When Speech Becomes Writing: The Case of Disfluencies

Anonymous ACL submission

Abstract

Occurrence of disfluencies such as "um" in spoken language may be intentional or unintentional. Intentional use of disfluencies can convey non-literal meanings, among other things. When written, disfluencies are intentional. Although disfluencies have frequently been dismissed as noise and, consequently, removed from the transcriptions of spoken language and training data, they are increasingly appearing in online writing. We investigate the potential role of written disfluencies in signaling non-literal meaning and readers' interpretation using two human experiments.

1 Introduction

With the advent of easy electronic communication and social media, written language has become more speech-like (e.g., Eisenstein et al., 2014). One aspect of this change is the use of written disfluencies, such as "um". There is controversy as to 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 the 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 in models to enhance their performance in human-computer interaction research, 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 the literal meaning with regard to the disruption caused by the disfluency. Disfluencies could be of potential significance in interpreting non-literal meanings. Several NLP studies have looked into non-literal language understanding by focusing on idiom, metaphors, and sarcasm (e.g., Sporleder and Li, 2009; D'Arcey et al., 2019; Desai et al., 2021; Hu et al., 2022). Disfluencies remain understudied in non-literal meaning interpretation.

We aim to investigate whether written disfluencies can signal non-literal meaning, or whether they can influence the ways in which readers interpret what they are reading. Our hypothesis stems from the idea that for human readers, written disfluencies should render the text more "speech-like" which in turn would open it up to less literal interpretation. In a previous study [CITATION AFTER REVIEW], we looked into how users were writing "um", "uh", "hmm", "erm", and "er" on Twitter. To test whether disfluencies could affect interpretation, as it has been reported for spoken disfluencies (Loy et al., 2017), we asked participants to rate randomly selected tweets with and without "um" and "hmm" for their potential sarcastic tone, offensiveness, language formality, and the emotions associated with them. Sarcasm scores were slightly, although not significantly, higher when "um" and "hmm" were kept in the tweets. "Um" and "hmm" also contributed to higher offensiveness scores, the tweets being perceived as less formal, and higher ratings of surprise.

Here, we present two experiments, one completed and one in process at the time of writing) designed to investigate readers' interpretations of written disfluency.

2 Experiment 1: Self-paced Reading

In an online word-by-word self-paced reading experiment, we hypothesized 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*), 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*).

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 asked 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. Each participant was shown only one version of each sentence. We also asked them to provide feedback on interpretability and readability of sentences. We kept 24 sentences for the self-paced reading experiment. Each sentence had 4 variations based on the meaning and whether the target word was preceded by "um": literal-fluent, literal-disfluent, sarcastic-fluent, and sarcastic-disfluent. In each sentence, 5-6 words were marked as regions for reading time depending on the number of the words after the target word. The target word (literal or sarcastic) was the 4th region in all of the sentences.

The self-paced reading task followed the structure of an ILS Labs moving window experiment using jsPsych[1]. We recruited 101 L1-English, UK-based, and non-dyslexic participants through Prolific[2]. Each participants read only one variation of the sentences, total of 24 items in a set. We analyzed reading time data of 99 participants. We removed 2 participants because they got fewer than 6 of the 8 attention-check questions correct. Moreover, 1 item was miscoded in the experiment, resulting in 28 missing trials (1.18% of the data).

We compared the reading times of the target word and of the target word + the next word (for spillover). Contrary to our hypotheses, words compatible with the sarcastic reading of the sentences were not faster to read when preceded by "um" and the maximally fitting models only showed the effect of fluency (β = -83.23, SE = 12.31, $p <$ .001). However, a closer look at the next-word region showed main effect of meaning (literal faster to read than sarcastic; β = 50.13, SE = 21.11, p = .017) and fluency (disfluent slower to read than fluent; β = -31.49, SE = 12.82, p = .014) but no interaction between the two.

3 Experiment 2: Eye Tracking

Whereas Experiment 1 failed to show that written disfluency indexes non-literal 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 *speech-like* prosody with which readers may have read the experimental sentences, reducing any effect that the traditionally spoken element "um" may have had in writing. For that reason, Experiment 2 is a replication of Experiment 1 using an eye-tracking methodology in which *natural* reading prosody is not disrupted.

We have made changes to the items for the eye-tracking experiment. The literal and non-literal words in the two versions of each item have the same number of characters (MERRY/FERAL). We have also counterbalanced the literal and non-literal readings of each word (MERRY (LITERAL)/FERAL (NON-LITERAL) and FERAL (LITERAL)/MERRY (NON-LITERAL)). Once we have ethical clearance, we will recruit 35 L1-British-English speakers to rate the items for potential sarcastic tone online. We will use the items above the cutoff point for the eye-tracking experiment and will have 4 variations of each based on the meaning and fluency.

We will use EyeLink 1000 Plus tracker with Experiment Builder[3] for presentation and Data Viewer[4] for analysis. We will recruit 80-100 neurotypical L1-British-English participants with normal/corrected-to-normal vision and no reading disorders for a 40-minute task in our eye-tracking laboratory. Each participant will see one variation of the items along with filler items.

4 Discussion

Our experiments suggest that readers may be sensitive to written disfluencies, and that LMs/LLMs may need to be trained on different data in order to take this into account. More generally, they point to the importance of (small) data sets stemming from behavioral experiments in evaluating the behaviors of models trained on (large), usually written, language corpora.

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

[2] https://prolific.co/

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

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

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.

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

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.

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*.

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.

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

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.