Disfluency in Spontaneous Speech (DiSS) Workshop 2023
28-30 August 2023, Bielefeld, Germany

Filled pauses and false starts do not reliably preface longer or more complex utterances across typologically diverse languages

Ludger Paschen[1]

[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft

paschen@leibniz-zas.de

Abstract

Disfluencies such as fillers and false starts provide a window into the cognitive processes underlying speech planning. One of the reasons for the occurrence of disfluencies is increased processing load due to the size or complexity of an upcoming utterance. Previous research indicates that disfluencies correlate with an above-average size of following utterances. This paper tests this hypothesis by comparing the size of inter-pausal units that follow a disfluency with those that do not, based on a sample of 51 diverse languages from a language documentation corpus. Results suggest that the presence of disfluencies has little to no effect on the size or complexity of upcoming utterances, be they measured in phonological or morphological terms. A discussion of potential alternative explanations is offered.

Index Terms: disfluency, filled pause, false start, inter-pausal unit, phonetic corpora

1. Introduction

An important finding from disfluency research is that the occurrence of disfluencies is not random, but rather reflects the speaker's cognitive state during language production. During speech planning, speakers may use fillers such as "uh" or "um" when they search for the right word or phrase. At the same time, disfluencies play an important role in interaction, where they can serve as pragmatically meaningful signals to interlocutors [1, 2].

The timing and distribution of disfluencies can reveal the cognitive load of language processing, with more disfluencies occurring during complex or difficult language tasks [3]. [4] showed that false starts or repetitions are more likely to occur the more complex the upcoming constituent. In a listening experiment, [5] found that both silent and filled pauses cause speakers of Japanese to anticipate a long or complex upcoming phrase. Similarly, [6] observed overall higher disfluency rates when speakers discussed abstract figures in a corpus of English task-oriented conversations. In an eye-tracking experiment, [7] found that fluent stimuli cause listeners to expect old referents while stimuli with fillers bias listeners toward new referents. Disfluencies also appear to affect the temporal properties of smaller domains; for example, [8] found that English function words were longer and less reduced in the vicinity of disfluencies.

However, the relation between disfluencies and utterance length has so far only been investigated for a small set of languages, which raises the question of how common potential dependencies between disfluencies and utterance size are across languages. This paper seeks to shed light on this question by comparing various measures of utterance complexity following disfluencies on a sample of 51 areally and genealogically diverse languages from the DoReCo corpus [9]. Utterances will be operationalized as *inter-pausal units* (IPUs), which are defined as chunks of connected speech delineated by two silent pauses. IPUs have the advantage of being a transparently defined and universally applicable unit for speech chunks that does not rely on language-specific categories such as "clause" or "intonational phrase". For disfluencies, this paper will focus on filled pauses and false starts, as annotations for these are readily available in the DoReCo corpus. It should be noted that while this study is mostly cognitive-oriented, I remain agnostic with respect to the signal-symptom debate (see [10] for discussion). The following sections present the methods and data (Section 2) as well as the results (Section 3), and offer some further discussion and suggestions for further research (Section 4).

2. Methods and data

2.1. The DoReCo corpus

DoReCo (Language DOcumentation REference COrpus) is a corpus containing spoken language documentation data on a worldwide sample of 51 languages [9]. Data in DoReCo originated from fieldwork-based documentation of small and endangered languages. DoReCo contains over 100 hours of audio-recorded, mostly narrative texts with transcriptions that are time-aligned at the phone and word level; in addition, time-aligned interlinear morphological glossing exists for a subset of 38 languages. DoReCo data are freely accessible under Creative Commons licenses via the website `https://doreco.huma-num.fr/`. DoReCo features both *core* and *extended* data, the former having completed a rigorous process of manual correction of word boundaries and annotation of disfluencies.

2.2. Annotation of silent pauses and disfluencies

Word boundaries were first created using the WebMAUS forced aligner [11] and subsequently corrected manually for all core datasets in DoReCo. The onset and offsets of silent pauses were also thoroughly checked manually to avoid falsely allocating the pause status to closure phases of stops or to hesitations. By extension, this means that the boundaries of inter-pausal units can be considered reliable for the purpose of this study. For more details on the data processing pipeline in DoReCo, see [12].

DoReCo employs a range of labels for marking special speech events that cannot be reasonably time-aligned on the phone level. An overview of these labels is given in Table 1. For labelling filled pauses, annotators relied on their auditory impression as well as existing annotations, where filled pauses were sometimes transcribed with special character strings such as *mmm*, or directly annotated in the glosses. In most cases, however, filled pauses were manually added during data processing. For annotating false starts, again, existing transcriptions were used when available, but in the majority of cases,

10.21437/DiSS.2023-3

false starts had to be manually identified and added by the human annotators.

Table 1: *Labels for disfluencies and other non-alignable speech events used in DoReCo 1.2.*

Speech event	Label
Filled pause	<<fp>>
False start	<<fs>>
Prolongation	<<pr>>
Foreign material	<<fm>>
Singing	<<sg>>
Backchannel	<<bc>>
Ideophone	<<id>>
Onomatopoeic	<<on>>
Word-internal pause	<<wip>>
Unidentifiable	<<ui>>

2.3. Measures of IPU length and complexity

Word-level CSV files for 51 languages in DoReCo 1.2 were downloaded from `https://doreco.huma-num.fr/`. Inter-pausal units were identified as any sequence of tokens situated between two silent pauses (indicated by the <p:> symbol). IPUs were extracted and exported to a separate CSV file using a custom-made Python script, which also calculated for each IPU (i) its total duration in ms, (ii) its duration excluding IPU-initial disfluencies, (iii) the number of phones, (iv) the number of morphs, (v) the number of meanings expressed by the morphs. In order to assess the effect of an IPU-initial disfluency on the length of the remaining IPU, only the duration of IPUs excluding IPU-initial disfluencies was considered in this study. To calculate the number of meanings for a given morph, each string separated by a dot, colon or "+" character was counted as one meaning. For example, the gloss "PROG" was counted as one meaning, while the glosses "2PL.GEN", "house:ACC" and "give+FUT" were each counted as contributing two meanings.

Word units were not included as a measure of length or complexity. The reason for this is the heterogeneity of transcription conventions found in language documentation data upon which DoReCo is built. These conventions are often non-standardized orthographies developed for practical purposes by the field linguists, usually in a joint effort with the speaker communities, and the criteria for deciding which lexical items and grammatical markers are separated by a space may differ vastly from case to case. Moreover, the comparative concept of "wordhood" is problematic from a typological point of view for various reasons [13]. Phones and morphs are less affected by arbitrary conventions and thus provide more stable units for bottom-up cross-linguistic comparison.

2.4. Data filtering

The CSV described in the previous paragraph was merged with another CSV containing metadata information, and that table was then read in R [14] using RStudio [15]. This initial dataset consisted of 151,874 IPUs. Only texts from core speakers, i.e. speakers for which manual annotation of pauses and disfluencies had been conducted within DoReCo, were considered, leaving 144,636 IPUs from 393 unique speakers. Then, IPUs containing zero word tokens were excluded, reducing the dataset to 131,719 IPUs. Zero-word IPUs contain only labelled content, which can be anything from short isolated backchannels to larger stretches of singing, or whole utterances in a language other than the object language. Another filter was used to exclude IPUs containing filled pauses or false starts at any position other than directly at the beginning of an IPU, as disfluencies occurring within an IPU are likely to influence the duration and complexity of that IPU beyond what can be reasonably controlled for in this study. Excluding "internally disfluent" IPUs cut the size of the dataset down to 124,949 IPUs, which constitutes the final sample upon which duration and phone count analyses were performed. For morph and meaning counts, 38 languages with interlinear morph-level annotation were considered, leaving a sample comprising 88,881 IPUs from 290 unique speakers.

2.5. Statistical modelling

Linear mixed models were fit to analyze the effect of the predictor variables `sp_dis_X` (= IPUs with an initial disfluency) and `dis_sp_X` (= IPUs preceded by a disfluency across a silent pause). Models were fitted using the `lmerTest` library [16]. The fixed effects had three levels: no disfluency (`no`), filled pause (`fp`), and false start (`fs`). The models included three random effects terms, which account for the variability between levels of the grouping factors: `Speaker`, `Genre`, and `Language`. Specifically, the model featured a random intercept for each unique speaker, as it is expected that each speaker has their own baseline for IPU length and complexity. Similarly, a random intercept for each language was set to account for potential language-specific effects. Finally, a random intercept for each unique genre of speech was included, too, as monologic narratives may be enacted differently than conversations involving multiple participants. It is worth pointing out, though, that the majority of recordings in DoReCo 1.2 consist of traditional and personal narratives, and that the number of conversations and other speech genres is relatively low. The effect of the predictor variables was tested against four measures of complexity: (log-transformed) IPU duration, number of phones, number of morphs, and number of meanings. The CSV table and R script are available at `https://osf.io/8ezj7/`.

3. Results

Visual inspection of the duration plots in Figure 1 suggests almost identical distributions and median values regardless of the presence or absence of a disfluency before an IPU. This holds for both filled pauses and false starts, and independently of whether a silent pause intervenes between the disfluency and the IPU onset. The mean duration for IPUs not beginning with a disfluency was 1384 ms, while the mean remaining duration of IPUs starting with fillers and false starts was 1455 ms and 1387 ms, respectively. Standard deviation values were quite substantial, with 11673 for the control group and 1029 and 911 for filled pauses and false starts, respectively. Disfluencies occurring across a silent pause displayed similar mean IPU duration and sd values (mean = 1386 ms, 1424 ms, 1326 ms; sd = 11865, 974, 868).

Linear mixed models revealed significant effects for false starts at the onset of IPUs, and for both false starts and filled pauses separated from an IPU by a silent pause (Table 2). However, effect sizes were negligible (`sp_dis_X`: Cliff's delta estimate = -0.059; `dis_sp_X`: Cliff's delta estimate = -0.037). When running separate smaller models for each genre, only

	sp_dis_X		dis_sp_X	
	fp	fs	fp	fs
Duration	.	***	***	*
N(phones)	ns	*	.	ns
N(morphs)	ns	ns	ns	ns
N(meanings)	ns	ns	ns	ns

Table 2: *Significance matrix for four predictors and four arrangements of disfluencies (*`dis`*) relative to silent pauses (*`sp`*) and (remainders of) IPUs (*`X`*).*

false starts in the `sp_dis_X` condition reached simple significance (*) in the majority of the models, while the other effects disappeared. Combined with the overall small absolute differences of less than 100 ms between categories, this suggests the effects have, in fact, little to no practical relevance. It is likely that the statistical significance here arises due to a large sample size and does not have meaningful implications [17, 18]. The results thus only lend weak support to the hypothesis that disfluencies preface longer upcoming utterances.

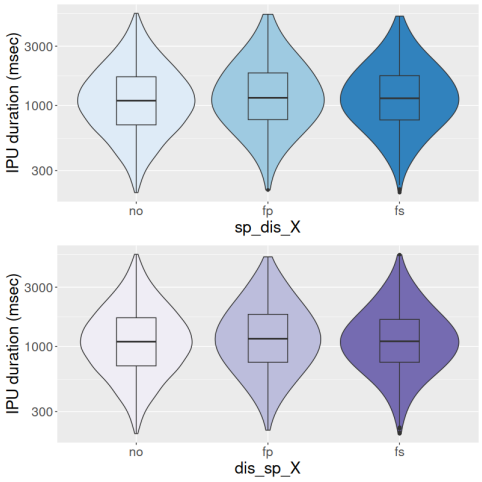

Figure 1: *Various types of arrangements of disfluencies in the vicinity of an IPU vs. IPU duration. The following abbreviations are used:* `dis_sp_X` *= disfluency (filled pause or false start) followed by a silent pause followed by an IPU, and* `sp_dis_X` *= silent pause followed by a disfluency followed by the remainder of an IPU.*

Among the random factors, `Speaker` showed the greatest variability, ranging between ±0.5 (corresponding to about 650 ms below or above the average IPU length). This attests to a considerable degree of inter-speaker variation with respect to IPU duration. Figure 2 shows the random effect sizes for `Genre` and `Language`. The effect of `Genre` was relatively small (±0.02 or 20 ms), but traditional narratives showed a tendency towards longer IPUs while conversations displayed the opposite effect. `Language` (±0.4 or 500 ms) had an effect comparable to that of `Speaker`. While there do not seem to be clear genealogical groupings emerging from Figure 2, one areal pattern can be observed: among the six languages with the largest negative effects, five were languages from the African continent (Baïnounk Gubëeher, Gorwaa, Kakabe, N||ng, Tabaq).

Figure 3 shows the distribution of the number of phones for

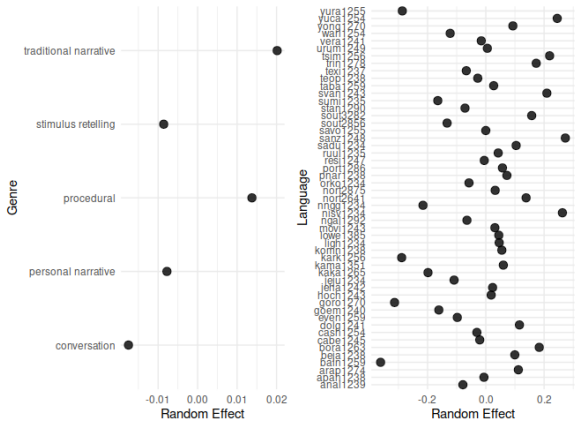

Figure 2: *Effect structure for* Genre *and* Language *(log scale).*

IPUs depending on whether or not an IPU follows a disfluency. As with duration, disfluencies appear to have little to no impact on the number of phones in an IPU. The mean number of segments for IPUs not beginning with a disfluency was 15.9, while the mean segment counts in IPUs starting with fillers and false starts were 16.6 and 16.4, respectively (sd = 12.1, 13.1, 11.8). Disfluencies occurring across a silent pause displayed virtually the same means and sd values (mean = 15.9, 16.4, 15.8; sd = 12.2, 12.2, 11.4). Linear mixed models revealed a significant effect for false starts at the onset of IPUs, but not across silent pauses, and no significant effect for filled pauses at all.

The finding that duration was slightly elevated after disfluencies but number of phones remained virtually stable raises an interesting question: Could the former effect be caused by a slower speech rate which merely creates the illusion of more contentful utterances? To answer this, local speech rate was calculated as phones for second for all IPUs, and additional linear mixed models were run with spech rate as the dependent variable. Results turned out to be significant (**) for filled pauses in `sp_dis_X`, but not for other contexts. This suggests that at least for utterances starting in a filler, the apparent lengthening effect could well be an epiphenomenon of an overall slower speech rate, which is compatible with a processing difficulty scenario.

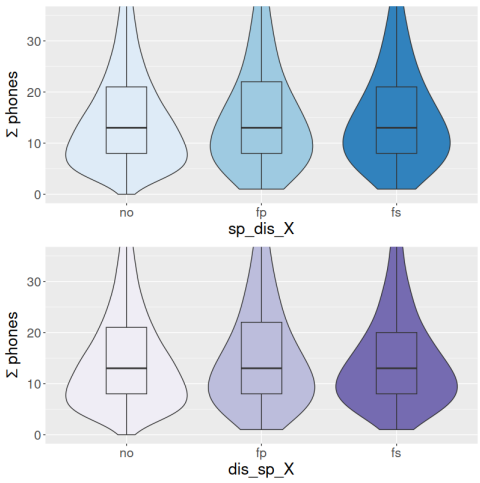

Figure 3: *Various types of arrangements of disfluencies before an IPU vs. number of phones in an IPU.*

For number of morphs, the distributions are highly similar across all conditions (Fig. 4). Means for the `sp_dis_X` condition were 5.4 (sd = 4.5) for no disfluencies, 5.8 for fillers (sd = 4.9) and 5.7 for false starts (sd = 4.4). Means for the `dis_sp_X` condition were 5.4 (sd = 4.5) for no disfluencies, 5.4 for fillers (sd = 4.3) and 5.2 for false starts (sd = 4.3). None of the effects reached statistical significance, and given sd values almost as high as the means, no notable impact of disfluencies on the complexity of following IPUs in terms of morph count could be confirmed.

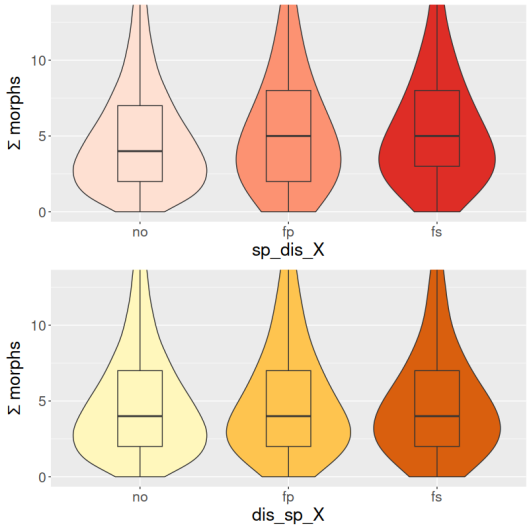

Figure 4: *Various types of arrangements of disfluencies before an IPU vs. number of morphs in an IPU.*

For number of meanings, the distributions resemble those of number of morphs (Fig. 5). Means for the `sp_dis_X` condition were 7.0 (sd = 6.1) for no disfluencies, 7.2 for fillers (sd = 6.4) and 7.3 for false starts (sd = 5.8). Means for the `dis_sp_X` condition were 7.0 (sd = 6.1) for no disfluencies, 6.9 for fillers (sd = 5.7) and 6.7 for false starts (sd = 5.6). Again, none of the effects reached statistical significance, thus no evidence for a meaningful impact of disfluencies on the complexity of following IPUs in terms of lexical or grammatical meanings could be found.

4. Discussion

The lack of a strong cross-linguistic trend for increased unit size following disfluencies is surprising in light of previous findings discussed in Section 1. There may be several reasonable explanations for this. First, languages may differ substantially in the use and distribution of disfluencies, and communicative strategies observed for English or Japanese may not be representative of broader cross-linguistic trends. Second, speakers have other means than fillers and restarts at their disposal to signal planning difficulties, the most obvious being silent pauses. In their study on English based on the Switchboard corpus, [8] found that disfluencies and silent pauses had a similar effect on the likelihood of function words being lengthened. Third, measures related to IPUs do not capture the potential impact that disfluencies have on other properties of spontaneous speech. [19] stated that *uh* and *uhm* foreshadow pauses of different lengths (but see [20] and [21] for critical discussion), which suggests that effects of disfluencies may be better visible outside rather than within IPUs. Overall, it seems reasonable to assume that conceptual

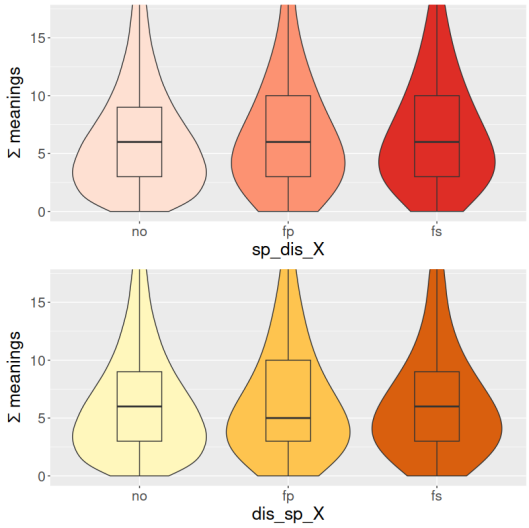

Figure 5: *Various types of arrangements of disfluencies before an IPU vs. number of meanings in an IPU.*

difficulties may manifest through a combination of filled and silent pauses, and also segmental and prosodic properties.

Another result of the present study was that there were only very limited differences between filled pauses and false starts. [6] note that in their sample, fillers were distributed differently than restarts, which they interpret as evidence that the former serve as a resource for interpersonal coordination. The closest to a distinction between fillers and false starts in the present study was the finding that when conflating duration and N(phones) into a measure of speech rate, filled pauses but not false starts showed a significant effect in the `sp_dis_X` condition, indicating that the former have a local slowing effect.

Future research should direct attention to more variables and possible confounds. Section 3 discussed IPU duration with respect to `Genre`, `Speaker`, and `Language`. It is reasonable to assume `Genre` could impact the size and complexity of IPUs beyond what was found in this study, especially with respect to conversations, which introduce an additional layer of complexity through turn and sequence organization. Inter-speaker variation was confirmed in the present study and is expected to impact all aspects of speech. Finally, the potential emergence of an areal clustering for African languages suggests a promising avenue to further scrutinize `Language` and its relation to other prosodic and grammatical features.

5. Conclusion

A corpus study on 51 diverse languages revealed no or only negligible effects of disfluencies on the size or complexity of an upcoming utterance, as measured in duration, number of phones, number of morphs, and number of meanings. This suggests that the occurrence of disfluencies alone is not a cross-linguistically reliable predictor for a following long, and hence cognitively demanding, chunk of speech. This raises serious questions about the functional typology of disfluencies and their interplay with other language-specific cues.

6. Acknowledgements

This research was supported by a grant by the German Research Foundation to Ludger Paschen (DFG-PA2368/1-1).

7. References

[1] G. Tottie, "On the use of uh and um in american english," *Functions of Language*, vol. 21, no. 1, pp. 6–29, 2014.

[2] L. Kosmala, "On the specificities of l1 and l2 (dis)fluencies and the interactional multimodal strategies of l2 speakers in tandem interactions," *Journal of Monolingual and Bilingual Speech*, vol. 3, no. 1, pp. 69–101, 2021.

[3] M. H. Christiansen and N. Chater, *Creating language: Integrating evolution, acquisition, and processing*. MIT Press, 2016.

[4] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, pp. 201–242, 1998.

[5] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, "Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners," *Speech Communication*, vol. 50, pp. 81–94, 2008.

[6] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and speech*, vol. 44, no. 2, pp. 123–147.

[7] J. E. Arnold, M. K. Tanenhaus, R. J. Altmann, and M. Fagnano, "The old and the thee, uh, new: disfluency and reference resolution," *Psychological Science*, vol. 15, no. 9, pp. 578–582, 2004.

[8] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," *Journal of the Acoustical Society of America*, vol. 113, pp. 1001–1024, 2003.

[9] F. Seifart, L. Paschen, and M. Stave, Eds., *Language Documentation Reference Corpus (DoReCo) 1.2*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022. [Online]. Available: https://doreco.huma-num.fr

[10] M. Belz, *Die Phonetik von äh und ähm: Akustische Variation von Füllpartikeln im Deutschen*. Berlin: J.B. Metzler, 2021.

[11] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using the maus technique," in *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, 2011.

[12] L. Paschen, F. Delafontaine, C. Draxler, S. Fuchs, M. Stave, and F. Seifart, "Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo)," in *Proc. LREC 2020, Marseille*, 2020, pp. 2657–2666. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.324

[13] M. Haspelmath, "The indeterminacy of word segmentation and the nature of morphology and syntax," *Folia Linguistica*, vol. 45, pp. 31–80, 2011.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, 2022, r version 4.2.1. [Online]. Available: https://www.R-project.org/

[15] R. Team, *RStudio: Integrated Development Environment for R*, 2022, rStudio 2022.07.2 Build 576. [Online]. Available: https://www.rstudio.com/

[16] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmertest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[17] A. P. Association, *Publication manual of the American Psychological Association (6th ed.)*. American Psychological Association, 2010.

[18] G. Cumming, *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2012.

[19] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[20] D. C. O'Connell and S. Kowal, "Uh and uhm revisited: are they interjections for signaling delay?" *Journal of Psycholinguistic Research*, vol. 34, no. 6, pp. 555–576, 2005.

[21] S. Betz and L. Kosmala, "Fill the silence! basics for modeling hesitation," in *Disfluency in Spontaneous Speech 2019*, Budapest, Hungary, 2019, pp. 12–14. [Online]. Available: https://hal.science/hal-02360611