INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Revisiting recall effects of filler particles in German and English

Beeke Muhlack, Mikey Elmers, Heiner Drenhaus, Jürgen Trouvain, Marjolein van Os, Raphael Werner, Margarita Ryzhova, Bernd Möbius

Language Science and Technology, Saarland University, Saarbrücken, Germany
muhlack@lst.uni-saarland.de

Abstract

This paper reports on two experiments that partially replicate an experiment by Fraundorf and Watson (2011, J Mem. Lang.) on the recall effect of filler particles. Their subjects listened to three passages of a story, either with or without filler particles, which they had to retell afterwards. They analysed the subjects' retelling in terms of whether important plot points were remembered or not. For their English data, they found that filler particles facilitate the recall of the plot points significantly compared to stories that did not include filler particles. As this seems to be a convincing experimental design, we aimed at evaluating this method as a web-based experiment which may, if found to be suitable, easily be applied to other languages. Furthermore, we investigated whether their results are found in German as well (Experiment 1), and evaluated whether filler duration has an effect on recall performance (Experiment 2). Our results could not replicate the findings of the original study: in fact, the opposite effect was found for German. In Experiment 1, participants performed better on recall in the fluent condition, while no significant results were found for English in Experiment 2.
Index Terms: filler particles, hesitations, recall, memory

1. Introduction

Filler particles (henceforth FPs) like *uh* and *um* (synonyms are 'hesitation particles', 'filler tokens', 'filled pauses' or simply 'fillers') occur frequently in spontaneous discourse [1, 2]. They are not regarded as words in a conventional sense [3, 4] and were described for many languages, where they can be observed in two main phonetic forms: as vowel only and as a sequence of vowel plus nasal consonant (e.g. [ɔː/ɔm] for German [5]).

Several studies tested the influence of FPs on the recall of the immediate context. To our knowledge, Fraundorf and Watson (2011) [6] (henceforth F&W) is one of few studies that extended their research of recall effects of FPs to the discourse level. As their experiments led to the promising results that FPs may improve recollection [6], it is reasonable to transfer their experimental design to other languages. In our study we attempt to verify whether the test paradigm used by F&W [6] is suitable for testing the recall effects across languages.

The FPs *uh* and *um* are usually classified as a type of disfluency, suggesting that these FPs interrupt the speech flow. They are considered as signalling a delay of speech and thus a difficulty of speech processing on the speaker's part [4, 7]. Using FPs such as *uh* and *um* is often seen as an undesirable aspect of speech production especially in public speaking [8]. From a speech perception perspective this view has been questioned. In fact, it has been shown that FPs can have several benefits for the listener. They support discourse structure in that they are more likely to occur at phrase boundaries [9]; they are often produced before a new topic or a new referent is introduced, which helps the listener to process the discourse structure [10]; and they may emphasise the following word [11].

Hesitations orientate listeners' attention towards the following speech material and lead to immediate processing of an utterance as well as to better recognition when the word (that previously occurred after a FP) appears again [12]. Additionally, [13] reported that hesitations are associated with longer-term consequences for the representation and processing of the message. These results are in line with the results reported by Fraundorf and Watson [6]. F&W [6] propose that the improved recollection of information after FPs is due to an attention-orienting effect towards upcoming material, rather than additional processing time for the speaker. However, whereas [13] found that words preceded by disfluency were more likely to be remembered, [14] found no effect of FPs on memory. This conflicting result might be attributed to a difference in experimental paradigm and stimuli. As such, the question whether FPs can facilitate recall is still inconclusive.

The method used by F&W seems appropriate for testing the impact of FPs on recall while also being simple enough to be employed for other languages. Furthermore, the design is suitable for a web-based study, which was necessary due to the Covid-19 pandemic. One aim of this study was to replicate the findings by F&W for another Germanic language, namely German, and thus validate that the method is in fact applicable for testing recall effects in different languages. The second aim of this study is to partially replicate Experiment 1 from F&W, but this time under remote test conditions and with a slightly different set of stimuli.

2. Methodological details

We were aiming at partially replicating the first experiment of F&W who presented three short passages from *Alice in Wonderland* [15] retold by a female speaker of (American) English to their participants, who were then asked to retell the three stories in their own words. Each story consisted of 14 plot points that represented one or two closely connected events that are important for the outcome of the story. The authors created three conditions for their experiment: a version where all FPs were removed (fluent condition), a version including FPs before six plot points (FP condition), and a version that included coughs (instead of FPs) that were matched in duration before the same six plot points (cough condition). For the analyses, the retellings of the subjects were checked for whether the plot points were correctly recalled or not. For the present studies, we hypothesise that the stories in the FP condition are recalled better compared to the fluent and long silence conditions.

Scott Fraundorf and Duane Watson kindly agreed to make the original recordings of their experiment available to us. We found that their FPs were longer than those observed in natural spontaneous speech [16]. Furthermore, the "fluent" version still contains disfluencies such as lengthenings. After informal inspection, an average of 17 lengthened syllables were found in

http://dx.doi.org/10.21437/Interspeech.2021-1056

Table 1: *Mean (SD) FP duration of stimuli in ms in the experiments reported here compared to stimuli by F&W.*

FP type	*uh*	*um*
Exp1 (FP condition)	482 (112)	679 (59)
Exp2 (short FP condition)	571 (77)	726 (74)
Fraundorf & Watson	1314 (252)	1441 (290)

each story. It may be the case that the duration of the FPs had an impact on the listeners' recall of the plot points. In order to check this claim we replicated the study in two experiments and made some changes to the experimental setup.

The insertion of the FPs and silences into the fluent version of the stories was done using Praat [17], modifications to intensity were done with Audacity [18]. Both experiments were designed as web-based experiments using Labvanced [19] for presenting the audio stimuli and recording the participants' answers. For recruitment and payment of the subjects we used the online participant pool Prolific [20].

3. Experiment 1

3.1. Method

The first experiment was created to replicate the findings of F&W for English with German data [1], specifically focusing on the benefit of FPs on recall. We also used three conditions: a fluent condition, an FP condition, and a long silence condition. Instead of using coughs in one of the conditions (as it was done by F&W), the pause between sentences was extended to match the duration of the FPs in the FP condition.

The stories and plot points were translated into German on the basis of the transcripts provided by F&W. Recordings of the stories were then made by a female speaker of German (first author) using an H1 Zoom recorder with a clip-on microphone in a quiet room. The speaker memorised each plot point and then retold it rather than reading from the script.

The best version (i.e., the most fluent) of each plot point was first concatenated into a fluent version of the story, which was then used for splicing in FPs or silence. The fluent version was approved by the co-authors as sounding natural; in particular, the concatenation was not noticeable. The fluent version did not contain any further disfluencies such as syllable lengthening. For the FP condition, six FP tokens (3 *uh*, 3 *um*) were chosen from disfluent versions of the story recorded by the same female speaker. The intensity of the FP was adjusted such as to make the splicing least noticeable. The duration of these FPs were shorter than those used by F&W (Table 1).

The long silence condition was created by splicing in "silent" segments taken from the same audio file, and inserting them before the same six plot points that were manipulated in the FP condition. The silences matched the FPs in duration. Each of the three stories was created in every condition (fluent, FP, long silence, see Fig. 1), resulting in a total of nine different versions. Each participant heard every story in one of the conditions. The stories were presented in randomised order. A total of 45 native German speakers (mean age 31.2 yrs; age range 18–63 yrs; 19 females, 25 males, 1 non-binary) participated in this experiment. None indicated the use of a hearing aid.

[1]The stimuli for Experiment 1 are available on www.pauseparticles.org.

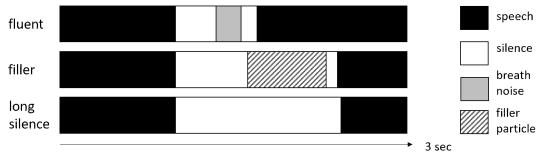

Figure 1: *Schematic representation of a pause in each of the three conditions for Experiment 1.*

Table 2: *Contrast matrix of Experiment 1 (German) and Experiment 2 (English) which differ in experimental conditions.*

Experiment 1 Experiment 2	fluent fluent	FP long FP	long silence short FP
Contrast1 (C1)	2/3	-1/3	-1/3
Contrast2 (C2)	0	1/2	-1/2

Two annotators (one of them the first author) listened to the retellings of the participants to make a binary choice whether or not each plot point was correctly recalled. Inter-rater reliability was calculated from a subset of stories that were annotated by both raters (about 20%), yielding only 5.3% disagreement between the two annotators in this subset. Where the decisions differed, the first author's annotations were used.

3.2. Results

Analyses were performed in R (Version 3.6.1) by fitting Generalised Linear Mixed-Effects Models using the lme4 [21] package (Version 1.1-21). Estimates, standard error and z-value are reported. We extracted confidence intervals by using the profile function of the stats package (Version 3.6.1). Model fit was used to determine by-model comparison applying the ANOVA function also implemented in the stats package such that models with a lower AIC (Akaike Information Criterion) were preferred. Models with maximal random structure were fitted following [22]. Contrasts were coded the following way (Table 2): The first contrast (C1) compares the fluent condition against the other two conditions, which is why the coded value for fluent is the same as the absolute sum of the values of the other two conditions. The second contrast (C2) only compares the FP condition to the long silence condition, which is achieved by setting the value of the fluent condition to equal 0.

Each story contained 14 plot points, only six of which were manipulated, i.e. a FP or longer silence was added before the beginning of the sentence. Statistical analyses using the full data set of 14 plot points per story did not reach statistical significance, indicating that the manipulations did not affect the global recall of the stories. All reported analyses focus on the subset of the six (potentially) manipulated plot points per story. Note that the plot points in the fluent condition remained unmanipulated. Participants listened to three different conditions, leading to 810 observations (45 subjects x 3 stories x 6 plot points) for our analysis.

The full random structure including the C1 and C2 under subject, story and plot point led to singular fit warnings. We followed the approach as described by [23] utilizing PCAs (principle component analyses) to reduce the random structure of the models. The resulting structure used in the final model is the following: $glmer(Answer \sim C1 + C2 + (1 + C1 + C2 \mid Subject) + (1 \mid Story) + (1 \mid Plotpoint), data =$

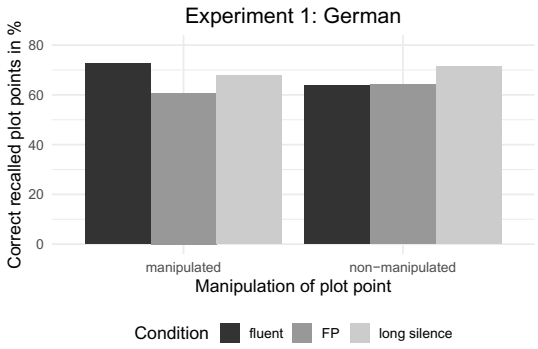

Figure 2: *Correct recall per condition in the German dataset for manipulated and non-manipulated plot points separately.*

$data, family = binomial$). This model explains the response variable (subject's recall of a plot point) by the coded contrasts. Random intercepts were assumed for subject, story and plot point, while random slopes were also added by subject. Our analysis revealed a main effect for C1 (*Estimate* = 0.5098, SE = 0.2021, z value = 2.522, Pr($> |z|$) $< .01$, CI = [0.11993; 0.95432]), and a trend for C2 (*Estimate* = -0.4520, SE = 0.2396, z value = -1.887, Pr($> |z|$) = .06, CI = [-0.96744; 0.01798]). Figure 2 supports the results of the statistical analysis. Recall of the manipulated plot points was best in the fluent condition and worst in the FP condition, with the long silence condition scoring in-between the other conditions. Peculiarly, the contrast between the long silence condition and the other two conditions in the non-manipulated plot points was not shown in the statistical analysis as the initial analysis with 14 plot points did not yield any results.

3.3. Discussion

The significant result for the first contrast, namely the contrast between the fluent condition on the one hand, and the long silence and FP conditions on the other hand, suggests that the fluent versions of the stories are recalled better than the two manipulated versions. The short FPs as well as the additional silence seem to have reduced participants' recall. Furthermore, a tendency for the second contrast can be observed. This trend suggests that the prolonged silence leads to better recall than the FPs. This finding is in contrast with what F&W found in their study for English data, where the FP condition facilitated recall. Participants were able to remember the story in more detail when FPs were included as opposed to the story with no FPs at all. Various factors could be the cause for these diverging results, such as the different languages of the experiments, a different speaker, the different filler duration, differences regarding the number of lengthenings in the fluent condition, the method (web-based vs lab), or the scoring method of the plot point annotations. As the present study developed an individual annotation scheme based on the information reported in F&W it may be possible that we employed stricter (or less strict) rules for judging whether or not a plot point was recalled. These factors will be explained in more detail in the general discussion.

4. Experiment 2

Our second experiment was designed to investigate the factors of Experiment 1 in more detail. It was conducted using the original stimuli from F&W in the fluent and the FP condition. A third condition (short FP condition) was created by shortening the FPs of F&W's FP condition.

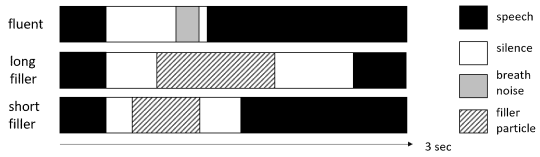

Figure 3: *Schematic representation of a pause in each of the three conditions for Experiment 2.*

4.1. Method

Experiment 2 employed a similar setup as Experiment 1, i.e., three retellings of *Alice in Wonderland* [15] were used which were manipulated by inserting a FP token (*uh/um*) before six plot points for each story. Importantly, the stories in the fluent condition remained unmanipulated (identical with those from the original study). Scott Fraundorf and Duane Watson were kind enough to provide us with their original recordings, which were made by a female native speaker of (American) English.

The three conditions for this experiment are as follows: 1) fluent, i.e., no further manipulations were made; the recordings, however, included syllable lengthening as a form of hesitation, 2) long FPs, as used in [6] (mean duration: 1378 ms), and 3) short FPs, i.e., the length of the FP tokens and the surrounding silence in condition 2 were reduced by 50% (mean duration: 648 ms, Fig. 3). The reduction of the FP was achieved by cutting a stable portion of the vowel and nasal consonant. All resulting stimuli still sounded natural. A total of 58 native English subjects (mean age 31.4 yrs; age range 18–57 yrs; 35 females, 23 males) participated in this experiment. The annotation of participants' answers was performed as reported for Experiment 1, the disagreement between both annotators was 5.2%. Inter-rater reliability was calculated from a subset of stories that were annotated by both raters (about 20%). In case of disagreement, the first author's annotation was selected.

4.2. Results

The statistical analysis and coding of contrasts were performed the same way as in Experiment 1. The first contrast compares the fluent condition to both FP conditions while the second contrast compares only the two FP conditions to one another (Table 2). Each story contained 14 plot points, six of which were manipulated. Again, statistical analyses using the full data set of 14 plot points per story did not reach statistical significance, so the reported analyses focus on the subset of the six manipulated plot points per story. Participants listened to three different conditions, leading to 1044 observations (58 subjects x 3 stories x 6 plot points) for our analysis.

The model with the best fit (following the procedure as described in Experiment 1) was the following: $glmer(Answer \sim C1+C2+(1+C1+C2 \mid Subject)+(1 \mid Story/Plotpoint), data = data, family = binomial)$. This model explains the response variable (subject's recall of a plot point) by the coded contrasts. Random intercepts are assumed for both subject and story, while random slopes are added by subject. Plotpoint was nested under Story. The model did not show significant effects for any of the contrasts: C1: (*Estimate* = 0.0674, SE = 0.1847, z value = 0.365, Pr($> |z|$) =

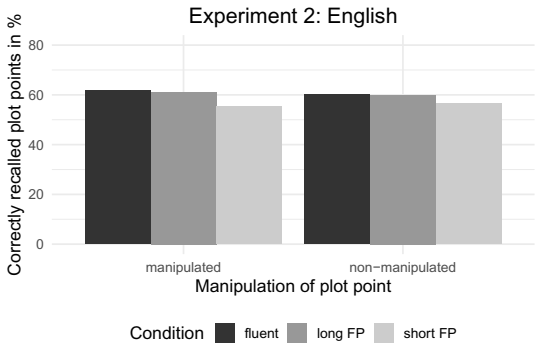

Figure 4: *Correct recall per condition in the English dataset for manipulated and non-manipulated plot points separately.*

0.72, CI = [-0.29912; 0.45693]), C2 (*Estimate* = 0.1944, SE = 0.2314, z value = , Pr($> |z|$) = 0.4, CI = [-0.27079; 0.66927]).

In concordance with the statistical analysis, Figure 4 only shows minimal differences in recall between the three conditions. Similar to Experiment 1, the short FP condition shows a tendency of scoring worst in the recall of the plot points.This tendency could not be shown with the statistical modelling.

4.3. Discussion

In contrast to our expectations, the results of F&W's first experiment could not be replicated in this experiment. The conditions of the stories (fluent, long FP, short FP) did not influence the recall of the manipulated plot points. This lack of significant results can be due to various factors, in particular the different experimental methods (web-based vs lab) and the method of scoring the participants' answers. Possible explaining factors will be discussed in the general discussion.

5. General discussion

In two experiments on the recall effect of FPs we attempted to replicate the results of F&W. Experiment 1, which was conducted in German, shows that the plot points in the fluent condition were better recalled than both the long silence condition and the FP condition. A tendency in the statistical analysis suggests that the long silence condition still leads to a better recall than the FP condition. These results were not obtained in Experiment 2, which was conducted in English. In this experiment, a fluent condition was compared to two FP conditions that differed only in duration of the FP tokens (long vs short condition). The analyses from Experiment 2 did not lead to statistically significant results, suggesting that neither of the conditions had an impact on participants' recall of the manipulated plot points. The results of both experiments are inconsistent with the results reported by F&W, who found that FPs facilitated recall. One possible explanation for the difference between the results from our first experiment and F&W is the language in which the experiments were conducted (German vs English). It is possible that speakers of English benefit from FPs while German speakers prefer a fluent version.

It is more likely, however, that the experimental design is an explaining factor for the different results. F&W conducted their experiment as a laboratory study, which required participants to visit the lab, whereas our experiments were conducted as web-based experiments. One drawback of remote designs is that the researcher cannot control the participant's surroundings and thus cannot exclude distracting factors such as ambient noise or interference from technical devices. These factors may affect the concentration of the participant on the task of the experiment, compared to laboratory experiments where distractors can usually be avoided and, furthermore, the experimental setup is consistent across all subjects. Another factor may be the scoring method employed by the annotators. While it was ensured to keep the disagreement between the annotators of the two experiments reported here to a minimum, it is possible that our procedure still differed from the annotation scheme employed by F&W. A problem the annotators had to face was making a set of rules as to when a plot point was considered as correctly recalled. A simple plot point would easily be annotated while a plot point that includes two closely connected events is more difficult to annotate when the subjects only partially recall the plot point. This set of rules may differ from the annotation method F&W employed.

Previous studies (e.g., [13, 14]) also showed different results with respect to the benefit of FPs in memory tasks, which have also been attributed to differences in experimental design. More research is needed to investigate the effects that design choices have on memory benefit, especially given the finding that listeners are sensitive to speaker-specific characteristics [7, 14] and able to adapt within a single experimental session to an atypical disfluency distribution, also depending on the speaker [24].

6. Conclusions

F&W [6] developed an experimental design to test the recall effect of filler particles (FPs) for American English. The aim of our study was to test this experimental design in a different language and see whether the results are also applicable to our German data. With the two reported experiments, we have shown that the web-based experimental method employed here is not comparable with a similar in-lab experiment as conducted by F&W [6]. While F&W reported results that suggest that filler particles facilitate recall, our Experiment 1 with German data shows the opposite. The fluent condition leads to better recall while the FP condition leads to worse recall. Furthermore, an experiment that closely matched F&W's experiment did not lead to any significant results. It seems that the web-based experiment is less suited to test recall effects since the participants' concentration cannot be ensured.

One motivation for this (partial) replication study was to use an easy-to-copy paradigm to test whether inserted FPs can have a facilitating recall effect for listeners on discourse level. However, for the development of a test paradigm that works across languages and can also be applied as a web-based method, further research has to be conducted.

7. Acknowledgements

We are grateful to Scott Fraundorf and Duane Watson for providing us with the original stimuli of their study and allowing us to use them in our second experiment. We also thank our student assistant Hanna Zimmermann for her reliable work. This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID MO 597/10-1 and Project-ID 232722074 – SFB 1102, as well as private funds from the Max Mangold estate.

8. References

[1] 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, 2001.

[2] J. E. Fox Tree, "The Effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, pp. 709–738, 1995.

[3] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Linguistics and Language Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[4] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[5] B. Muhlack, "L1 and L2 production of non-lexical hesitation particles of German and English native speakers," in *Proceedings of Laughter and Other Non-Verbal Vocalisations Workshop*, B. Ludosan, P. Wagner, M. Rychlowska, and G. McKeown, Eds., Bielefeld, Germany, 2020, pp. 44–47.

[6] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of Memory and Language*, vol. 65, no. 2, pp. 161–175, 2011.

[7] J. E. Arnold, C. L. H. Kam, and M. K. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 5, pp. 914—930, 2007.

[8] J. E. Fox Tree, "Interpreting pauses and ums at turn exchanges," *Discourse Processes*, vol. 34, no. 1, pp. 37–55, 2002.

[9] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, pp. 485–496, 1998.

[10] D. J. Barr, "Trouble in mind: Paralinguistic indices of effort and uncertainty in communication," in *Oralité and gestualité: communication multimodale, intéraction*, S. Santi, I. Guaïtella, C. Cave, and G. Konopczynski, Eds. Paris, France: L'Harmattan, 2001, pp. 597–600.

[11] G. Kjellmer, "Hesitation. In defence of ER and ERM," *English Studies*, vol. 84, no. 2, pp. 170–198, 2003.

[12] P. Collard, M. Corley, L. J. MacGregor, and D. I. Donaldson, "Attention orienting effects of hesitations in speech: Evidence from ERPs." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 34, no. 3, p. 696, 2008.

[13] M. Corley, L. J. Macgregor, and D. I. Donaldson, "It ' s the way that you , er , say it : Hesitations in speech affect language comprehension," *Cognition*, vol. 105, pp. 658–668, 2007.

[14] H. R. Bosker, H. Quené, T. Sanders, and N. H. de Jong, "Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not," *Journal of Memory and Language*, vol. 75, pp. 104–116, 2014.

[15] L. Carroll, *Alice's Adventures in Wonderland*. New York: Sam'l Gabriel Sons & Company, 1916.

[16] E. de Leeuw, "Hesitation markers in English, German, and Dutch," *Journal of Germanic Linguistics*, vol. 19, no. 2, pp. 85–114, 2007.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[18] Audacity Team, "Audacity(r): Free audio editor and recorder [computer application]," 2020. [Online]. Available: https://audacityteam.org/

[19] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "LabVanced: A unified JavaScript framework for online studies," *2017 International Conference on Computational Social Science IC2S2*, pp. 2016–2018, 2017.

[20] "Prolific," Oxford, UK, 2014, accessed: 23.11.20/11.02.21. [Online]. Available: https://www.prolific.co

[21] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[22] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

[23] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," *arXiv preprint arXiv:1506.04967*, 2015.

[24] H. R. Bosker, M. van Os, R. Does, and G. van Bergen, "Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension," *Journal of Memory and Language*, vol. 106, pp. 189–202, 2019.