# Perspectives on Language Model and Human Handling of Written Disfluency and Nonliteral Meaning

**Aida Tarighat**
UKRI CDT in NLP
School of Informatics
University of Edinburgh
tarighat.aida@gmail.com

**Patrick Sturt**
Department of Psychology
School of PPLS
University of Edinburgh
patrick.sturt@ed.ac.uk

**Martin Corley**
Department of Psychology
School of PPLS
University of Edinburgh
martin.corley@ed.ac.uk

## Abstract

When written, disfluencies are intentional. Despite frequently being considered irrelevant noise and consequently excluded from transcriptions and training data of spoken language, disfluencies are now more commonly present in online writing. While humans can process the meanings conveyed by written disfluencies, language models struggle to understand them, mainly due to being trained on filtered data. We test BERTweet's capability to make human-like predictions in fluent and disfluent cases. We find that the model performs better than expected when handling fluent sentences; however, its performance significantly worsens when the context includes a written *um*. We believe that this decline in performance is related to sarcasm. We present two, not wholly successful, reading experiments to test our theory. We suggest that incorporating disfluencies into training data could improve model performance. We invite further comment.

## 1 Introduction

With the advent of easy electronic communication and social media, written language has taken on a more conversational and speech-like quality (e.g., Eisenstein et al., 2014). One aspect of this change is the use of written disfluencies, such as *um*. There is disagreement on whether these tokens are produced deliberately in speech (Clark and Tree, 2002; Corley and Stewart, 2008); however, in written language, they *must* be intentionally produced. This opens up the question of what their *meaning* might be, and whether language models (LMs) and large language models (LLMs) might fail to capture that meaning, and any distinction between spoken and written disfluency.

Although, to date, LMs/LLMs have tended to treat disfluency as noise, there has been growing interest in incorporating both spoken and written disfluencies into models to enhance their performance in applications such as real-time dialogue systems (e.g., Passali et al., 2022), autonomous vehicles (e.g., Large et al., 2017), question answering systems (e.g., Gupta et al., 2021), and stuttering detection (e.g., Al-Banna et al., 2022). However, the main focus of the recent comprehension and detection studies has been on retrieving the literal meaning with regard to the 'disruption' caused by the disfluency. This approach misses the fact that disfluencies could be of potential significance in interpreting nonliteral meanings. Whereas natural language processing (NLP) studies have looked into nonliteral language understanding by focusing on idiom, metaphors, and sarcasm (e.g., D'Arcey et al., 2019; Desai et al., 2021; Hu et al., 2022; Sporleder and Li, 2009), disfluencies remain understudied.

### 1.1 Our Way of Approaching Disfluencies

We previously studied the use of written disfluencies (*um*, *uh*, *hmm*, *erm*, and *er*) on Twitter and found that humans rated tweets containing *um* and *hmm* as slightly more, although not significantly more, sarcastic when fillers were in tweets compared to when the fillers were excised from the same tweets. Humans also considered the tweets containing fillers to be less formal (Tarighat et al., 2022). Therefore, we aimed to investigate the potential role of the written filler *um* in signaling nonliteral meanings using a set of materials to be tested in both LMs and behavioral experiments.

Although written disfluency has not been experimentally investigated to date, a number of studies have focused on the comprehension of spoken hesitations. Fillers such as *um* and *uh* speed up the processing of the word which follows them (Corley and Hartsuiker, 2003; Fox Tree, 2001), and help with the integration of unexpected words into their discourse (Corley et al., 2007). They bias expectations toward new rather than given information (Arnold et al., 2003). Importantly, spoken fillers influence listeners' pragmatic interpre-

*Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.*

| Meaning-Fluency | Item: word-by-word self-paced reading experiment |
|---|---|
| literal-fluent | `Well, blue whales are an endangered species; so I'd say hunting them is a really` **`bad`** `move.` |
| sarcastic-fluent | `Well, blue whales are an endangered species; so I'd say hunting them is a really` **`wise`** `move.` |
| literal-disfluent | `Well, blue whales are an endangered species; so I'd say hunting them is a really um` **`bad`** `move.` |
| sarcastic-disfluent | `Well, blue whales are an endangered species; so I'd say hunting them is a really um` **`wise`** `move.` |
| **Meaning-Fluency** | **Item: masked language modeling task – cloze test – eye-tracking reading experiment** |
| literal-fluent | `Sitting through an hour of sermon would make most children` **`feral`** `on any day. You can ask them.` |
| nonliteral-fluent | `Sitting through an hour of sermon would make most children` **`merry`** `on any day. You can ask them.` |
| literal-disfluent | `Sitting through an hour of sermon would make most children, um,` **`feral`** `on any day. You can ask them.` |
| nonliteral-disfluent | `Sitting through an hour of sermon would make most children, um,` **`merry`** `on any day. You can ask them.` |

Table 1: Examples of the 4 versions of an item used in the four experiments. The target counterparts are in bold: BAD - WISE and FERAL - MERRY. In the second edition of materials, commas were used to enclose *um*. In the SPR experiment, ETR experiment, and cloze test, each participant saw only one version of an item. Target words were not bold in the experiments. In the MLM task and cloze test, the word denoting the literal/nonliteral meaning was masked and replaced by a blank space.

tations, guiding them toward particular meanings (Loy et al., 2017, 2019). For example, Loy et al. (2019) showed that, in a situation where interpreting *some* in its definitional sense as encompassing *all* would cause speakers to lose face ("I ate some cookies"), listeners were more likely to make that interpretation following a disfluency.

Our hypothesis that *written* disfluency might be used to make sarcasm easier to comprehend is related to the Graded Salience hypothesis (Giora, 2003; Giora and Fein, 1999). This hypothesis suggests that humans have difficulty understanding nonliteral meaning because salient (default) meanings have cognitive priority in language comprehension, and accessing an alternative (such as an ironic or sarcastic interpretation) is cognitively effortful. In line with this suggestion, Filik et al. (2014) found N400-like effects and disruptions in eye movements when participants encountered unfamiliar ironies. We hypothesize that the use of *um* in a sarcastic context (in speaking or in writing) signals an interruption of the salient context, making it easier for listeners or readers to access the intended, nonliteral, meaning.

As computers are increasingly being used to communicate with humans, it is important that the nuances of meaning are shared between them, on the surface level at least. Although an LM does not 'understand' disfluency, if it makes different assumptions about how *um* affects the words that are likely to be produced, then it will not communicate effectively. This matters when meaning is nuanced, because achieving human-like performance in LMs increases their ability to better reflect human cognitive processes, and address the complexities of language understanding and generation.

The present study is an investigation inspired by these considerations. We wanted to know how well LMs could handle written disfluencies, whether written disfluencies could signal nonliteral meaning, and whether they could influence the ways in which readers interpret what they are reading.

Our investigation has two parts. First, we compare the performances of an LM trained on informal speech-like data and of humans in predicting nonliteral meanings in the presence of written disfluencies for a set of carefully crafted sentences. Second, we study human behavior in controlled reading experiments using the same set of sentences. Here, we present results from a masked language modeling (MLM) task with BERTweet and a cloze test, conducted to compare meaningful word prediction between the LM and humans. We also report on a self-paced reading (SPR) experiment and an eye-tracking reading (ETR) experiment designed to investigate readers' handling of written disfluency.

## 2 Materials

We made the materials in two rounds. There were 32 items in the first round, 24 of which we used in the SPR experiment. There were 70 items in the second round, 48 of which we used in the MLM, cloze, and ETR experiments. Table 1 shows examples of the items used in the four experiments reported in this paper.

We made 32 grammatically correct speech-like sentences, each with its literal and sarcastic variations (*If you have a butler and a nanny, your life must be* EASY (LITERAL)/HARD (SARCASTIC) *to bear.*). We then recruited 12 L1-English speakers to rate the sentences for sarcastic tone (*How sarcastic do you think the author of this sentence was being?*) on a 7-point Likert scale (*not sarcastic at all*

| Item | BERTweet top word | Cloze top word (count) |
|---|---|---|
| 1. Keep speaking nonsense and people will think you are <mask> /, um, <mask> at some point. I'm telling you. | stupid - fluent<br>stupid - disfluent | stupid (28) - fluent<br>stupid (31) - disfluent |
| 2. Having to listen to people's munching noise when I am trying to eat makes me <mask>/, um, <mask> about my life. It really does. | think - fluent<br>think - disfluent | annoyed (12) - fluent<br>think (17) - disfluent |
| 3. A guy in the audience kept clearing his throat throughout the whole lecture. It was a truly <mask>/, um, <mask> distraction for all of us obviously. | unnecessary - fluent<br>painful - disfluent | annoying (24) - fluent<br>annoying (23) - disfluent |

Table 2: Fluent and disfluent example items used in the MLM and cloze tasks followed by BERTweet's and cloze top word for each fluency version. The number next to the cloze word is the count of it in 80 responses. The critical tokens denoting the literal/nonliteral meanings were removed in the two tasks: 1. STUPID (LITERAL)/BRAINY (NONLITERAL); 2. ANNOYED (LITERAL)/PLEASED (NONLITERAL); 3. DISGUSTING (LITERAL)/DELIGHTFUL (NONLITERAL). In the cloze test, each participant saw only one version of an item. The critical token appeared as a blank space to be filled in with a word.

- *definitely sarcastic*).[1] Each participant was shown only one version of each sentence. We also asked them to provide feedback on interpretability and readability of sentences. We used a sarcastic-literal mean score difference of above 2.7 as a cutoff point. We kept 24 sentences and used them in the SPR experiment (Section 5).

We made changes to the items and increased their number before using them in the MLM task, cloze test (Section 3), and ETR experiment (Section 6). The literal and nonliteral words in the two versions of each item had the same numbers of characters (MERRY/FERAL). We also counterbalanced the literal and nonliteral readings of each word across items (MERRY (LITERAL)/FERAL (NONLITERAL) and FERAL (LITERAL)/MERRY (NONLITERAL)). In the revised materials, we used commas to enclose *um*, to help with readability and increase the salience of the disfluency. Lastly, we added more words after each target word, often in the form of a short second sentence, to minimize gaze regressions out of the target interest area in the ETR experiment. To rate the newly made and edited 60 items for potential sarcastic tone on a 7-point Likert scale online,[2] we recruited 36 neurotypical L1-British-English speakers between the ages of 18 and 34 with no reported reading disorders. We only kept the counterbalanced items with a nonliteral-literal mean score difference above 2. For the items with good scores in only one reading, we repeated the procedure with 10 items rated by 20 other participants. Overall, we kept 48 counterbalanced items for the ETR experiment. There were 4 variations of each of the 48 experimental items based on meaning and fluency (Table 1).

## 3 BERTweet Masked Language Modeling and Cloze Test

We compared the LM and human predictions of meaningful words and how they might be influenced by written disfluencies. We expected that, given a context, the LM would perform better in predicting words in utterances which did not contain *um*.

### 3.1 MLM task

We first ran an MLM task on BERTweet (Nguyen et al., 2020). The tokens denoting the literal and nonliteral meanings were excised. We had 48 fluent items without *um* and 48 disfluent items with *um*, totaling 96 items. The critical tokens assigned to signify literal/nonliteral meanings in the items were masked.

We chose BERTweet due to the presence of fillers such as *um* in its training data and the higher structural similarity between the tweets and the speech-like materials we created for our experiments. We obtained BERTweet's top 10 predictions for 96 materials, using the first eligible predicted word in each list in further analyses (details below in 3.3).

### 3.2 Cloze test

We then conducted a cloze test using the same materials to study the humans' predictions and the possible effect of written disfluencies on their predictions. There were 48 fluent items without *um* and 48 disfluent items with *um*, totaling 96 items. The critical tokens were replaced by a blank space.

For the cloze test, we recruited 160 neurotypical L1-English participants between the ages of 18 and 34 with no reading disorders.[3] We asked them to fill in the blanks using the first word (only a single word without a space or a hyphen) that came to

[1] Informatics Research Ethics Process, RT number 789617.

[2] Informatics Research Ethics Process, RT number 789617.

[3] Informatics Research Ethics Process, RT number 789617.

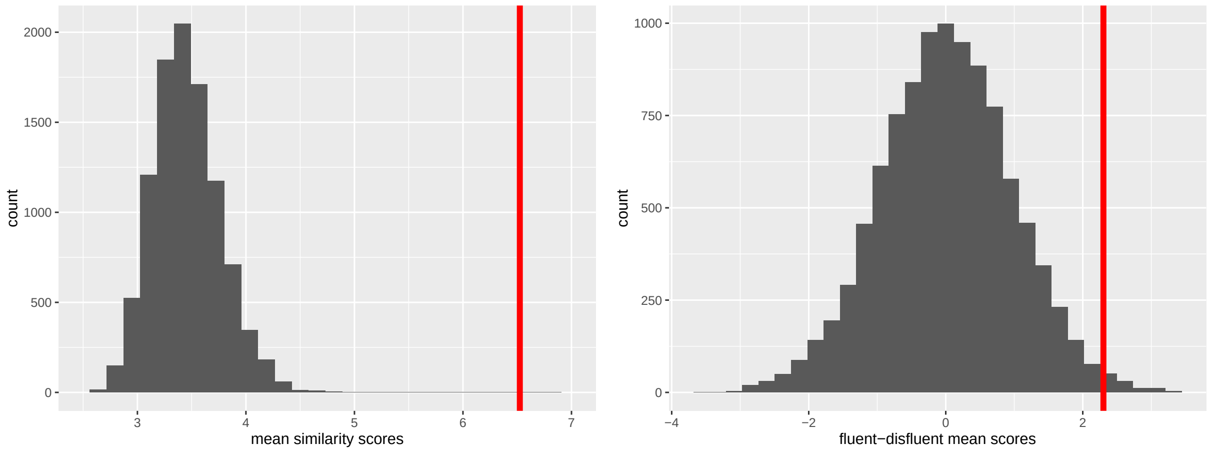

Figure 1: Left: simulated mean similarity scores, with the red vertical line indicating the mean similarity score of 6.52. Right: simulated fluent-disfluent mean scores, with the red vertical line indicating the observed fluent-disfluent difference in similarity score of 2.30. We ran 10,000 permutations of the scores to recalculate the means. BERTweet's continuations were better matches to human continuations following fluent items compared to disfluent items.

mind (autocompletion and autocorrection options were disabled on the participants' devices). Each participant saw only one version (fluent or disfluent) of an item. The participants were remunerated £3.70 for completing 48 items which took about 10 minutes on average.

### 3.3 Analysis

The first step was obtaining the most frequent response for each item in the cloze test to compare with the MLM predictions. However, for six items, there were ties where equal numbers of participants provided two words equally often in either the fluent or disfluent condition. To resolve the ties, we selected the word that was not used in the other condition for the relevant item. If both words were not used in the other condition, the selection was made at random. For the MLM data, we ensured that for each item, we had the most popular predicted word while adhering to the following criteria: no punctuation; no symbols (e.g., @); and no stop words such as “a”, “as”, “and”, “be”, or “not”.

Next, we standardized spellings of the completions to American, and calculated Latent Semantic Analysis (LSA) cosine similarities by making pairwise comparisons using word2vec (Google News, 300 dimensions: University of Colorado). We obtained a similarity score between cloze completions and BERTweet predictions for each item by multiplying the number of identical cloze completions by the BERTweet confidence scores and then by the LSA cosine similarity between words. For example, in one item (*Well, blue whales are an endangered species. So, I'd say hunting them is a really <mask> choice environmentally speaking.*), the most popular cloze completion BAD was chosen by 30 participants, while the highest ranked LM completion was GOOD, which had a confidence rating of 0.333. The word2vec similarity score between BAD and GOOD was 0.719. Therefore, the overall score was $30 \times 0.333 \times 0.719 = 7.183$.

Calculated in this way, the mean similarity score between BERTweet and human cloze completions was 6.52. To assess BERTweet's performance against chance, we permuted the LSA and BERTweet scores 10,000 times, recalculating the mean similarity for each permutation. The visualization of these scores (Figure 1) shows that BERTweet predicted what humans would write significantly better than a baseline of random guessing ($p < .0001$). Table 2 shows fluent and disfluent example items used in the MLM and cloze tasks along with the BERTweet's and cloze top word for each fluency version.

Importantly, we also assessed the effect of fluency, by calculating the difference between mean similarity scores for fluent and disfluent items. The observed difference in similarity scores (fluent-disfluent similarity score = 2.30) was compared to the distribution of mean differences derived from 10,000 permutations of the data (Figure 1). BERTweet's continuations were better matches to human continuations following fluent items compared to disfluent items ($p = 0.0096$).

## 4 Behavioral Experiments

One possible explanation for BERTweet's significantly worse performance in the presence of *um* could be the role of the filler in implying nonliteral, namely sarcastic, meaning. We conducted two reading experiments to test whether written disfluencies could signal nonliteral meaning. We predicted that (1) words compatible with a sarcastic reading of a sentence (*hunting blue whales is a really* WISE *move*) should be easier to read when preceded by *um* (*really* UM WISE *move*) than when not preceded by *um*, and (2) words compatible with a literal reading of a sentence (*hunting blue whales is a really* BAD *move*) might be harder to read when preceded by *um* (*really* UM BAD *move*) than when not preceded by *um*. This leads to the prediction of an interaction between fluency and meaning, with longer reading times and/or more regressions for fluent literal items than disfluent sarcastic ones. To summarize, the disfluency *um* could signal a shift toward a nonliteral or sarcastic interpretation. Example materials for both experiments are in Table 1.

## 5 Experiment 1: Self-paced Reading

We implemented the online word-by-word SPR experiment (Mitchell and Green, 1978) as a moving-window reading task, using jsPsych.[4] Each item had 4 variations based on (a) meaning (whether the critical word was literal or sarcastic in context), and (b) fluency (whether the target word was preceded by *um* or not): literal-fluent, literal-disfluent, sarcastic-fluent, and sarcastic-disfluent (Table 1). In each item, we were interested in reading times for a target word and the following word (for spillover). The target word was a word selected to be [in]consistent with a sarcastic/literal interpretation. We predicted an interaction between fluency and meaning; i.e., disfluency would signal a nonliteral or sarcastic meaning whereas fluency would signal a literal meaning.

### 5.1 Participants and procedure

We recruited 101 L1-English, UK-based, and non-dyslexic participants through Prolific.[5] Participants were remunerated £1.75 for reading 26 items: 2 practice items and 24 experimental items. There were 4 variations of the 24 experimental items based on meaning and fluency (Table 1). Each participant read only one variation of each experimental item, pressing the space bar to reveal each new word of the sentence. Items were selected such that participants read 6 items in each experimental condition. There were 8 attention checks. Experiment settings ensured that the target word was never the last word of the sentence and was followed by at least one word. The experiment took about 10 minutes to complete.

| | target | | target+next | |
|---|---|---|---|---|
| **Meaning-Fluency** | **Mean** | **SD** | **Mean** | **SD** |
| literal-fluent | 381.02 | 184.16 | 809.95 | 351.46 |
| sarcastic-fluent | 388.49 | 204.51 | 869.17 | 544.28 |
| literal-disfluent | 470.19 | 367.25 | 940.54 | 539.39 |
| sarcastic-disfluent | 465.15 | 344.96 | 985.18 | 618.05 |

Table 3: SPR experiment: mean and standard deviation of reading times in milliseconds for target and target+next regions.

### 5.2 Data preparation

We analyzed the reading time data from 99 participants. We removed 2 participants because they got fewer than 6 of 8 attention-check questions correct. Moreover, 1 item was miscoded in the experiment, resulting in 28 missing trials (1.18% of the data).

### 5.3 Analysis

We compared the log-transformed reading times of the target word and of the target word plus the next word (for spillover). Mean and standard deviation of reading times in milliseconds for target and target+next regions are in Table 3. Contrary to our hypotheses, words compatible with the sarcastic interpretation of the sentences were not faster to read when preceded by *um*. Maximally-fitting linear mixed-effects models only showed an effect of fluency, indicating that fluent sentences were faster to read in both literal and sarcastic versions (target, $\beta$ = -0.13, SE = 0.02, $p$ < .001; target+next, $\beta$ = -0.10, SE = 0.01, $p$ < .001). We found no interaction between meaning and fluency (target, $\beta$ = 0.03, SE = 0.03, $p$ = .27; target+next, $\beta$ = 0.02, SE = 0.02, $p$ = .31).

## 6 Experiment 2: Eye-tracking Reading

Whereas the SPR experiment failed to show that written disfluency indexes nonliteral meaning (at least, in the form of sarcasm), it did show that readers were sensitive to written *um*. One possibility is that the artificial segmentation needed for self-paced reading disrupted the rhythm with

[4] https://github.com/UiL-OTS-labs/jspsych-spr-mw

[5] https://prolific.co/

which readers might have read the experimental sentences, reducing any interruption effect that the traditionally spoken element *um* might have had in writing. For that reason, the ETR experiment was a replication of the SPR experiment using an eye-tracking methodology in which natural reading prosody was not disrupted. Our hypotheses were the same: (1) words compatible with a nonliteral reading of a sentence (*hunting blue whales is a really* WISE *move*) should be easier to read when preceded by *um* (*really* UM WISE *move*), and (2) words compatible with a literal reading of a sentence (*hunting blue whales is a really* BAD *move*) might be harder to read when preceded by *um* (*really* UM BAD *move*), and that this would predict longer reading times and/or more regressions for fluent (relative to disfluent) literal items, and vice versa for nonliteral items. Once again, this predicts an interaction between fluency and meaning.

We used Experiment Builder[6] version 2.4.1 to set up the experiment for presentation on an EyeLink 1000 Plus tracker for in-person data collection.

### 6.1 Participants and procedure

We recruited 60 neurotypical L1-English participants between the ages of 18 and 34 with normal/surgically-corrected-to-normal vision and no reported reading disorders.[7] Participants were remunerated £10 for reading 152 items: 2 practice items, 48 experimental items, and 102 filler items. Each participant read only one variation of each experimental item, selected such that they read 12 items in each experimental condition. There were 32 attention checks, 16 for experimental items and 16 for filler items. Experiment settings ensured that the target word was always followed by at least two words before a line break, and that the target word never fell at the beginning of a line and was always preceded by at least two words. The experiment took about 35 minutes to complete, and participants were given breaks after items 50 and 100.

### 6.2 Data preparation

We used Data Viewer[8] to prepare and summarize the eye-tracking data, and did the statistical modeling in R. Since all participants had answered 80% (26) or more of the attention checks correctly, their data was included in the analyses. Data preparation included removing the filler trials, merging nearby fixations, removing fixations less than 80 milliseconds, aligning the fixations vertically within the preassigned interest area bounds, and monitoring the number of horizontally misaligned trials for each participant for removal. If more than 20% (10) of the experimental trials for a participant needed to be removed due to severe horizontal misalignment, that participant's data was excluded from analysis. This left us with 59 participants.

|  | target |  | target+next |  |
|---|---|---|---|---|
| **Meaning-Fluency** | **Mean** | **SD** | **Mean** | **SD** |
| **regression path time** |  |  |  |  |
| literal-fluent | 297.41 | 188.51 | 612.09 | 376.32 |
| nonliteral-fluent | 313.23 | 200.00 | 684.62 | 471.09 |
| literal-disfluent | 299.50 | 211.19 | 629.34 | 408.57 |
| nonliteral-disfluent | 318.34 | 217.30 | 695.69 | 443.70 |
| **first pass time** |  |  |  |  |
| literal-fluent | 235.92 | 109.13 | 472.02 | 186.38 |
| nonliteral-fluent | 237.28 | 107.59 | 490.46 | 202.84 |
| literal-disfluent | 254.70 | 121.28 | 500.14 | 215.25 |
| nonliteral-disfluent | 268.91 | 126.48 | 538.85 | 217.96 |
| **total dwell time** |  |  |  |  |
| literal-fluent | 297.40 | 193.84 | 612.73 | 343.77 |
| nonliteral-fluent | 329.37 | 201.87 | 681.62 | 386.43 |
| literal-disfluent | 323.36 | 195.31 | 652.62 | 358.07 |
| nonliteral-disfluent | 365.96 | 226.34 | 730.76 | 393.75 |

Table 4: ETR experiment: mean and standard deviation of regression path time, first pass time, and total dwell time in milliseconds for the target interest area (target) and the summation of target and next interest areas (target+next).

### 6.3 Analysis

We focused on the target and target+next interest areas and compared the log-transformed reading times for 3 measures: (1) *regression path time* (go-past time) which is the summed fixation duration from when the current interest area is first fixated until the eyes enter a later interest area; (2) *first pass time* which is the sum of the duration of all fixations before the interest area is exited for the first time; and (3) *total dwell time* which is the summed duration of all fixations on the current interest area. Table 4 shows the mean and standard deviation of the 3 measures in milliseconds for the target and target+next interest areas. We also compared the proportions of *first pass regressions out* for the target and next regions; i.e., whether regression(s) were made from the current interest area to the earlier interest area prior to leaving the interest area in a forward direction (Table 5).

Consistent with our prediction of an interaction between fluency and meaning, we expected the

[6] https://www.sr-research.com/experiment-builder/

[7] PPLS Research Ethics Committee, reference number 392-2223/1.

[8] https://www.sr-research.com/data-viewer/

presence of the word *um* to signal a nonliteral or sarcastic meaning, while fluency would signal a literal meaning. However, the results of the ETR experiment did not fully support this hypothesis, as was the case for the SPR experiment. Analyses of regression path time, first pass time, and total dwell time revealed significant effects of both fluency and meaning across the interest areas. Specifically, words signaling literal meanings were consistently read faster than those signaling nonliteral meanings, indicating an overall effect of meaning on reading behavior. Additionally, fluency also influenced reading speed, with target words generally being read faster in the fluent sentences than disfluent ones.

The maximally-fitting linear mixed-effects models of regression path time showed an effect of meaning for the target interest area ($\beta$ = 0.06, SE = 0.02, $p$ = .01), and target+next interest areas ($\beta$ = 0.10, SE = 0.03, $p$ < .001) indicating that literal words were faster to read than nonliteral ones. We found no interaction between meaning and fluency (target, $\beta$ = -0.03, SE = 0.04, $p$ = .35; target+next, $\beta$ = -0.01, SE = 0.04, $p$ = .70)

The maximally-fitting linear mixed-effects models of first pass time showed an effect of fluency in the target interest area ($\beta$ = -0.10, SE = 0.02, $p$ < .001) indicating that fluent sentences were read faster than disfluent ones. For the target+next interest areas, the models showed an effect of fluency ($\beta$ = -0.07, SE = 0.01, $p$ < .001) and one of meaning ($\beta$ = 0.06, SE = 0.02, $p$ = .004) indicating that fluent sentences were read faster than disfluent ones and that literal meanings were read faster than nonliteral ones. However, there was no interaction between meaning and fluency in target+next interest areas (target, $\beta$ = -0.05, SE = 0.03, $p$ = .06; target+next, $\beta$ = -0.04, SE = 0.03, $p$ = .11).

As for total dwell time, the maximally-fitting linear mixed-effects models showed the effects of meaning ($\beta$ = 0.10, SE = 0.03, $p$ < .001) and of fluency ($\beta$ = -0.10, SE = 0.02, $p$ < .001) for the target interest area indicating that literal meanings were read faster than nonliteral ones and that fluent items were read faster than disfluent ones. However, there was no interaction between meaning and fluency in the target interest area. The models also showed the effect of meaning ($\beta$ = 0.11, SE = 0.03, $p$ < .001) and of fluency ($\beta$ = -0.07, SE = 0.02, $p$ < .001) in the target+next interest areas indicating that literal meanings were faster to read than nonliteral ones and fluent items were read faster than disfluent ones. However, there was no interaction between meaning and fluency in the target+next interest areas (target, $\beta$ = -0.03, SE = 0.04, $p$ = .40; target+next, $\beta$ = -0.01, SE = 0.03, $p$ = .74).

| | **target** | **next** |
|---|---|---|
| **Meaning-Fluency** | **Mean** | **Mean** |
| literal-fluent | 0.18 | 0.16 |
| nonliteral-fluent | 0.19 | 0.21 |
| literal-disfluent | 0.08 | 0.16 |
| nonliteral-disfluent | 0.10 | 0.18 |

Table 5: ETR experiment: proportions of first pass regression out, i.e., the regressions that were made from the target and next interest areas to the earlier interest area prior to leaving the interest area in a forward direction.

Lastly, for the proportions of first pass regressions out, the maximally-fitting logistic mixed-effects models only showed an effect of fluency for the target interest area ($\beta$ = 1.01, SE = 0.16, $p$ < .001) indicating that regressions were more likely to be made following a fixation on the target word when the items were fluent. No other effects were reliable, for the target word or the word which followed, and there was no interaction between meaning and fluency (target, $\beta$ = -0.11, SE = 0.25, $p$ = .66; next, $\beta$ = 0.25, SE = 0.22, $p$ = .26).

The results suggest that the effects of fluency and meaning on reading behavior were independent of each other, contrary to our initial prediction of an interaction. However, it is important to note that fluency and meaning each had distinct effects on reading behavior, underscoring the complexity of their influence on comprehension.

## 7 Discussion

We investigated the handling of written disfluencies, which could indicate nonliteral meanings like sarcasm, by an LM and humans. We found that although BERTweet made human-like predictions, its performance was significantly worse when the disfluency *um* was present. Additionally, in our reading experiments, we found that readers were faster to read fluent sentences without *um* and sentences compatible with literal meanings rather than nonliteral or sarcastic ones. We found no interaction between fluency and meaning in the sense that disfluency did not signal a nonliteral or sarcastic meaning and fluency did not signal a literal meaning.

Our results suggest that BERTweet's difficulty with written disfluencies may be due to training on filtered data that excludes disfluencies. The decline in performance, especially in contexts involving sarcasm, highlights the model's limitations in understanding the subtleties of human communication. Previous research has often dismissed disfluencies as irrelevant noise. However, our findings align with more recent studies that recognize the communicative value of disfluencies in online writing. The observed challenges in BERTweet's performance are consistent with other studies that highlight the limitations of LMs in NLP.

## 8 Limitations

Our experiments to date have investigated a specific disfluency in a specific language and context. Our results may have been influenced by the specific design and sample size. Whereas we have established that written disfluencies are worth investigating, with LMs as well as humans sensitive to their presence, this study is just a starting point. To gain a more complete picture, attention should be paid to the naturalness of the stimuli used, and work should be generalized to other languages and disfluencies.

## 9 Future Steps

Future studies should explore more sophisticated methods for integrating disfluencies into LM training. Our next step would involve manipulating the filler placement and removing the commas on the LM to monitor any changes in model behavior. The model could produce different output if disfluency occurred earlier in the sentence and not immediately preceding the masked token, and it would treat *um,* as a very different token from *um*. A later approach could be for us to further pre-train BERTweet using a data set of tweets containing fillers from our previous study, since its performance could potentially be improved. Then, another masked-token prediction task could follow to evaluate the model's improved ability to handle disfluencies.

Another major aspect of future research would be testing disfluencies in an LLM (e.g., Llama) to check differences and potential improvements in performance, which could be the result of the set parameters and/or training data. Since LLMs are different from BERT-type models and are increasingly preferred, it would be important to know if and how they would produce better outputs.

We would also need to compare our findings with other psycholinguistic and computational experiments that focus on licensing nonliteral interpretation. This comparison could identify strengths and weaknesses in current approaches and guide future improvements in human experiments as well as model training and evaluation, especially for developing purpose-built models and data sets for specific tasks. For instance, we know that not all humans understand disfluency in the same way (Li et al., 2022; McKenna et al., 2015), or that nonliteral and sarcastic interpretation is influenced by social and cultural factors (Katz et al., 2004). Therefore, a simple model-training approach might not work when considering how computers should interact with humans.

## 10 Conclusion

Our findings highlight the challenges LMs face in handling disfluencies and probably also in interpreting nonliteral meanings conveyed by disfluencies. Incorporating such elements into training data could improve model performance. Future research should explore more sophisticated methods for integrating disfluencies and other nonliteral indicators into LMs. Additionally, investigating the nuances of sarcasm detection in written text remains a promising area for further study. Well-designed behavioral experiments can capture fine-grained variations in comprehension by focusing on specific psycholinguistic features. Such evidence would be beneficial in evaluating the behaviors of models trained on large, usually written, language corpora. With more information, we can determine how and to what extent to reintroduce disfluencies into data sets.

## Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics, and School of Philosophy, Psychology & Language Sciences. Aida thanks Amr Keleg for his technical advice on the initial steps of the MLM task on BERTweet. Aida also thanks Oli Liu and Coleman Haley for the discussions on the validity of LM/LLM processing techniques.

## References

Abedal-Kareem Al-Banna, Eran Edirisinghe, Hui Fang, and Wael Hadi. 2022. Stuttering disfluency detection using machine learning approaches. *Journal of Information & Knowledge Management*, page 2250020.

Jennifer E Arnold, Maria Fagnano, and Michael K Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, 32:25–36.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Martin Corley and Robert J Hartsuiker. 2003. Hesitation in speech can… um… help a listener understand. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

Martin Corley, Lucy J MacGregor, and David I Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2021. Nice perfume. How long did you marinate in it? Multimodal sarcasm explanation. *arXiv preprint arXiv:2112.04873*.

J Trevor D'Arcey, Shereen Oraby, and Jean E Fox Tree. 2019. Wait signals predict sarcasm in online debates. *Dialogue & Discourse*, 10(2):56–78.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Ruth Filik, Hartmut Leuthold, Katie Wallington, and Jemma Page. 2014. Testing theories of irony processing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):811.

Jean E Fox Tree. 2001. Listeners' uses ofum anduh in speech comprehension. *Memory & cognition*, 29(2):320–326.

Rachel Giora. 2003. *On our mind: Salience, context, and figurative language*. Oxford University Press.

Rachel Giora and Ofer Fein. 1999. Irony comprehension: The graded salience hypothesis. *Humor: International Journal of Humor Research*, 12:425–436.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. *arXiv preprint arXiv:2106.04016*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Albert N Katz, Dawn G Blasko, and Victoria A Kazmerski. 2004. Saying what you don't mean: Social influences on sarcastic language processing. *Current Directions in Psychological Science*, 13(5):186–189.

David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics*, 63:53–61.

Wei Li, Hannah Rohde, and Martin Corley. 2022. Veritable untruths: Autistic traits and the processing of deception. *Journal of Autism and Developmental Disorders*, 52(11):4921–4930.

Jia E Loy, Hannah Rohde, and Martin Corley. 2017. Effects of disfluency in online interpretation of deception. *Cognitive Science*, 41:1434–1456.

Jia E Loy, Hannah Rohde, and Martin Corley. 2019. Real-time social reasoning: the effect of disfluency on the meaning of some. *Journal of Cultural Cognitive Science*, 3(2):159–173.

Peter E McKenna, Alexandra Glass, Gnanathusharan Rajendran, and Martin Corley. 2015. Strange words: Autistic traits and the processing of non-literal language. *Journal of autism and developmental disorders*, 45:3606–3612.

Don C Mitchell and David W Green. 1978. The effects of context and content on immediate processing in reading. *The Quarterly Journal of Experimental Psychology*, 30(4):609–636.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

University of Colorado. Word embedding analysis tools. Accessed: 2024-05-27.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. Lard: Large-scale artificial disfluency generation. *arXiv preprint arXiv:2201.05041*.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Fatemeh S Tarighat, Walid Magdy, and Martin Corley. 2022. Understanding fillers may facilitate automatic sarcasm comprehension: A structural analysis of twitter data and a participant study. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 215–217.