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. [əz/ən] 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] (henceforward 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 results to stories that did not include filler particles. As their experiments led to the promising results that FPs tend their research of recall effects of FPs to the discourse level. To our knowledge, Fraundorf and Watson [7] proposed 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
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, 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 \textit{uh}, 3 \textit{um}) were chosen from different 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.

<table>
<thead>
<tr>
<th>FP type</th>
<th>\textit{uh}</th>
<th>\textit{um}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp1 (FP condition)</td>
<td>482 (112)</td>
<td>679 (59)</td>
</tr>
<tr>
<td>Exp2 (short FP condition)</td>
<td>571 (77)</td>
<td>726 (74)</td>
</tr>
<tr>
<td>Fraundorf &amp; Watson</td>
<td>1314 (252)</td>
<td>1441 (290)</td>
</tr>
</tbody>
</table>

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

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: \textit{glmer}(Answer \sim C1 + C2 + (1 + C1 + C2 | Subject) + (1 | Story) + (1 | Plotpoint), data =

![Figure 1: Schematic representation of a pause in each of the three conditions for Experiment 1.](image)

<table>
<thead>
<tr>
<th>Contrast1 (C1)</th>
<th>Experiment 1 fluent</th>
<th>FP long silence</th>
<th>short FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 2 fluent</td>
<td>2/3</td>
<td>-1/3</td>
<td></td>
</tr>
<tr>
<td>Contrast2 (C2)</td>
<td>0</td>
<td>1/2</td>
<td>-1/2</td>
</tr>
</tbody>
</table>

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

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

Experiment 2

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 ~ C1+C2+(1+C1+C2 | Subject)+(1 | 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 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]).

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 ~ C1+C2+(1+C1+C2 | Subject)+(1 | 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 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 3: Schematic representation of a pause in each of the three conditions for Experiment 2.
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 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


