Exploring the Transfer of Irony Explanation Generation from English to Dutch

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

Language and Translation Technology Team (LT3), Ghent University

firstname@lastname.ugent.be

Abstract

Explanation generation has gained increasing attention in the field of NLP because it makes the output of classification models more intuitively understandable for humans. This is particularly relevant for complex semantic tasks such as irony detection, where there may not be any explicit linguistic markers. Generative models have shown great potential for irony explanation in earlier work, but most studies have been limited to English. Since this is the highest-resourced language, these capabilities may not be available in languages other than English. To address this gap, this paper analyses the performance of generative models for explanation generation in Dutch, a lower-resourced but closely related language to English. Our work shows that larger proprietary models, like GPT-4, can generate meaningful explanations based on relevant world knowledge, whereas smaller open-source models still struggle to perform this task. Besides quality evaluation, we also analyse the limitations of these models, showing that GPT models struggle most with verbosity and that both open-source and proprietary models exhibit circular reasoning ("this text is ironic because the person expresses this in an ironic way"). Finally, open-source models struggle in particular for Dutch because they fail to produce the relevant world knowledge that is required to understand the irony. All models and data used for the experiments is available at iRONNIE on Hugging Face.

1. Motivation & Related Work

In everyday conversation, we usually assume that people mean what they say. Still, figurative language lets us play with this expectation in creative ways, allowing us to express ideas that go beyond the literal meaning of our words. Metaphors, for example, let us exaggerate or reshape reality: when we say that someone is "as big as a mountain," the statement expresses a comparison built on a shared characteristic. Irony, another form of figurative language, instead presents the listener with an expression that clashes with reality or with social expectations. For example, when a researcher remarks, "I sure hope my paper will be rejected," most listeners recognize that it is highly unlikely that this statement should be interpreted literally. With verbal irony, people convey a meaning that contradicts or significantly differs from the literal interpretation (Sperber and Wilson, 1986). To understand verbal irony, and to identify which expressions should be interpreted literally, people rely on world knowledge, like knowing that people who submit an article for publication in an NLP venue may spend months coding, analysing and writing it, and would usually hope for their efforts to be rewarded. This type of common or shared background knowledge is often not explicitly mentioned in the expression itself but is grounded in real-world facts, cultural conventions, and social beliefs (Lewis, 1969; Schiffer, 1972).

These linguistic insights into irony highlight two main factors that make the automatic processing of this type of language particularly challenging. First, most language is intended to be understood literally, which biases computational systems toward literal interpretation. Second, these systems often process sentences in isolation, even though understanding irony requires access to relevant background knowledge that is connected to the input but not explicitly stated. Together, these factors make the modelling of irony a demanding task. Nevertheless, irony is an important topic of research in its own right and is also highly relevant to several high-impact NLP applications, including emotion detection, aspect-based sentiment analysis, and the moderation of harmful content. Ironic expressions can obscure a speaker's true intent, conveying negative or harmful opinions while maintaining plausible deniability (Brown and Levinson, 1987), which makes them difficult to detect automatically. When used to express hostility or prejudice, irony further complicates the detection of hate speech (Frenda et al., 2023).

Within the NLP community, irony has mostly been approached as a classification task, i.e., to detect whether a text is ironic or not (Wallace et al., 2014; Van Hee et al., 2018). The detection of irony has therefore become a popular topic of research, covering not only English, but also a broad range of languages, including Dutch (Van Hee et al., 2016; Maladry et al., 2024), Italian (Cignarella et al., 2018), Chinese (Xiang et al., 2020) and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6795–6807
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Arabic (Farha and Magdy, 2020). In addition, these efforts have extended to multiple modalities, analysing irony in memes (Kumari et al., 2024) and videos (Băroiu and Trăușan-Matu, 2023). With recent advances in generative AI, modelling of irony is shifting from detection-only approaches toward deeper analysis through the inclusion of explanation generation. Here, we note that explanation generation has the primary goal to provide an intuitive justification for a label that is meaningful for human interpretation. However, the produced explanation may not necessarily be faithful to the model's internal reasoning (Atanasova et al., 2023). Therefore, explanation generation evaluates the quality of the produced explanations, but not whether they align with the model's internal "reasoning" process. In some cases, explanations for irony are conceptualized as paraphrases of the intended meaning (Saakyan et al., 2025). More linguistically-grounded approaches, however, recognize that the ironic contrast should be made explicit and that generating an adequate explanation often requires some form of reasoning or world knowledge (Yi et al., 2025). Comparing generated explanations to human-written references, recent work has shown that generative models can produce high-quality explanations, and that explanations from fine-tuned models can even be indistinguishable from their human counterparts (Maladry et al., under review). This work also found that the produced explanations contain relevant world knowledge that was not explicitly mentioned in the text, and that a separate set of generative models can be used to automatically extract this world knowledge from the explanations. While these results are promising, current work on irony explanation generation remains limited to English. As a result, existing evaluations only describe the upper bound of model capabilities for the language with the most resources, leaving open the question of whether these capabilities transfer to less-resourced languages.

In this paper, we explore whether (1) generative models for irony explanation can also produce high-quality explanations that contain relevant world knowledge for Dutch. In addition, we evaluate (2) whether generative models can also be used to automatically extract this world knowledge for Dutch, and (3) whether automatic evaluation metrics align with human evaluation of generated explanations. Besides evaluating the quality of generated explanations, this paper (4) also includes an exploratory analysis of the world knowledge that is generated for irony explanation, and (5) a manual error analysis to identify the limitations of current models for both English and Dutch.

2. Data, Annotation and Evaluation

To assess whether generative models can explain irony in languages other than English, we built a Dutch dataset on the sample principles for data collection and annotation as the English study (Maladry et al., under review). We sampled tweets from a Dutch dataset for irony detection (Van Hee et al., 2016) and annotated 395 Dutch ironic tweets with human-authored explanations. Since genuine tweets should be interpreted literally, explaining them becomes extremely trivial and obsolete. Therefore, this work targets only the ironic tweets in the dataset.

Each of these 395 ironic tweets is annotated with a human-written explanation that first presents the relevant world knowledge needed to understand the text and then describes how this world knowledge contradicts the literal statement or reveals a violation of social conventions and cultural expectations. This first annotation therefore includes both the underlying world knowledge and the explicit contrast. A second annotation is then conducted to isolate the relevant world knowledge from the discussion, as illustrated in Example 2.1.

Example 2.1 `Ironic tweet`*: Had an insomnia cookie and now I can't sleep*
`Explanation`*: The tweet is ironic because an insomnia cookie is not a cookie that causes insomnia (not being able to sleep). Cookies do not normally cause insomnia. The name refers to the bakery being open all night.*
`Background knowledge`*:*

1. *Insomnia cookies is a bakery that is open all night.*
2. *Cookies do not normally cause insomnia*

In this setup, the first annotation is used for explanation generation, while the second serves to extract world knowledge from text-explanation pairs. Notably, this approach does not attempt to enumerate all possible world knowledge associated with every concept in the text, as doing so would result in an unnecessarily large collection of knowledge items. Instead, we focus on generating contextually relevant world knowledge that contributes directly to understanding the irony. Details about the annotation guidelines, which align with the earlier study for English, along with a corresponding data statement following the principles proposed by Bender and Friedman (2018), are provided in a technical report (Maladry et al., 2025).

To evaluate explanation quality, we make use of three established criteria (Desai et al., 2022; Saakyan et al., 2025): (1) **Adequacy**, (2) **Human-likeness** and (3) **Relative comparison**. **Adequacy** evaluates whether the explanation successfully explains the irony and whether the produced world knowledge is correct and relevant to the explanation. **Human-likeness** describes whether the flow of the text is natural, covering aspects like grammaticality, fluency and whether the produced knowledge is not too generic or extensive. Both adequacy and human-likeness are annotated on a 1-5 Likert scale. The third category involves **relative comparison** between the different explanations. We present all AI-generated explanations together with the human-written explanation, and ask our annotators to **rank them all from best to worst**. Practically, this ranking is converted into a scoring system in which each explanation receives a score between 0 and 4 depending on its position in the ranking.

For the experiments, the dataset is split into 276 training samples (70%) and 119 test samples (30%), which will be made publicly available on Hugging Face along with fine-tuned models as the iRONNIE collection. Model evaluation is performed on the test set of 119 tweets and each sample is evaluated 5 times. Since the evaluation of generated output constitutes an annotation task in itself, we include additional guidelines in a technical report (Maladry et al., 2025) and assess the consistency of the evaluation with inter-annotator agreement. Using linear weighting, the results show Krippendorff's α scores of 70% for adequacy, 61% for human-likeness, and 73% for ranking, meaning that adequacy and ranking exhibit high agreement and that human-likeness is more subjective but still reliable.

3. Experimental Setup

This study evaluates both Llama 3 (Dubey, 2024) and GPT-4-Turbo to include open-source and closed-source models and building on the methodology used in the English study. Open-source models are fine-tuned with QLoRA (Hu et al., 2021; Dettmers et al., 2023) and a single in-context example to illustrate the expected output format and goal of the task. Fine-tuning was performed over 10 epochs, using the following parameters: 128 adapter ranks, an alpha value of 16, with an effective batch size of 4, a learning rate of 5e-5, and AdamW as the optimiser. Proprietary models are not fine-tuned but are provided with three in-context examples. Both proprietary and open-source models perform explanation generation and knowledge extraction as two separate tasks (prompts available in Appendix A and Appendix B).

The experimental setup involves three types of fine-tuning. The first type, **monolingual fine-tuning** trains only on Dutch data. In addition, we explore cross-lingual transfer using English resources. The second type, **mixed fine-tuning**, combines English and Dutch training samples into a single, shuffled dataset. This approach enables the model to learn from both languages simultaneously, treating the task as inherently multilingual. The third type, **sequential fine-tuning**, adopts a two-stage process. The model is first fine-tuned on English to learn the explanation task, and then further fine-tuned to transfer this skill to Dutch. These strategies yield the following set of explainers for evaluation: (1) Llama 3 fine-tuned on monolingual Dutch data, (2) Llama 3 fine-tuned on a mixed English-Dutch dataset, (3) Llama 3 sequentially fine-tuned in two separate stages, (4) GPT-4 Turbo (in-context examples) and (5) a human standard explainer.

3.1. Results

To analyse the **results for explanation generation**, we calculate the average score for each metric and conduct statistical testing to analyse whether the score differences are statistically significant, as displayed in Table 1.

The results show that GPT-4 Turbo ranks the highest, even being preferred over the human reference. However, its explanations are readily identified as not human-like. Moreover, the scores for adequacy indicate that human explanations are more consistent than GPT's explanations, although the difference is not statistically significant. All three Llama-based models (monolingual, mixed, and sequential) score significantly lower than the human baseline across all three evaluation criteria. While differences between the Llama variants are modest, the sequentially fine-tuned model receives higher adequacy ratings and the monolingual model is rated highest in human-likeness. Notably, neither of the cross-lingual strategies, mixed nor sequential fine-tuning, leads to substantial improvements over the monolingual Dutch setting, despite the inclusion of additional English training data.

A comparison of explanation lengths further contextualizes the evaluation criteria. Human-produced explanations average 61 words, while GPT-4 outputs are considerably longer, averaging 103 words. This length discrepancy partially explains why GPT's outputs are consistently recognized as machine-generated. In contrast, the three Llama models produce explanations averaging 57 (monolingual), 55 (sequential) and 61 (mixed) words, closely matching the human baseline. While the human-like explanation length gives Llama models an edge over GPT in human-likeness, the Llama explanations are still far from indistinguishable, with human-likeness scores of 28%. This means that GPT explanations stand out from human explanations because of their excessive length, whereas Llama explanations stand out due to their content. The Llama models score considerably worse on ranking score and adequacy compared to the other explainers. The same open-source models produced high quality explanations for English that are indistinguishable from human explanations and attain adequacy scores over 50% (Maladry et al.,

	Adequacy			Human-likeness			Ranking Score		
	Avg.	Δ	Sig.	Avg.	Δ	Sig.	Avg.	Δ	Sig.
GPT4	0.68	-0.05	–	0.16	-0.36	***	3.22	+**0.60**	***
→ human	**0.73**	0.00	–	0.52	0.00	–	2.62	0.00	–
Llama3_sequential	0.43	-0.30	***	0.28	-0.24	***	1.56	-1.06	***
Llama3_monolingual	0.39	-0.34	***	0.27	-0.25	***	1.31	-1.31	***
Llama3_mixed	0.38	-0.35	***	0.28	-0.24	***	1.31	-1.31	***

Table 1: Comparison of explanations against the `human` reference across three evaluation metrics: Adequacy, Human-likeness, and Ranking Score. Each metric includes the average score, the mean difference (Δ) from `human`, and the significance level from paired t-tests. The asterisks mark the degrees of significance, with * indicating p<0.05, ** indicating p<0.01, and *** indicating p<0.001. A dash (–) indicates that the score differences are not statistically significant.

under review). This performance gap between English and Dutch highlights the limitations of current open-source models in adapting to complex tasks such as irony explanation in lower-resourced languages. It also underscores the importance of stronger pre-training and more robust multilingual instruction-tuning. Nevertheless, the results demonstrate that larger proprietary models like GPT-4 are capable of generating high-quality explanations in Dutch.

For **knowledge extraction**, Llama 3 produces an average of 3.04 knowledge items per explanation and GPT4 produces an average of 2.14 knowledge items, both exceeding the human reference of 1.28. In terms of length, humans use an average of 18 words to express these knowledge items, compared to 27 words for GPT-4 and a substantially higher 90 words for the monolingually fine-tuned Llama 3, indicating considerable overgeneration.

For a more fine-grained assessment of knowledge extraction performance in Dutch, we conduct further manual evaluation according to two criteria: soft recall, which assesses whether all essential background knowledge from the explanation has been successfully captured, and soft precision, which checks whether any of the extracted items introduce information not supported by the explanation. These two metrics are considered "soft" because they do not require an exact string match, relying instead on semantic similarity. In some cases, the information is more or less present in the explanation, but does not exactly match the desired output. Therefore, we also analyse two common deviations from ideal precision: `content drift`, as shown in Example 3.1 where the extracted item differs slightly in meaning (e.g., due to over-generalization or added specificity), and `reasoning`, as shown in Example 3.2 where the knowledge items contains reasoning to connect the background knowledge to the input text, as opposed to purely background knowledge.

Example 3.1

`Tweet:` *Nit pick at every little thing guys. Keep doing that. That'll show the team you're behind them.*

`Explanation:`

***Nitpicking (pointing out a lot of minor faults) is often perceived as** providing too much negative feedback. This does not help to support a team**, i.e. showing the team you're behind them**.*

`Extracted Knowledge:`

(1) Supporting a team means giving them positive reinforcement and encouragement.

Example 3.2

`Tweet:` *highly inflamed stomach, kinda enjoying this*

`Explanation:`

*Having a highly inflamed stomach is an irritable and unpleasant medical condition. **It is highly improbable that someone would genuinely enjoy such a medical condition**.*

`Extracted Knowledge:`

(1) Having a highly inflamed stomach is an irritable and unpleasant medical condition.

*(2) **People are highly unlikely to enjoy experiencing medical conditions that cause irritation and discomfort**.*

As shown in Table 2, GPT-4 achieves perfect soft recall (100%) and high soft precision (94.39%), indicating that it successfully captures all relevant background knowledge with minimal extraneous content. Llama 3 also performs well in recall (97.44%) but lags behind in precision (73.55%), suggesting that a notable proportion of its extracted items are not grounded in the explanation.

To better understand these precision errors, we examine the rates of `content drift` and `reasoning`. Among Llama 3's extracted items, 23.68% exhibit content drift and 24.56% contain reasoning. By contrast, only 5.94% of GPT-4's extractions show content drift, and 18.18% include reasoning. Notably, we observe that both models exhibit more reasoning errors than content drift, indicating an increased tendency to blend background knowledge with interpretive content. Taken together, these results point to GPT-4 as the more reliable model for knowledge extraction in Dutch. It remains more faithful to the source explanations, producing concise and relevant knowledge items with fewer semantic deviations than Llama 3.

3.2. Semantic Similarity for Automatic Evaluation

Since manual evaluation is resource-intensive and limits scalability, we also assess whether automatic measures can serve as a reliable proxy for human judgments. SemScore (Aynetdinov and Akbik, 2024) is a semantic metric that is widely used for the evaluation of generative models. This evaluation metric is preferred over other metrics because it captures the meaning of entire sentences, as opposed lexical similarity metrics, and because it correlates more with human judgement than other semantic metrics like BERTscore (Zhang et al., 2020) an BLEURT (Sellam et al., 2020) that is reported to correlate with human judgment. In our experiments we use the multilingual sentence embedding model `paraphrase-multilingual-MiniLM-L12-v2` to calculate SemScore.

As shown in Figure 1, which presents **explanation similarities** across two languages and four systems, GPT-4 and the fine-tuned Llama variants achieve nearly identical SemScores (roughly 72%) in Dutch. This finding is surprising given that GPT-4 explanations are preferred by human annotators, while Llama outputs are rated adequate in only 42% of cases and human-like in just 23%. While SemScore aligned well with qualitative evaluations in English (Maladry et al., under review), it appears less effective at differentiating explanation quality in Dutch.

In addition to explanation similarity, we also compute SemScore for **knowledge extraction**, excluding the reasoning component. Here, we compare system- and human-extracted knowledge by concatenating all extracted items into a single string. When ignoring the reasoning component, SemScores reflect human judgement more closely preferring GPT-4 with a score of 86% over monolingually fine-tuned Llama 3 with score of 71%. Altogether, these results suggest that SemScore is only partially reliable across languages and tasks. For Dutch, SemScore only aligns with human judgement for knowledge extraction, but does not align when evaluating the full explanations (including reasoning).

4. Analysis and System Limitations

The models developed for English and Dutch irony explanations and knowledge extraction show strong performance. As a result, they provide a useful basis for further linguistic analyses of irony and the world knowledge involved in understanding it. However, these models should not be used and released without further analysis. To shed some light on the model capabilities and raise awareness about potential limitations, we propose three analyses. First, we explore what kind of knowledge these models can produce in the best-case scenario with the best-performing models. Secondly, we conduct an additional qualitative error analysis on the explanations of the best-performing model, highlighting the remaining issues linguists should be wary for when using this model. As the best-performing setup for **English**, we use GPT-4 for explanation generation, given its superior ranking and adequacy scores, and Llama 3 for knowledge extraction to reduce overgeneration. As the best-performing setup for **Dutch**, we use GPT-4 for both explanation generation and knowledge extraction, since qualitative evaluation favours this model over the fine-tuned Llama 3 models. Finally, we conduct an error analysis on the explanations of Llama 3 for Dutch. We perform this error analysis on the best-performing setup because it is likely to be adopted for its high-quality output and also on open-source models because they may be preferred for sensitive data or when API costs would be too high.

4.1. Knowledge used in Irony Explanation

To better understand what kind of knowledge is incorporated into model-generated explanations at scale, we extract knowledge items from all ironic tweets in both the English corpus (1,732 tweets) and the Dutch corpus (2,083 tweets). This results in 6,644 knowledge items for English and 6,573 for Dutch. Given the size of these datasets, we require an interpretable, high-level overview of the content distribution. To this end, we apply topic clustering using BERTopic (Grootendorst, 2022), leveraging a KeyBERT-inspired topic model. To guide the clustering process toward broadly interpretable and domain-relevant categories such as "Politics", "Economy", "Sports", and "Culture", we initialize the model with a set of zero-shot seed lists. These lists are generated by prompting GPT-4o to produce representative keywords associated with common news categories. For example, the "Economy" cluster is seeded with terms such as ["economy", "inflation", "GDP", "job market", "tax"]. While these seed lists are not intended to be exhaustive, they serve to loosely steer the clustering process toward broad, interpretable categories. This helps

	Recall	Precision	Content drift	Reasoning	Total items
Llama 3	97.44%	73.55%	23.68%	24.56%	155
GPT	100.00%	94.39%	5.94%	18.81%	107

Table 2: Soft recall and soft precision of extracted knowledge items based on manual validation on 90 human-written explanations. Content drift and reasoning refer to knowledge items that are considered correct for soft precision, but exhibit deviation from the ideal desired output.

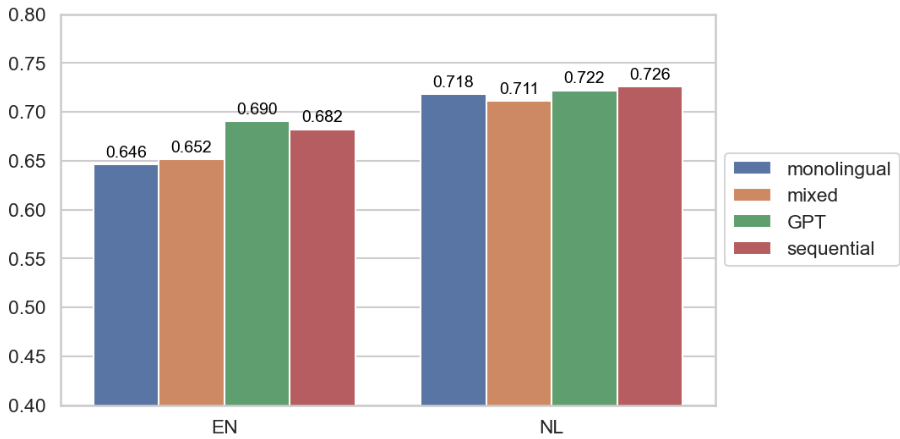

Figure 1: SemScore calculated between human and system explanations.

support a high-level overview of the types of knowledge used in model-generated explanations, while avoiding the fragmented groupings that may result from fully unsupervised topic discovery.

The topic model produced 120 clusters for each language (included in Appendix C). Several clusters in both languages contain explicit reasoning, explaining contradictions between a positive sentiment expressed in the text and a negatively connoted situation. Common themes across both languages include sports, school, and physical health. Many clusters also reflect everyday concerns and inconveniences, such as waking up early, traffic, family life, and the weather. In both corpora, we also observe clusters related to political and societal issues. For English, this includes topics such as racism and discrimination; for Dutch, clusters include themes like war, terrorism, and extremism. These topics reflect prominent public concerns during the time span in which the tweets were written (2012–2017), suggesting that the models can incorporate knowledge relevant to the socio-political context of that period. Additionally, both languages feature knowledge about digital communication norms, including the use of emoticons, emojis, and social media conventions. In addition, specific knowledge about social media conventions, like knowing that a skull emoji is often used in the sense "dying from laughter", is also present for both languages.

To further analyse the specificity of the extracted knowledge, we applied Named Entity Recognition using spaCy (Honnibal et al., 2020). As shown in Figure 2, the extracted knowledge frequently references named entities such as people, locations, and organizations.

In Dutch, the most frequently mentioned individuals include Geert Wilders and Fred Teeven, both known for their roles in Dutch politics. Organizations such as the NS (Dutch Railways), political parties (PvdA, VVD), and news outlets (RTL, NOS) also appear prominently. For English, the most commonly mentioned people are Jesus Christ (often appearing in expressions of surprise or frustration) and Barack Obama. This shows that the model can generate knowledge about well-known entities within the time span of the corpus (between 2014 and 2015). Beyond these more general references, we also find instances of specific, historically grounded knowledge. For English, we found mentions of Eric Garner, whose death became a pivotal moment in the Black Lives Matter movement. In the Dutch corpus, multiple explanations mention Charlie Hebdo, a staple event of Islamic terrorism in Europe. This shows that the model can also access knowledge about specific events that were central to public debate during the 2014–2015 period.

4.2. Error Analysis for Highest-Quality Output

The scores demonstrate what the state of the art can achieve, but limitations remain. To better understand where these limitations lie, we manually

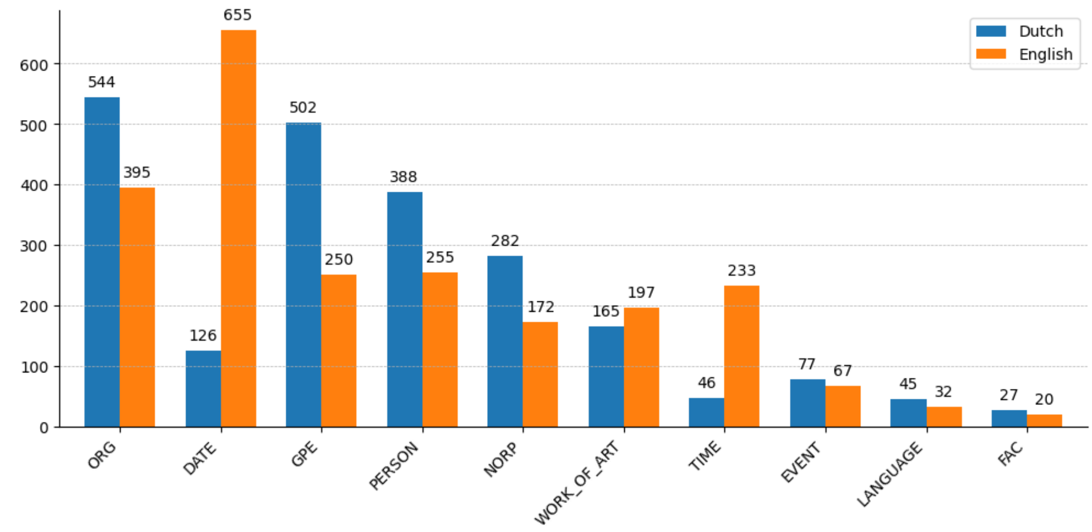

Figure 2: Named entity labels present in Dutch and English extracted knowledge.
NER label clarification: ORG = Organization, GPE = Geo-political entity (e.g., country, city, state), NORP = Nationality, religious, or political group, FAC = Facility (e.g., buildings, airports, highways).

examined the explanations generated by GPT-4, the best-performing explainer, which was generally preferred over human explainers. Our analysis focused on instances where the majority of evaluators deemed the explanation inadequate. For English, this results in a total of 23 explanations (out of the 80 in the test set) and for Dutch, this results in a total of 35 explanations (out of the 119 in the test set). As shown in Table 3, the most frequent issue across both languages is verbosity. Many explanations are unnecessarily long, particularly when the ironic contrast relies on relatively simple world knowledge. To illustrate, Example 4.1 demonstrates how the model over-elaborates on a straightforward contradiction.

	EN (n=23)	NL (n=35)
Verbose	18 (78.26%)	23 (65.71%)
Sarcasm	2 (8.70%)	12 (34.29%)
Reasoning	3 (13.04%)	5 (4.29%)

Table 3: Issues identified in English and Dutch GPT explanations.

While this issue occurs in both languages, overgeneration appears more frequently in English than in Dutch. A second recurring issue involves explanations that describe an expression as being "used sarcastically", "referring ironically", or "strengthening the irony". These statements result in circular reasoning, restating that something is ironic without explaining why. In addition to offering little interpretive value, they also contribute to excessive explanation length (see Example 4.2). Such explicit mentions of irony and sarcasm are more frequent for Dutch than for English.

Finally, the third systematic issue involves flawed reasoning structure. In these cases, the model produces generally accurate background knowledge and successfully highlights a contrast, but places disproportionate emphasis a less relevant component. As shown in Example 4.3, the explanation repeatedly focuses on the individual "Charlie", which weakens the overall coherence of the explanation.

Example 4.1 `Text:` *Love it when your typing at work and a spider comes out from the keyboard and hangs out between your fingers*
`Explanation:` *Typically, encountering a spider unexpectedly, especially one emerging from a personal space like a keyboard, tends to evoke feelings of surprise and often fear or disgust in many people. The notion of enjoying such an experience contradicts the usual reaction of discomfort or alarm associated with spiders. Therefore, the use of "Love it when" to describe an encounter that most people would find unpleasant or startling is an example of irony. The statement suggests affection or pleasure when the anticipated and conventional response would likely be the opposite.*

Example 4.2 `Text:` *cleaning up spew is a definite highlight of my job!!*
`Explanation:` *In typical job environments, tasks like cleaning up spew (vomit) are considered unpleasant and undesirable. Most people do not enjoy dealing with bodily fluids because it can be unsanitary and often has a strong, disagreeable odour. Labelling such a task as a "highlight" of one's job* ***employs heavy sarcasm*** *because it contradicts the common understanding and general attitude towards this type of menial and distasteful work. Therefore, the use of "definite highlight" in this context is meant to be ironic, expressing the speaker's actual dissatisfaction or frustration with this aspect of their job.*

Example 4.3 `Text:` *Love how I came into work at 8 because Charlie said we were busy.... 3 people in 45 minutes,yeah we got this place packed Charlie.*
`Explanation:` *The tweet sarcastically remarks on the situation where the speaker was told by "Charlie" to come in early due to being busy, which implies a heavy workload or many customers. However, the subsequent information about only three people being present in 45 minutes clearly contradicts Charlie's claim, suggesting an extremely slow day instead. The phrase "yeah we got this place packed" is ironic because the situation is the opposite of packed, highlighting the disparity between expectation (set by Charlie) and reality (very few people). The use of term "Love how" is also sarcastic, as it is unlikely that someone would genuinely love being misled about the need to start work early.*

4.3. Open-source Error Analysis for Dutch

To analyse the limitations of open-source models, we focus the evaluation on the irony explanations generated by the Llama models for Dutch, where the performance gap was most notable. Although the differences between sequential fine-tuning (first on English, then on Dutch), monolingual fine-tuning and mixed fine-tuning (on both English and Dutch simultaneously) are minimal, we select the outputs of the sequentially fine-tuned model because it achieved higher average scores for adequacy and was ranked higher in comparative judgement. These are further filtered to only include the explanations for which at least one annotator labelled the explanation as "not adequate", suggesting that the text may not fully meet the minimal requirements.

For this analysis, each explanation was labelled along a variety of categories (see Table 4). This was performed in a multi-label setup, as the different types of errors are not mutually exclusive and can co-occur. The analysis shows that almost 70% of inadequate responses did not contain sufficient, correct and relevant world knowledge, indicating that generating the relevant knowledge remains the main challenge. While the absence of relevant world knowledge undoubtedly causes reasoning issues further down the line, reasoning issues also occurred in 64% of the inadequate explanations, and in 17% even when there were no knowledge issues. Finally, about 60% of the explanation exhibited formal issues, most due to verbosity.

Error Type	Percentage (%)
Knowledge	**68.75**
Too broad	6.25
Irrelevant	22.92
Wrong	27.08
Missing	57.29
Reasoning	**63.54**
Circular	38.54
Conclusion	50.00
Form	**60.42**
Verbose	51.04
Fluency	30.21

Table 4: Error analysis of explanations generated by the sequentially fine-tuned Llama 3 model for Dutch in a multi-label setup. Percentages (including those in sub-categories) refer to the complete set of inadequate explanations.

5. Conclusion

In this study, we assess whether generative models can explain irony using world knowledge in languages other than English. To this end, we transfer the methodology and models that performed well for English irony explanation (Maladry et al., under review) to Dutch, a closely related language with more limited resources, and publish the models and corpus at iRONNIE on Hugging Face.

Our analysis shows that large proprietary models can perform on-par with human explainers for Dutch, but that they do not outperform humans like they do in English. Whereas the gap for irony explanation between English and Dutch is rather small for proprietary models, smaller open-sourced models like Llama 3 exhibit a larger performance gap, both for explanation generation and knowledge extraction. Similarly, automatic metrics for sentence similarity did align with human evaluation for English (Maladry et al., under review), but sentence embeddings for Dutch do not seem to capture sufficient nuance to reflect human judgement for irony explanation. This finding confirms the need for further manual evaluation for generative AI and suggests it may be especially important for languages other than English.

Analysis of the knowledge used for irony explanation in both languages shows that high-performing models express affective information to explicitly highlight the contrast between expected and expressed sentiment, but they also include more factual information. The explanations and extracted knowledge span a range of general world knowledge topics, such as daily life, political parties, and well-known public figures. In some cases, this knowledge extends to time- and place-specific individuals and events that played a major role in public discourse during the 2014–2015 period. Notably, this pattern also holds for Dutch, despite its lower-resource setting. For high-performing proprietary models, we identified issues in the model argumentation and reasoning with outputs being too verbose, using circular reasoning and employing uninformative statements like “used sarcastically”, “referring ironically”, or “strengthening the irony”. These limitations apply to irony explanation in both English and Dutch. Although open-source models also exhibit this behaviour, the explanations are mostly found inadequate because they do not contain sufficient correct and relevant world knowledge.

6. Acknowledgements

This work was supported by Ghent University under grant BOF.24Y.2021.0019.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Limitations

For future work, we believe it would be worthwhile to further explore data augmentation for this task. In the current evaluation setup, we fine-tuned a model using a mixed English and Dutch dataset and explored sequential fine-tuning (first for English, then for Dutch). In addition, it would also be possible to translate the English data to Dutch. However, we did not include this into our experimental setup, as it would have significantly increased the manual workload since all generated explanations are manually evaluated.

This study focused on Llama 3 and GPT 4 as representative models, but other models may yield different results. We performed a manual error analysis for some of these models, but the results may be different for newer versions of the models or different architectures and models.

Future research could also analyse the generated explanations in more detail, aligning them with argumentation theory to provide a more fine-grained classification of explanation types.

Finally, the dataset only contains explanations written by a single human reference. While it would be fascinating to investigate whether humans explain irony in different ways and to analyse these differences, this requires large-scale annotation beyond the scope of this work. Although we performed explanation evaluation with multiple annotators, we did not investigate why annotators disagree about the quality of specific explanations.

7. Bibliographical References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. *ArXiv*, abs/2305.18029.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity.

Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. 2023. Comparison of deep learning models for automatic detection of sarcasm context on the mustard dataset. *Electronics*, 12(3):666.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:26.

Poorav Desai, 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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

et al. Dubey, Abhimanyu. 2024. The llama 3 herd of models.

Ibrahim Abu Farha and Walid Magdy. 2020. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *The 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 32–39. European Language Resources Association (ELRA).

Simona Frenda, Viviana Patti, and Paolo Rosso. 2023. When sarcasm hurts: Irony-aware models for abusive language detection. In *International conference of the cross-language evaluation forum for european languages*, pages 34–47. Springer.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Gitanjali Kumari, Chandranath Adak, and Asif Ekbal. 2024. Mu2sts: a mu ltitask mu ltimodal s arcasm-humor-differential t eacher-s tudent model for sarcastic meme detection. In *European Conference on Information Retrieval*, pages 19–37. Springer.

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

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Véronique Hoste. 2024. The limitations of irony detection in dutch social media. *Language Resources and Evaluation*, 58(4):1355–1386.

Aaron Maladry, Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2025. Guidelines for writing and evaluating explanations for irony in tweets.

Aaron Maladry, Cynthia Van Hee, Els Lefever, and Veronique Hoste. under review. Understanding irony through explanations and background knowledge. In *Proceedings of the 1st Workshop on Computational Affective Science (CAS 2026), co-located with the Language Resources and Evaluation Conference (LREC) 2026*.

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

Stephen Schiffer. 1972. *Meaning*. Clarendon Press, Oxford,.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

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

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. 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.

Cynthia Van Hee, 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. Association for Computational Linguistics.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu, Chu-Ren Huang, et al. 2020. Ciron: a new benchmark dataset for chinese irony detection.

Peiling Yi, Yuhan Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A. Prompt Template for Explanation Generation

GPT-4 Turbo

Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.
Here are some examples of the expected output:

Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.

Example 2 -
Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous

Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.

Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.
Explanation: The stereotypical phishing email from a supposed "nigerian prince" is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).

New sample:
Text: {text placeholder}
Explanation: {system fills in explanation}

Llama 3

[Human Says][BoS]
You are an expert trained in identifying irony and sarcasm in social media text and explaining the underlying reasoning. Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.
Explain why this tweet is ironic:
Text: nit pick at every little thing guys. keep doing that. that'll show the team you're behind them.[EoS]

[System Says:][BoS]
Explanation: nitpicking (pointing out a lot of minor faults) is often perceived as providing too much negative feedback. this does not help to support a team, i.e. showing the team you're behind them.[EoS]

[Human Says:][BoS]
Explain why this tweet is ironic:
Text: {text placeholder} [EoS]

[System Says][BoS]
Explanation: {system fills in explanation}

B. Prompt Templates for Knowledge Extraction

GPT-4o

You will be provided with an ironic tweet along with an explanations for why they are ironic. These explanations are based on extra textual background knowledge. Your task is to extract which knowledge is used in the explanations but which is not present in the text itself. Split up this information into multiple items where this is appropriate. Make sure that the knowledge you provide is mentioned explicitly in the explanation. Here are some examples of the expected output:

Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
Background knowledge:
1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone.

Example 2 -
Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous
Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.
Background knowledge:
1. When people say that someone has “the guts to do something”, it means they are fearless, confident and brave.
2. The cartoon character Scooby Doo is known for his cowardice.

Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.
Explanation: The stereotypical phishing email from a supposed “nigerian prince”is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and

sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).
Background knowledge:
1. The email from a Nigerian prince or heir is a well-known stereotype of phishing.

New sample:
Text: {text placeholder}
Explanation: {explanation placeholder}
Background knowledge: {system fills in knowledge items}

Open-Source Models

[Human Says][BoS]
For each of the following input texts, we have an explanation for why they contain irony. Your task is to identify and extract any background knowledge used for the explanation.
Extract the background knowledge from this text and explanation pair:
Text: Loooovvveeeeeee when my phone gets wiped.
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic. [EoS]

[System Says][BoS]
Background knowledge:
1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone. [EoS]

[Human Says][BoS]
Extract the background knowledge from this text and explanation pair:
Text: {text placeholder}
Explanation: {explanation placeholder} [EoS]

[System Says][BoS]
Background knowledge: {system fills in knowledge items}

C. Topic Modelling

Dutch Topics		English Topics	
Count	Representation	Count	Representation
2,168	[sarcastic, ironic, irony]	2,357	[irony, contradiction, fact, knowledge]
333	[school, pupils, students, education]	121	[nfl, teams, fans, team, football, sports]
204	[hashtags, sarcasm, sarcastic, criticism]	101	[debt, money, pay, taxes, economy]
155	[budgets, income, expenses, financial, subsidy]	93	[Facebook, twitter, social, media]
97	[tweet, twitter, facebook, social, media, instagram]	93	[evaluation, positive, contradiction, negative]
92	[parents, children, upbringing, mother, household, family]	89	[annoys, uninteresting, complaining, frustrating]
84	[news, journalism, media, world news]	85	[exam, study, test, writing, student]
82	[traffic issues, driving behaviour, traffic, drivers]	82	[tweet, twitter, retweet, text, writer]
69	[sarcasm, humorous, irony, funny]	81	[diet, fat, healthy, unhealthy, eat, exercise]
67	[irony, humor, frustration, disappointment, funny]	78	[emoji, smile, indicates, emotion]
67	[symptoms, painful, inconvenience, stomach ache]	70	[irony, situational, sentiments, contradiction]
64	[weekend, vacation, Fridays]	70	[chores, work, lazy, job, office]
61	[food, meal, snacks, diet, lunch, eat]	62	[morning, sleep, 8am, night, tired]
57	[warm, temperature, summer, climate]	59	[police, officer, civilians, violence, crimes]
54	[impolite, rude, negative, unfair, criticism]	58	[weekend, Friday, Saturday, Monday, Sunday]
54	[music taste, pop music, music, festival]	53	[tweets, genuine, positive, evaluation]
53	[absurd, illogical, contrast, contrary]	53	[woken, sleep, asleep, alarm]
53	[sleeplessness, sleep, lack of sleep]	52	[sentiment, evaluation, positive, contradiction]
52	[shoes, clothing, feet, outfit, stylish]	52	[irony, knowledge, reasoning, inconsistency, factual]
47	[war, violence, extremism, terrorist attacks]	41	[discrimination, racism, unfair, stereotype]

Table 5: Top Dutch and English topic clusters with columns for representations and count for number of knowledge items that contain the cluster. Dutch topics are manually translated to English for accessibility.