotators to write an explanation for the irony in a tweet and then immediately follow this up knowledge extraction. By closely linking these two steps, we encourage annotators to ground their explanations in background knowledge.

After writing an explanation, the annotators are asked to **isolate the world knowledge** they used in their explanation. Where appropriate, this background knowledge should be split into separate knowledge items. These items may be slightly rephrased to ensure they are meaningful and coherent when presented out of context. This allows us to reconstruct the body of relevant shared world knowledge, excluding any reasoning steps and explicit evaluations from the tweet.

After receiving instructions and being introduced to several examples, annotators participated in a **training session** where they explained the irony in 30 tweets. During this phase, they were encouraged to ask questions and request clarification as needed. After completing the initial annotations, we reviewed and discussed the results, providing additional guidance where necessary. Annotators then proceeded to annotate the remaining 100 tweets, which were spot-checked to ensure quality and consistency. The practical implementation of the annotations was conducted using the **Label Studio** platform.

Given the subjective nature of irony and the possibility that different annotators might have varying interpretations, specific guidelines are provided for handling texts where the contrast was either too explicit (not requiring world knowledge) or the annotator felt the ironic intent didn't need an explanation. These instructions included **skipping or modifying** texts.

Example 3 *Knowledge Extraction 1*

Tweet: Accidentally breaking your own computer when your tech support! employee of the month #techSupport

Explanation: Someone who works as tech support is responsible for fixing computers, the fact that this person broke their computer contrasts the fact that their job is to fix other people's computers.

Extracted knowledge: 1. Tech support is responsible and often contacted for fixing computers.

Example 4 *Knowledge Extraction 2*

Tweet: Sat on yet another stationary @user train for 15 minutes with no explanation from the driver. Your service is just top drawer.

Explanation: When a train is does not leave on time and is delayed, passengers expect to know why. When these expectations are not met, they are likely not satisfied about the service. This makes the evaluation that the service is "top drawer" (i.e. very good) likely not genuine.

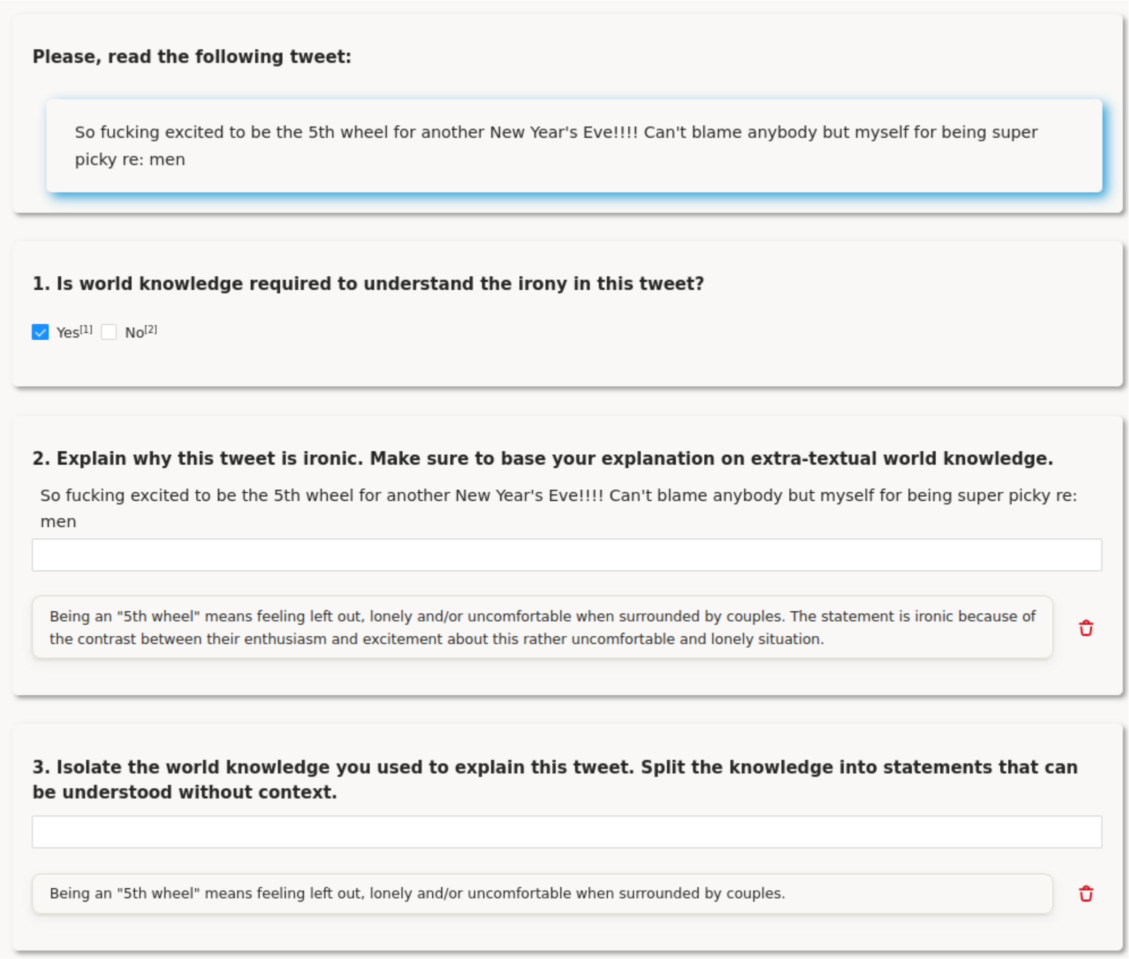
Extracted knowledge: 1. When a train is does not leave on time and is delayed, passengers expect to know why. 2. When their expectations are not met, customers are not satisfied.

Example 5 *Knowledge Extraction 3*

Tweet: Nose is stuffed. Awesome.

Explanation: A stuffed nose is a common symptom of a cold. A stuffed nose can be annoying because it makes it hard to breathe through your nose. This is especially annoying when you try to sleep. Therefore, a positive evaluation is likely ironic.

Extracted knowledge: 1.A stuffed nose is a common symptom of a cold. 2. Breathing through a stuffed nose is uncomfortable while trying to sleep.



Please, read the following tweet:

So fucking excited to be the 5th wheel for another New Year's Eve!!!! Can't blame anybody but myself for being super picky re: men

1. Is world knowledge required to understand the irony in this tweet?

☑ Yes[1] □ No[2]

2. Explain why this tweet is ironic. Make sure to base your explanation on extra-textual world knowledge.

So fucking excited to be the 5th wheel for another New Year's Eve!!!! Can't blame anybody but myself for being super picky re: men

Being an "5th wheel" means feeling left out, lonely and/or uncomfortable when surrounded by couples. The statement is ironic because of the contrast between their enthusiasm and excitement about this rather uncomfortable and lonely situation.

3. Isolate the world knowledge you used to explain this tweet. Split the knowledge into statements that can be understood without context.

Being an "5th wheel" means feeling left out, lonely and/or uncomfortable when surrounded by couples.

Figure 1: Example of annotating irony explanations in Label Studio.

Chapter 3

Explanation Evaluation

In related work, explanations for irony or sarcasm have been evaluated along the dimensions **adequacy** and **fluency**. Adequacy assesses whether the content is accurate and appropriate, while fluency evaluates the grammatical quality and overall readability of the explanation (Desai, Chakraborty, and Akhtar, 2022; Saakyan et al., 2025).

Building on the work of Desai et al. (2022), the irony explanations are evaluated along three dimensions: **(1) adequacy**, **(2) human-likeness** and **(3) overall ranking**, where annotators rank the explanations from best to worst based on personal preference. All evaluation annotations were conducted independently by multiple annotators using the Label Studio platform (Tkachenko et al., 2020-2022). A screenshot of the annotation tool for explanation evaluation illustrated with Figure 2 (evaluation of adequacy and human-likeness) and Figure 3 (ranking evaluation).

Adequacy Adequacy evaluates whether the explanation is **accurate, relevant and effectively clarifies** the irony. Explanations should ideally highlight the violated communication expectations and the contradictions that need to be identified to understand the ironic intent. If the explanation is based on incorrect or irrelevant information, it should be rated lower in adequacy. Nevertheless, annotators are encouraged to apply their own judgment and may apply additional criteria based on their intuition.

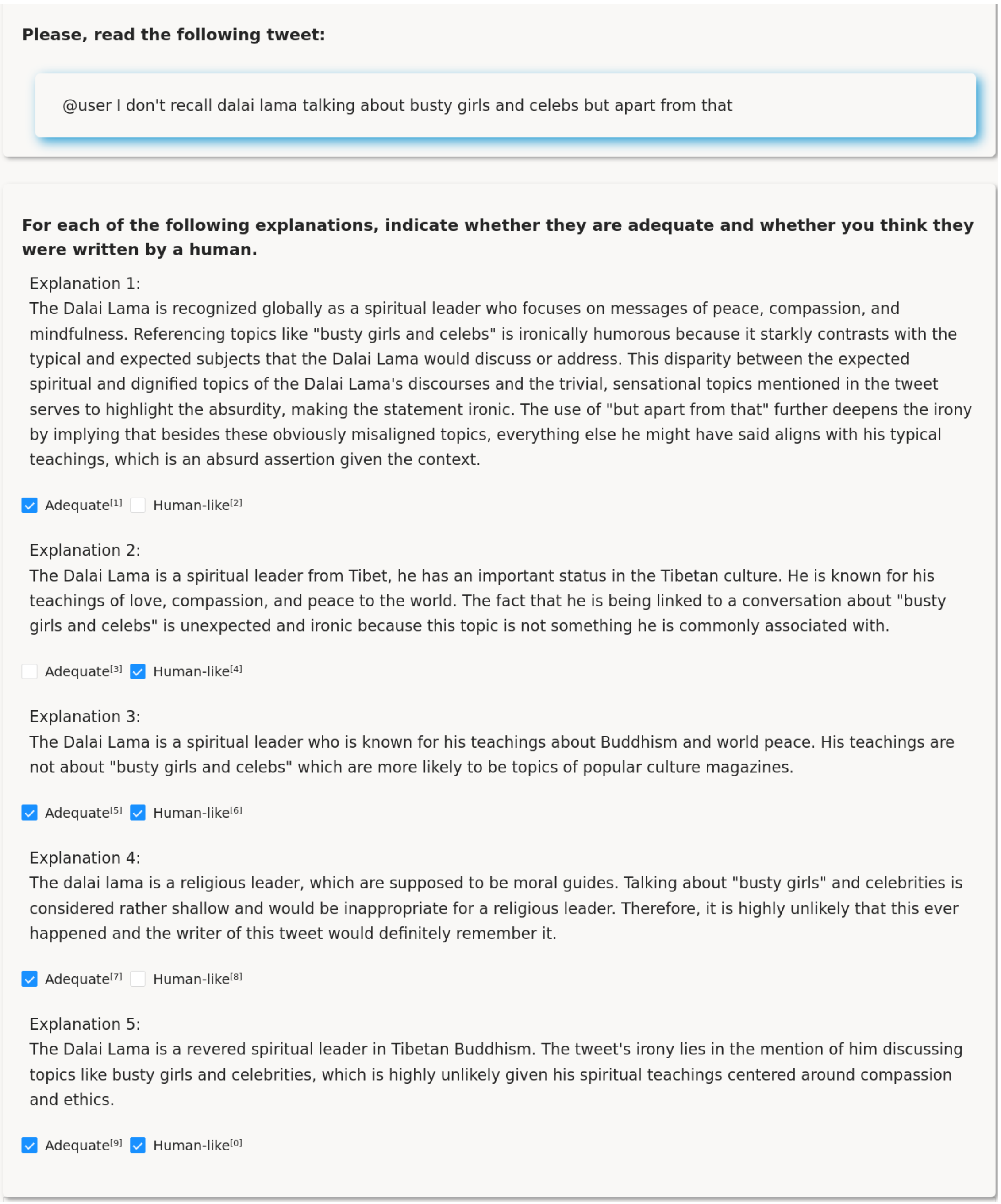
Human-likeness With human-likeness, annotators are asked to assess which explanations appear to be written by humans. This category primarily concerns **fluency**, but also encompasses issues like excessively long outputs and strange phrasings that are technically correct, but may feel unnatural or unintuitive.

While we do not disclose the exact number of human-written explanations in the dataset, annotators are informed that a mix of human and machine-generated explanations is presented to them, with at least one human-written explanation. Annotators are encouraged to apply their own judgment in identifying human-like explanations, as no specific criteria were provided. Interestingly, after completing the task, several annotators remarked that the longest explanations were often easy to identify as machine-generated, specifically by GPT-4-Turbo, because its level of elaboration was far beyond what a human would typically write.

Explanation ranking For explanation ranking, annotators are given the same freedom to establish their own criteria as in the human-likeness evaluation. They are encouraged to apply their

personal judgment in ranking the explanations, without relying on any predefined guidelines. Several annotators expressed uncertainty about whether to prioritize human-like explanations over machine-generated ones. In response, we clarified that "human-likeness" and overall quality should be treated as **separate considerations** from our perspective. The inclusion of this criterion was entirely up to their discretion, but it was not a required criterion for ranking.

Annotation-level interactions Annotators are instructed to treat each **evaluation dimension** –adequacy, human-likeness and overall ranking – as a **separate task**. An explanation may be factually correct and appear human-written, it may still fail to explain the irony adequately. In other words, adequacy is not a prerequisite for human-likeness, and vice versa: an explanation may seem machine-generated, but still be accurate and appropriate. Importantly, for the ranking task, annotators are not required to prioritize explanations that were either adequate or human-like. Rankings are based on personal preference, allowing for **flexibility** in how individual criteria are weighed.



Please, read the following tweet:

@user I don't recall dalai lama talking about busty girls and celebs but apart from that

For each of the following explanations, indicate whether they are adequate and whether you think they were written by a human.

Explanation 1:
The Dalai Lama is recognized globally as a spiritual leader who focuses on messages of peace, compassion, and mindfulness. Referencing topics like "busty girls and celebs" is ironically humorous because it starkly contrasts with the typical and expected subjects that the Dalai Lama would discuss or address. This disparity between the expected spiritual and dignified topics of the Dalai Lama's discourses and the trivial, sensational topics mentioned in the tweet serves to highlight the absurdity, making the statement ironic. The use of "but apart from that" further deepens the irony by implying that besides these obviously misaligned topics, everything else he might have said aligns with his typical teachings, which is an absurd assertion given the context.

☑ Adequate[1] □ Human-like[2]

Explanation 2:
The Dalai Lama is a spiritual leader from Tibet, he has an important status in the Tibetan culture. He is known for his teachings of love, compassion, and peace to the world. The fact that he is being linked to a conversation about "busty girls and celebs" is unexpected and ironic because this topic is not something he is commonly associated with.

□ Adequate[3] ☑ Human-like[4]

Explanation 3:
The Dalai Lama is a spiritual leader who is known for his teachings about Buddhism and world peace. His teachings are not about "busty girls and celebs" which are more likely to be topics of popular culture magazines.

☑ Adequate[5] ☑ Human-like[6]

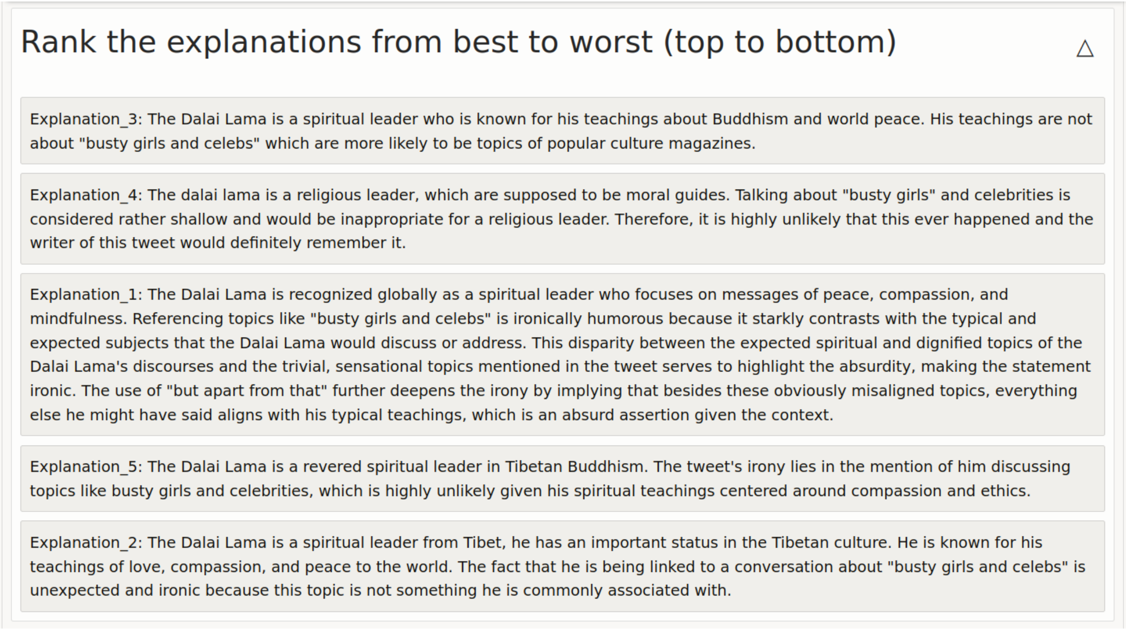
Explanation 4:
The dalai lama is a religious leader, which are supposed to be moral guides. Talking about "busty girls" and celebrities is considered rather shallow and would be inappropriate for a religious leader. Therefore, it is highly unlikely that this ever happened and the writer of this tweet would definitely remember it.

☑ Adequate[7] □ Human-like[8]

Explanation 5:
The Dalai Lama is a revered spiritual leader in Tibetan Buddhism. The tweet's irony lies in the mention of him discussing topics like busty girls and celebrities, which is highly unlikely given his spiritual teachings centered around compassion and ethics.

☑ Adequate[9] ☑ Human-like[0]

Figure 2: Example of the human-likeness annotation (before ranking) in Label Studio.



Rank the explanations from best to worst (top to bottom)

Explanation_3: The Dalai Lama is a spiritual leader who is known for his teachings about Buddhism and world peace. His teachings are not about "busty girls and celebs" which are more likely to be topics of popular culture magazines.

Explanation_4: The dalai lama is a religious leader, which are supposed to be moral guides. Talking about "busty girls" and celebrities is considered rather shallow and would be inappropriate for a religious leader. Therefore, it is highly unlikely that this ever happened and the writer of this tweet would definitely remember it.

Explanation_1: The Dalai Lama is recognized globally as a spiritual leader who focuses on messages of peace, compassion, and mindfulness. Referencing topics like "busty girls and celebs" is ironically humorous because it starkly contrasts with the typical and expected subjects that the Dalai Lama would discuss or address. This disparity between the expected spiritual and dignified topics of the Dalai Lama's discourses and the trivial, sensational topics mentioned in the tweet serves to highlight the absurdity, making the statement ironic. The use of "but apart from that" further deepens the irony by implying that besides these obviously misaligned topics, everything else he might have said aligns with his typical teachings, which is an absurd assertion given the context.

Explanation_5: The Dalai Lama is a revered spiritual leader in Tibetan Buddhism. The tweet's irony lies in the mention of him discussing topics like busty girls and celebrities, which is highly unlikely given his spiritual teachings centered around compassion and ethics.

Explanation_2: The Dalai Lama is a spiritual leader from Tibet, he has an important status in the Tibetan culture. He is known for his teachings of love, compassion, and peace to the world. The fact that he is being linked to a conversation about "busty girls and celebs" is unexpected and ironic because this topic is not something he is commonly associated with.

Figure 3: Example of the ranking annotation (after indicating human-likeness) in Label Studio.

Chapter 4

Data Statement

Dataset Title: **iRONNIE** *woRld knOwledge eNhaNced Irony Explanations)*
Dataset Curator: Aaron Maladry (`aaron.maladry@ugent.be`)

Data Statement Authors: Aaron Maladry, Cynthia Van Hee
Data Statement Version: v1, May 2025
The data is available on Hugging Face as:
Amala3/iRONNIE_train
Amala3/iRONNIE_EN_evalset
Amala3/iRONNIE_NL_evalset
Amala3/iRONNIE_Knowledge

1. ABOUT THIS DOCUMENT:

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 2 Schema. The template was prepared by Angelina McMillan-Major, Emily M. Bender, and Batya Friedman and can be found at: `http://techpolicylab.uw.edu/data-statements`.

2. CURATION RATIONALE

The importance of world knowledge in understanding irony is widely recognized. However, this importance is not reflected in current datasets for irony (and figurative language) explanations. To address this, we present the iRONNIE corpus, a dataset dedicated to explaining irony based on world knowledge in English and Dutch. The iRONNIE corpus contains two main components:

- a corpus of human-written, knowledge-based explanations to be used for training and evaluation of automatic systems. For English, the gold standard component consists of 130 tweets (50 train and 80 test), annotated by two separate annotators. For Dutch, the gold standard consists of 395 tweets (276 train, 119 test,) with explanations of a single annotator.
- a large-scale corpus of irony explanations that are automatically generated for 1,732 English ironic tweets and for 2,083 Dutch ironic tweets.

3. DOCUMENTATION FOR SOURCE DATASETS

The iRONNIE corpus is a collection of social media texts that originate from two existing benchmark datasets, SemEval 2018 Task 3 (Van Hee, Lefever, and Hoste, 2018) for English and a Dutch twitter corpus for irony (Van Hee, Lefever, and Hoste, 2016b). These corpora only contain the ironic tweets from the original dataset. Both irony datasets were collected using the Twitter API searching for irony-related hashtags (#irony/#ironie, #sarcasm/#sarcasme, #not). These explicit irony markers were removed from the text and all tweets were anonymized, replacing urls, images and usernames by placeholders.

4. LANGUAGE VARIETIES

The iRONNIE dataset includes two language varieties, each annotated using the appropriate BCP-47 tag: `en`, and `nl`. These correspond respectively to English, and Dutch as used in general online communication. While the dataset does not target specific regional dialects, it reflects a range of registers commonly found on social media platforms, such as Twitter and Reddit.

`en`: English, primarily international varieties used in online discourse, including contributions from speakers in the United States.

`nl`: Dutch as used in the Netherlands and Flanders (Belgium), with spelling and lexical choices representative of standard written Dutch, but also including informal expressions common in online environments.

Each language variety reflects naturally occurring user-generated content and may include features from regional or informal registers. Although the tweets in the dataset are in Dutch from both Flanders and the Netherlands, the explanations are only written and evaluated by Native speakers of Flemish variety.

6. ANNOTATOR DEMOGRAPHICS

For this project, two separate sets of annotators were employed: one set of annotators wrote the irony explanations and another set of annotators evaluated the explanations. This decision was made because we assume that annotators could not evaluate their own explanations objectively.

6a. ANNOTATOR DEMOGRAPHICS: Explanation Annotation

For English, the annotators include a male PhD student (the PI of this work) and a female master student in linguistics aged between 25-30. Their native languages are Flemish Dutch and Arabic respectively with advanced grasp on academic English. The annotators are between the age range of 25-30.

For Dutch, both explanation annotators are native speakers of Flemish Dutch, who hold a master degree in linguistics. The annotated are between the age range of 24-26. Both annotators for Dutch are male.

6b. ANNOTATOR DEMOGRAPHICS: Explanation Evaluation

For English, explanations are evaluated by fifteen PhD researchers and post-doctoral researchers from the LT3 research group at Ghent University. The native language of most annotators is Flemish Dutch, but other native languages include Italian, German and Chinese. Annotators are between the age range of 23 and 35 with a majority female gender distribution.

For Dutch, the explanations are evaluated by five PhD students with expertise in NLP (including the PI), and one master student. All annotators are native speakers of Flemish Dutch aged between 24 and 30, with a balanced gender distribution.

7. SPEECH SITUATION AND TEXT CHARACTERISTICS

The **tweets in the dataset** includes linguistic data published between 2016 and 2018. The texts were entirely collected from Twitter, and therefore exhibit many typical features of user-generated content. These include pictograms (e.g., emojis), meta-tokens such as hashtags and user handles, and a wide range of non-canonical linguistic phenomena including slang, abbreviations, neologisms, unconventional spellings, punctuation for emphasis (e.g., repeated exclamation marks or ellipses), and informal syntax. Posts often reflect spontaneous, informal communication styles, and many contain irony expressed through multimodal or context-dependent cues.

The explanations in the dataset are written in 2023-2024 by trained linguists with the goal to make world knowledge that is not present in the text explicit. Based on the guidelines, explanations usually start with explicitation and description of world knowledge (facts, assumptions and experiences), followed by reasoning steps that connect this knowledge to a literal expression in the tweet. The style of these explanations is objective and academic. Explanations often contain assumptions and opinions but convey these in a descriptive rather than expressive manner.

8. DATA FORMATTING

The data instances in the iRONNIE dataset have been organized in comma-separated value (CSV) format and encoded using UTF-8 to ensure compatibility and ease of use across systems. For privacy and consistency, user handles and URLs within tweets were anonymized by replacing them with standardized placeholders (@user and [URL], respectively).

9. RECORDING QUALITY

N/A

References

[Cignarella et al.2018] Cignarella, Alessandra Teresa, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *CEUR Workshop Proceedings*, volume 2263, pages 1–6. CEUR-WS.

[DeepSeek-AI et al.2025] DeepSeek-AI, Aixin Liu, Bei Feng, and et al. 2025. Deepseek-v3 technical report.

[Desai, Chakraborty, and Akhtar2022] Desai, Poorav, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

[Farha et al.2022] Farha, Ibrahim Abu, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814.

[Grattafiori, Dubey, and et al.2024] Grattafiori, Aaron, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models.

[Grice1975] Grice, H. P., 1975. *Logic and Conversation*, pages 41–58. Brill, Leiden, The Netherlands.

[Hao and Veale2009] Hao, Yanfen and Tony Veale. 2009. Support structures for linguistic creativity: A computational analysis of creative irony in similes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

[Kreuz et al.1999] Kreuz, Roger J, Max A Kassler, Lori Coppenrath, and Bonnie McLain Allen. 1999. Tag questions and common ground effects in the perception of verbal irony. *Journal of Pragmatics*, 31(12):1685–1700.

[Kreuz and Roberts1995] Kreuz, Roger J and Richard M Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and symbol*, 10(1):21–31.

[Lewis1969] Lewis, David Kellogg. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.

[Riloff et al.2013] Riloff, Ellen, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

[Saakyan et al.2025] Saakyan, Arkadiy, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment.

[Schiffer1972] Schiffer, Stephen R. 1972. Meaning.

[Sperber and Wilson1986] Sperber, Dan and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

[Tkachenko et al.2020-2022] Tkachenko, Maxim, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

[Van Hee, Lefever, and Hoste2016a] Van Hee, Cynthia, Els Lefever, and V'eronique Hoste. 2016a. Exploring the realization of irony in twitter data. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1795—-1799, Paris, France, may. European Language Resources Association (ELRA).

[Van Hee, Lefever, and Hoste2016b] Van Hee, Cynthia, Els Lefever, and Véronique Hoste. 2016b. Exploring the realization of irony in twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1794–1799.

[Van Hee, Lefever, and Hoste2018] Van Hee, Cynthia, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana, June. Association for Computational Linguistics.

[Wilson and Sperber2012] Wilson, Deirdre and Dan Sperber. 2012. Explaining irony. *Meaning and relevance*, pages 123–145.