From local hesitations to global impressions of the listener

Tanvi Dinkar, Beatrice Biancardi, Chloé Clavel

From local hesitations to global impressions of the listener

Tanvi Dinkar
LTCI, Télécom Paris
IP Paris, France
tanvi.dinkar@telecom-paris.fr

Beatrice Biancardi
LTCI, Télécom Paris
IP Paris, France
beatrice.biancardi@telecom-paris.fr

Chloé Clavel
LTCI, Télécom Paris
IP Paris, France
chloe.clavel@telecom-paris.fr

Abstract

The listener's interpretation of a speaker's utterance includes estimates about the speaker's commitment to what they are saying. Previous works have shown that fillers (e.g. "um") are linked to both the speaker's metacognitive state, and the listener's impression of a speaker's state. However, these results are limited to contexts that may not apply to spontaneous speech. Additionally, there is a lack of hierarchical analysis of the discourse; i.e. how a speaker's local use of fillers could lead to a listener's overall impression. In this work, we address these limitations by studying how does a speaker's use of fillers relate to the incoming message, and consequently, what is the resulting impression formed by the listener. We do so by analysing a dataset of English monologue movie reviews, where the speakers voluntarily and naturally recorded themselves giving a movie review. Our findings show that speakers tend to stylistically use fillers in the incoming message before introducing new information related to the review, and that listeners may not associate this specific use of fillers with their estimate of the speaker's expressed confidence. Our results highlight that there are potentially different metacognitive effects from the speaker's use of fillers on the listener.

1 Introduction

There is a complex relationship between what a speaker says versus the way that a speaker says it; and consequently, what is the resulting impression left on the listener. Consider the following example, taken from Brennan and Williams (1995):

A: Can I borrow that book?
B: . . . {F um} . . . all right.

In the above example, speaker **B** used a filler {F...} (Clark and Fox Tree, 2002), which is a sound filling a pause in an utterance or a conversation. We see that the filler causes **A** to note that **B** might have had a different intention compared to if **B** answered "all right" immediately. While **B** in essence says "yes" to lending the book, the way **B** said this *implicitly* indicates some uncertainty or hesitation.

The aim of this work is to empirically study on a real-life dataset, whether the utterance level use of fillers can help in understanding/ interpreting the perception of the speaker that was formed by a listener. The present work is based on the following observations: According to Brennan and Williams (1995), the listener's interpretation of the speaker's utterance includes estimates about the speaker's commitment to/ expressed confidence in what they are saying. Flavell (1979) termed these processes (of the speaker) as **metacognitive** ones, that is cognition about cognitive phenomena, or more simply "thinking about thinking". While the field of metacognition was initially in the context of children's development and education; the idea of metacognitive states is applicable a wide variety of communicative scenarios. When considering the comprehension of disfluent speech for e.g., research has linked fillers to the listener's assessment of a speaker's metacognitive state (Brennan and Williams, 1995). However, these results may not apply to spontaneous speech datasets collected in real-life contexts, or non-QA datasets. Additionally, the focus of analysis tends to be on utterances as if they occur in isolation, rather than part of an overall discourse.

Thus existing studies do not focus on the connection between the hierarchical levels of discourse; i.e. how a speaker's local use of fillers could lead to a listener's overall (global) impression of the speaker. In this work, we study how does a speaker's use of fillers relate to the incoming message from the speaker, and consequently, how does that relate to a listener's perception of the speaker.

We do so by studying a dataset of publicly available English monologue movie reviews, where the speakers voluntarily and naturally recorded themselves giving a movie review. These video reviews were collected from a social media platform that was created for the purpose of enabling speakers to upload their unbiased opinions towards products (in this case, a movie) to a large, but unseen audience. Annotators (listeners) were asked to label the reviews for attributes such as "confidence"; without explicitly being told to pay attention to the speaker's use of fillers. Our findings suggest that speakers stylistically do tend to use fillers in the incoming message, when introducing a new entity (to indicate new information), rather than an entity already introduced into the discourse. Our results also suggest that the occurrence of fillers before new entities may not have an effect on the listener's perception of the speaker's expressed confidence, despite previous works that suggest the link between fillers and expressed confidence. This does not discount other possible metacognitive aspects, such as the listener may expect a speaker to use fillers typically when the speaker is introducing new information in the incoming message. The rest of the paper is organised as follows: in section 2, we overview the theoretical foundations and research questions of our study, section 3 describes the dataset, section 4, the methodology, section 5, the results and discussions of the work, and section 6, the conclusion.

2 Background and Research Questions

2.1 Metacognition and the listener's perspective

When a speaker says an utterance, this articulation process includes an estimation of their commitment/ certainty about what they are saying. Research suggests that fillers and prosodic cues are linked to a speaker's metacognitive state, specifically; their *Feeling of Knowing (FOK)* or **expressed confidence** — a speaker's certainty or commitment to a statement (Smith and Clark, 1993). A speaker may *encode* meaning into their utterance using fillers, but the onus is on the listener to *decode* this information; making the interpretation of fillers contextual and dependent on the listener. Brennan and Williams (1995) observed that fillers and prosodic cues contribute to the listener's perception of the speaker's metacognitive state; which they refer to as the *Feeling of Another's Knowing (FOAK)*.

Other studies also focus on the comprehension of disfluent speech, i.e. taking into account the listener's understanding of the speaker's disfluencies (Corley and Stewart, 2008), and not on why the disfluency itself was produced (Nicholson, 2007). For example, Vasilescu et al. (2010) observe that the French "euh" has both *disfluent* (signalling production difficulties of the speaker) and *fluent* (as a discourse marker – to bracket lexical units that may aid in listener comprehension) properties. Related to metacognition, research suggests that following fillers, listeners may expect a speaker to shift topics, as they carry information about larger topical units (Swerts, 1998), that the use of fillers biases listeners towards new referents rather than ones already introduced into the discourse (Arnold et al., 2004), relax listener's expectations when hearing an unpredictable word (Corley et al., 2007), and that listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle) (Barr and Seyfeddinipur, 2010). In the present paper, we focus on the listener's comprehension of disfluencies. As Corley and Stewart (2008) state, "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech". We analyse the speaker's use of fillers from the incoming message from a corpus of previously recorded speech, and then observe what effect this may have on the listener's perception.

Drawbacks of current works Corley and Stewart (2008) illustrate that the results observed in Brennan and Williams (1995) that link fillers to FOAK, could have been influenced by the listener's being asked explicitly to rate speaker confidence/certainty on the speaker's short answer to a question (which may have included a filler). While this effect has been observed in other scenarios, for e.g. in human-machine interaction (Wollermann et al., 2013), it was still based on single utterance responses. These studies were appropriately targeted towards a QA setting. In a similar line of reasoning, Schrank and Schuppler (2015) show the drawbacks in research on automatic uncertainty detection[1], due to the narrow range of question-answering (QA) datasets commonly utilised. In general, this shows that when listener's are asked

[1]Which among other features, can use prosodic cues and the presence of fillers

to evaluate a speaker's certainty on shorter utterances, it could direct the listener towards paying attention to the fillers used by the speaker. Moreover, perceived uncertainty of the speaker in *local* utterances could still lead to a different *global* impression. Thus, there is a lack of evidence to support this effect on more spontaneous speech datasets.

Recently, Dinkar et al. (2020a) found that in an unsupervised manner, fillers can indeed be a discriminative feature in the automatic prediction of a listener's impression of a speaker's confidence. These results empirically solidified an effect that was often assumed to be true (and indeed, fillers are sometimes interchangeably used with the term "hesitations" in certain works (Pickett, 2018; Corley and Stewart, 2008)). However, the study simply focused on the overall impression the listener had of the speaker, i.e. the global, and did not account for more fine-grained information shared by the speaker.

2.2 Research Questions and Hypothesis

While work such as in Dinkar et al. (2020a) is important as preliminary analysis, they do not account for how fillers locally interact with the rest of the message in a holistic way. Clark (1996); Clark and Fox Tree (2002) proposed that speakers are able to utilise fillers as *collateral signals* in communication, in addition to the *primary signal* of the message. We colloquially refer to the primary signal of the message as *what* was said (in essence) and the collateral signal as *how* it was said. In Spoken Language Understanding (SLU), a similar phenomenon occurs of separating these two signals. However, in this context, reducing an input utterance into its primary signal (or *what* was said in essence) is standard practice (e.g. as seen in Tur and De Mori (2011), chapter 13. Speech Summarization). Indeed, in dialogue systems, the output transcripts of automatic speech recognisers are often cleaned of disfluencies such as fillers in post-processing, despite work relevant to the area that shows for e.g. the link between fillers and opinions (Le Grezause, 2017; Levow et al., 2014; Dinkar et al., 2020a), or the rich linguistic literature to suggest otherwise (Clark and Fox Tree, 2002). And yet, even recent work such as Barr and Seyfeddinipur (2010) support the collateral signal account, specifically that the listener is able to process fillers as a collateral signal (even if unclear whether the speaker (un)intentionally used them as such). This is an important finding, as it shows that perhaps the listener's attention is drawn to the cognitive state of the speaker. The problem then, as stated in Clark and Fox Tree (2002), remains about how to merge the two signals. Given the rapid advancements of dialogue systems, and growing interest in SLU, there is a need to move towards an automatic but holistic analysis of both together; if we hope to move towards better models and understanding of spontaneous speech. Thus the research questions are as follows:

RQ1: (Local effect of fillers): How does a speaker's use of fillers relate to the incoming message from the speaker? From the findings of Barr and Seyfeddinipur (2010); Arnold et al. (2004) as discussed in section 2, we would like to empirically analyse the role fillers play in a dataset of spontaneous speech, specifically related to new information from the incoming message of the speaker. Since the dataset we choose to study is a dataset of English monologue movie review videos (please refer to section 3), we consider the speaker's mention of terms related to the movie annotated from metadata, such as actors and directors.

- **H1** Fillers are more likely to occur before the introduction of new and upcoming information in the review.

RQ2: (Global effect of fillers): How does the speaker's use of fillers relate to a listener's perception of the speaker? We would like to empirically analyse whether the speaker's use of fillers has an impact on the listener's overall impression of the speaker.

- **H2** From H1, the speaker's use of fillers preceding new information in the incoming message contributes to the listener's perception of the speaker's confidence.

  Specifically, we hypothesise that when fillers are predominantly used in the context of preceding new information, listener's may judge the expressed confidence of the speaker as high, and listeners may only notice when fillers are used in other contexts (for e.g. as seen in Tottie (2014), listeners notice fillers when they are overused or used in the wrong context) which consequently will decrease the expressed confidence rating.

3 Materials

Persuasive Opinion mining (POM) dataset For this work, we choose the POM dataset (Park et al., 2014), a dataset of 1000 (American) English monologue movie review videos. Speakers recorded themselves (video and audio) giving a movie review, which they rated from 1 star (most negative) to 5 stars (most positive). The movie review videos are freely available on ExpoTV.com, and are completely in the wild; speakers were simply reviewing a movie without the knowledge that their review would eventually be annotated for such a context. 3 annotators (or listeners) per video were then asked to label the movie reviews for high level attributes, such as confidence. We think this dataset is particularly relevant for the following reasons: 1. Since this is a dataset of monologues, it allows us to focus uniquely on the role of fillers (Swerts, 1998). This is because the speaker is conscious of an *unseen* listener, but is not interrupted by the listener with other dialogue related disfluencies, such as backchannels ("Uh-huh"). This also minimises some turn-taking properties of fillers, such as when they are used to hold the speaker turn. Additionally, the annotators were never asked to pay special attention to the speaker's use of fillers. 2. Filler annotations of "uh" and "um" have been manually transcribed. Each transcription of a movie review video was reviewed by experienced transcribers for accuracy after being transcribed via Amazon Mechanical Turk (AMT) (Park et al., 2014). The experience of the transcriber is important, as Zayats et al. (2019) shows that transcribers tend to misperceive disfluencies and indeed, this can affect the transcription of fillers (Le Grezause, 2017). The filler count of this dataset is high (roughly 4% of the transcriptions, for comparison, the Switchboard (Godfrey et al., 1992) dataset of human-human dialogues, consists of $\approx$ 1.6% of fillers (Shriberg, 2001)). Sentence markers have been manually transcribed, with the practice of the filler being annotated sentence-initially, if the filler occurs between sentences (in this dataset, utterance segmentation is not available, and is interchangeable with sentence). 3. The inter-annotator agreement for several attributes is high; with confidence (which we use to denote the *FOAK*, or the listener's perception of the speaker's expressed confidence) (Krippendorff's alpha = 0.73), (Park et al., 2014). For confidence annotators were asked "How confident was the reviewer", and had to rate the speaker on a Likert scale of 1-7 with given labels: 1 (not confident), 3 (a little confident), 5 (confident) and 7 (very confident). Additional details can be found in Park et al. (2014). Summary statistics, that have been taken from Dinkar et al. (2020b), are given in Table 1.

Description	Value
Reviews that contain fillers	**792**
Total number of review used	892
Total *um* fillers in the corpus	4969
Total *uh* fillers in the corpus	4967
Total fillers in the corpus	**9936**
Number of tokens in the corpus	230462
% of tokens that are fillers	**4.31**
Average length (in tokens) of a review	255.9

Table 1: Details about the POM dataset.

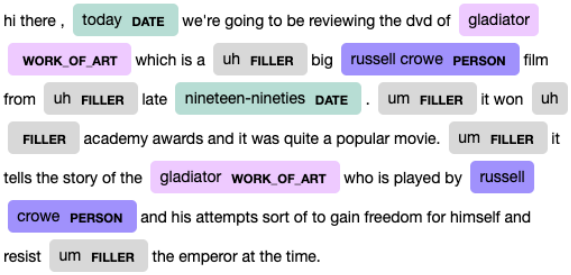

Figure 1: An example transcript that has annotated entities (in colour) using the EntityRuler. As shown, patterns from the metadata (e.g. "russell crowe") are added to the existing set (e.g. "nineteen-nineties"). Fillers are marked in grey. The first mention of "russell crowe" would be considered a new entity mentioned, while the second, an old one. Note, while the entity annotation is fairly reliable given the metadata, it is not exact. For e.g. the EntityRuler sometimes mislabels entities (the second mention of the word "gladiator".)

4 Methodology

4.1 RQ1 How does a speaker's use of fillers relate to the incoming message from the speaker?

H1 Fillers are more likely to occur before the introduction of new information in the review.

We consider the speaker's mention of entities related to the movie, that we extract from metadata files[2]. These entities could be categorised into actor, director or title of the movie. We then add these custom entities to SpaCy's EntityRuler, a rule

[2]The complete code and processed data will be made available online for reproducibility here https://github.com/tdinkar/fillers_in_POM.git

based named entity recogniser[3]. We prepossess the files (e.g. so that the filler annotations match the fillers in the existing model's vocabulary). We map the entities to match the existing patterns in the EntityRuler, for e.g "actor" is converted to "PERSON", by adding to the already existing entity patterns (please refer to Figure 1). The tagging of entities follows the *BIO* format (beginning, inside and outside of an entity).

To investigate H1, we inspect for each transcript, the distribution of filler positions, in relation to the automatically annotated entities in the discourse (denoted by Ent). We split these entities into Ent_new; i.e. entities newly introduced in the discourse, to indicate new information in the incoming message, and Ent_old to indicate entities already introduced in the discourse. We specifically note the order of the tokens in the transcripts for the filler positions and the first token of the 1. Ent_new (the first occurrence of the Ent) and 2. Ent_old (the second and following occurrences of each Ent), using the *B* tag of the Ent. Then, we check whether the distributions of filler positions (by its token position in the transcript) are significantly different compared to the distributions of 1. Ent_new and 2. Ent_old positions (by its first token's position), by utilising a Kruskal-Wallis H test[4] and use the Benjamini-Hochberg procedure for multiple testing correction. We then estimate the effect size by computing Cliff's Delta δ[5]. Lastly, we compare the δ distributions of the two experiments, i.e. fillers with Ent_new versus fillers with Ent_old using a Wilcoxon signed-rank test, to see if they significantly differ.

4.2 RQ2 How does the speaker's use of fillers relate to a listener's perception of the speaker and review?

H2 From H1, the speaker's use of fillers preceding new information contributes to the listener's perception of the speaker's confidence.

To investigate H2, we take the mean of the three confidence labels provided by the three annotators as the final rating of the speaker giving the review. We then consider reviews that are categorised as high-confidence (HC) and low-confidence (LC). Since confidence ratings are positively skewed[6] we take ratings of 3 (a little confident) and below to denote LC speakers, and 6 and above to denote HC speakers. The resulting size of the categories are 130 HC and 116 LC speakers. To calculate the percentage of fillers preceding new information (denoted by a new entity), we first consider the Ent_new labels that were automatically annotated in H1. We then count the number of fillers in the review that occur before (but not after) an Ent_new, constrained to a maximum distance of 1 token in between the filler and Ent_new. We normalise by dividing this count by the total number of fillers used in the review. From this, we obtain the percentage of fillers that occur before an Ent_new versus the percentage of fillers used in any other context that is not Ent_new. We then sum these two values for all HC and LC reviews, to get a cumulative percentage (please see Table 2).

Table 2: OR contingency table, where NE stands for the cumulative percentage of fillers that occur preceding an Ent_new for all HC (a) / LC (b) reviews, and OC the remaining cumulative percentage of fillers used in other contexts ((c) and (d) respectively).

		Outcome	
		HC	LC
Exposure	NE	a	b
	OC	c	d

We compute Odds Ratios (ORs) in order to investigate whether the use of fillers around new entities is associated with confidence. Odds ratios are an association measure that represents the odds that an outcome will occur given a particular exposure, compared to the odds that the outcome will occur in the absence of that exposure. Here, the odds denote the outcome of HC or LC, given the occurrence of fillers before new entities, compared to the occurrence of fillers that do not occur before new entities. We expect that the more fillers are used in the context of preceding new entities, the greater the odds of HC.

$$OR = \frac{odds_{HC}}{odds_{LC}}$$

where $odds_{HC} = a/c$ and similarly $odds_{LC} =$

[3] https://spacy.io/api/entityruler

[4] We utilise this method according to the guidelines given in the scipy software (https://scipy.org/) where the test is only run if the samples for each category ≥ 5. We calculate Cliff's delta regardless of this criteria.

[5] Utilising effect size tools from https://github.com/ACCLAB/DABEST-python/blob/master/dabest/_stats_tools/effsize.py

[6] This is shown both in the annotation guidelines as discussed in section 3, and the ratings itself, as annotator's may have hesitated to rate the speaker 1 (not confident). and preferred instead to use the label 3 (a little confident).

Table 3: Results of the Kruskal-Wallis H test, to compare the distributions of filler positions (by its token position in the transcript) compared to Ent_new/Ent_old positions, where "corrected" indicates the p-value after the Benjamini-Hochberg procedure. Note: Each cell indicates the number of reviews

	$p > .05$	$p \leq .05$
Ent_new	322	59
Ent_new corrected	381	0
Ent_old	477	70
Ent_old corrected	547	0

b/d using Table 2 for reference.

5 Results and Discussion

5.1 RQ1 How does a speaker's use of fillers relate to the incoming message from the speaker?

H1 Fillers are more likely to occur before the introduction of new information in the review.

Results for H1 are given in Table 3 for the Kruskal-Wallis H test, to compare the distributions of filler positions compared to 1. Ent_new and 2. Ent_old positions. By Kruskal-Wallis H test the distributions are significantly different for $\approx 15-20\%$ of the reviews (where $p \leq .05$). However, after utilising the Benjamini-Hochberg procedure for multiple testing correction, the distributions using this method do not significantly differ. This test is calculated using the sum of the ranks of each distribution. Given that the average review length is short ($\approx$ 256 tokens), and considering the close average median of fillers, Ent_new and Ent_old as given in Table 4, on reflection, this test may not capture nuances of the positional effects of fillers. We further discuss the limitations in section 7.

While significance testing focuses on a dichotomous result (i.e. significant versus not), we utilise Cliff's Delta δ to gain further insight into the magnitude of the effect. To interpret the results, Cliff's Delta δ ranges from -1 to 1, where 0 would indicate that the group distributions overlap completely; whereas values of -1 and 1 indicate a complete absence of overlap with the groups. For e.g. in H1 Ent_new, -1 indicates that all fillers in the review occur before new entities, and 1 indicates that all fillers in the review occur after new entities. This means that the smaller the effect size (close to zero) the larger the overlap, and the larger the effect size, the smaller the overlap.

By computing δ to estimate effect sizes as given in Figure 2, we see that for most reviews, fillers do occur visibly before Ent_new (median = -0.30 , $SD = 0.41$), but not before Ent_old (median = 0.20, $SD = 0.37$, given in Table 4), where the distributions of the δ values significantly differ ($Z = 27578.0$, $p < .05$ using Wilcoxon signed rank test). We see further evidence for this in Table 5[7], where majority of the reviews (565) have fillers occurring before Ent_new (sum of "nLarge" to "nSmall" δ sizes), compared to 163 reviews that had negligible effect size, and 139 reviews that had positive effect size (reviews that had fillers occurring after the introduction of new entities). We see the opposite δ effect sizes for Ent_old, where most of the reviews have fillers occurring after entities already introduced in the discourse (with predominately positive δ values as shown in Table 5), but not before. Fillers occurring after Ent_old is entirely plausible given that new entities can occur throughout the review, and not just at the start of one (as shown in Table 4, where the average median of Ent_new is roughly the same as Ent_old). Given the larger group with negligible effect size (247) for Ent_old, this does show that speakers may sometimes use fillers when repeating entities already introduced into the discourse. Dinkar et al. (2020a) used a language model (LM) trained on spontaneous speech to observe the probability of a filler appearing at a certain position; and found that the learnt word distribution shows that the LM places fillers predominantly at the start of sentences. However, sentence boundary annotation is dependent on the perspective of the transcriber, which in turn is certainly based on the presence of prosodic cues and fillers itself. Our findings suggest that there is more nuance to the way speakers utilise fillers (and indeed, our methodology is agnostic to sentence boundaries) in spontaneous speech. Therefore, regarding H1, stylistically speakers do tend to use fillers in the incoming message when introducing a new entity rather than one already introduced[8] (whether intentionally or not remains an open question), and the positions of fillers with

[7] The magnitude of Cliff's Delta δ can be interpreted by using the thresholds from Romano et al. (2006), i.e. $|\delta| < 0.147$ "negligible", $|\delta| < 0.33$ "small", $|\delta| < 0.474$ "medium", and otherwise "large".

[8] and indeed, this is the case for a dataset of spontaneous speech.

Table 4: Average median and SD for Ent_new, Ent_old (by first token position) and Fillers, and median and SD for effect size of the two δ distributions respectively.

	Avg. Median	Avg. SD
Ent_new	66.32	88.21
Ent_old	67.84	156.91
Fillers	66.05	125.95
δEnt_new	-0.30	0.41
δEnt_old	0.20	0.37

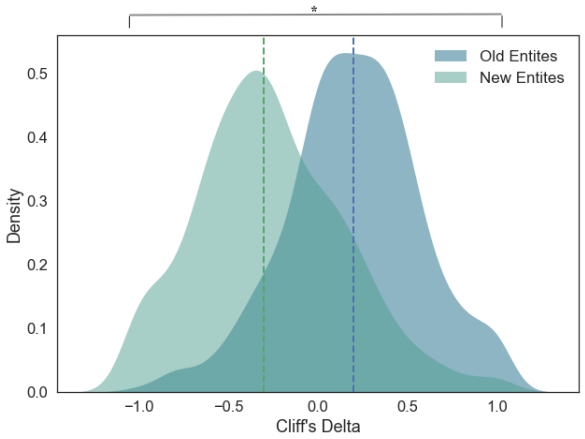

Figure 2: Distribution of Cliff's delta δ for fillers with Ent_new (New Entities) and fillers with Ent_old (Old Entites). Wilcoxon signed rank test has been performed to test whether the distributions significantly differ, with p $< .05$ given by $*$. The dotted line denotes the median (given in Table 4).

respect to Ent_new significantly differ from positions of fillers with respect to Ent_old.

5.2 RQ2 How does the speaker's use of fillers relate to a listener's perception of the speaker and review?

H2 From H1, the speaker's use of fillers preceding new information contributes to the listener's perception of the speaker's confidence.

To investigate the presence of fillers occurring before new information among confidence ratings, we computed ORs. To interpret the results, when $OR = 1$, the presence of the percentage of fillers that occur before new entities (exposure) does not affect the odds of neither HC nor LC (i.e. no association of the expo-sure with outcome). When $OR > 1$, the presence of the exposure is associated with higher odds of HC (positive association). When $OR < 1$, the presence of the exposure is associated with higher odds of LC (positive association with decrease of HC).

Table 5: Counts of Cliff's delta δ for fillers with Ent_new and fillers with Ent_old for all reviews, where the "n" or "p" before each row value indicates negative or positive values respectively.

	Ent_new	Ent_old
nLarge	277	36
nMedium	142	36
nSmall	146	66
Negligible	163	247
pSmall	62	156
pMedium	35	138
pLarge	42	189

The results of the test show $OR = 0.72$ ($p < .001$, 95% CI : 0.6-0.8)[9]. While $OR < 1$ in this case, indicating that the presence of fillers occurring before new entities gives a higher odds of LC, it is closer to 1, showing that the presence of the stimulus on the outcome is small. Interestingly, these findings are the opposite of what was hypothesised, which was that the speaker's use of fillers preceding new information contributes to the listener's perception of confidence; i.e. the more fillers are used in this way, the greater the odds of HC. According to the results of the *OR*s test, fillers occurring before new entities do not have a great effect on the odds of HC (only 28% lower given the presence of new entities) of the rating that the listener gives the speaker. This is consistent with the existing psycholinguistic literature on fillers as discussed in section 2. Arnold et al. (2004) for e.g. showed that fillers bias listeners towards new referents rather than ones already introduced into the discourse. In a study of the two fillers "um" and "uh" in American English, Tottie (2014) found that in natural conversation, listener's are not aware of the use of fillers, unless overused or used in the wrong context. Barr and Seyfeddinipur (2010) found that listener's expect the speaker to refer to something new following a filler (although they also found this to be specific to what was new for the speaker, and not only the listener), showing that listeners interpret fillers as delay signals, and infer plausible reasons for the delay by taking the speaker's perspective. While we cannot account for whether the annotator had rated the same speaker

[9]Risk Ratio $RR = 0.826$ with $p = .001$, 95% CI : 0.7-0.9

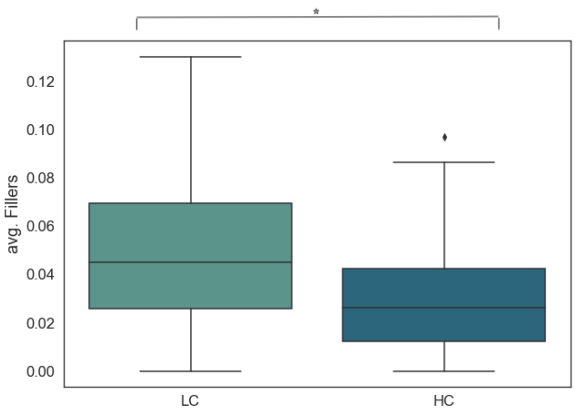

Figure 3: The speaker's average use of fillers (given by the percentage of fillers used compared to tokens in the review) with the categories of confidence using the divisions given in section 4 RQ2, $*$ denotes $p < .05$.

in multiple reviews, the annotator thus may expect the speaker to use fillers before new entities, or generally, before new expressions. This may not be considered usage in the "wrong context", and indeed, could simply indicate an increase in the number of entities in the review.

Looking at Figure 3, to show the average rate of fillers in the review (given by the percentage of fillers used compared to tokens in the review), it is clear that the use of fillers differs between HC and LC rated speakers (median filler rate of 0.026 and 0.045 respectively, with $U = 3873.0$ and $p < .05$ by Mann-Whitney U test). These results do not contradict Brennan and Williams (1995), i.e. there could be impressions formed by the listener about the speaker's expressed confidence based on fillers in spontaneous speech (as found in Dinkar et al. (2020a)). However, these results would suggest that the effect may not be from fillers used in the context of introducing new entities. This is an interesting finding; as fillers in these contexts may still have a metacognitive function as discussed above, but not necessarily related to FOAK. We cannot reject the null hypothesis, because there isn't sufficient evidence using our methodology to suggest that the occurrence of fillers before new entities has an effect on confidence (neither HC nor LC). Thus, these results suggest that fillers used in the context of introducing new entities in the discourse has little effect on the listener's rating of confidence that they attribute to the speaker.

6 Conclusion

The aim of this study was to empirically study on a real-life dataset, whether the utterance level use of fillers can help in understanding/ interpreting the perception of the speaker that was formed by a listener. We do so by studying a dataset of publicly available English monologue movie reviews, where the speakers voluntarily and naturally recorded themselves giving a movie review. Our findings show that speakers generally do tend to use fillers in the incoming message when introducing a new entity, rather than an entity already introduced into the discourse. Our results also suggest that the occurrence of fillers before new entities may not have an effect on the listener's perception of the speaker's expressed confidence, despite previous research to suggest otherwise (although these findings were validated in a different QA context). **Thus, local hesitations need not always lead to global impressions of uncertainty.** To the best of our knowledge, we are the first to contribute an in depth study of fillers accounting for hierarchical levels of analysis, i.e the sentence level and discourse level on real life data. In the *perspective taking account* of language comprehension as discussed in Barr and Seyfeddinipur (2010); the listener might be drawn to the mind of the speaker and infer possible reasons for delays in speech. Our analysis shows the possibility of different metacognitive functions in this perspective taking account that are brought about by the use of fillers on the listener. We hope that by using real-life data (reviews are available on ExpoTV.com, a social media platform where speakers can directly upload to an (unseen) audience videos of themselves giving an unbiased review), this study will both contribute to and encourage research on fillers in SLU.

7 Limitations

Our study is constrained to a dataset of monologues as mentioned in section 3. However, fillers can be used differently by the speaker (and consequently, processed differently by the listener) in dialogues. Furthermore, when considering the use of fillers, an important aspect is the acoustic information – as fillers are ubiquitous to spontaneous speech. While our measures focus on the transcripts and use ranking, it loses this temporal information, for e.g. distances in time, durations of fillers etc. However, it is difficult to calculate H1 in terms of time (rather than position), due to the poor results of the

forced alignment algorithms on this dataset. Since speaker's recorded themselves voluntarily and naturally using their own equipment, it is hardly surprising that the audio data is noisy. However, considering that SLU is often done on the output transcripts of ASR without considering acoustic information (except for the purposes of speech recognition), we consider these results as a preliminary analysis towards integrating fillers for SLU tasks.

References

Jennifer E Arnold, Michael K Tanenhaus, Rebecca J Altmann, and Maria Fagnano. 2004. The old and thee, uh, new: Disfluency and reference resolution. *Psychological science*, 15(9):578–582.

Dale J Barr and Mandana Seyfeddinipur. 2010. The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, 25(4):441–455.

Susan E Brennan and Maurice Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3):383–398.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

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

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.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020a. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online. Association for Computational Linguistics.

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, and Chloé Clavel. 2020b. How confident are you? exploring the role of fillers in the automatic prediction of a speaker's confidence. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8104–8108. IEEE.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520. IEEE.

Esther Le Grezause. 2017. *Um and Uh, and the expression of stance in conversational speech*. Ph.D. thesis.

Gina-Anne Levow, Valerie Freeman, Alena Hrynkevich, Mari Ostendorf, Richard Wright, Julian Chan, Yi Luan, and Trang Tran. 2014. Recognition of stance strength and polarity in spontaneous speech. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 236–241. IEEE.

Hannele Buffy Marie Nicholson. 2007. Disfluency in dialogue: attention, structure and function.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI 2014, page 50–57, New York, NY, USA. Association for Computing Machinery.

Joseph P Pickett. 2018. *The American heritage dictionary of the English language*. Houghton Mifflin Harcourt.

Jeanine Romano, Jeffrey D Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the nsse and other surveys? In *annual meeting of the Florida Association of Institutional Research*, volume 177.

Tobias Schrank and Barbara Schuppler. 2015. Automatic detection of uncertainty in spontaneous german dialogue. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Elizabeth Shriberg. 2001. To 'errrr' is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Vicki L Smith and Herbert H Clark. 1993. On the course of answering questions. *Journal of memory and language*, 32(1):25–38.

Marc Swerts. 1998. Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, 30(4):485 – 496.

Gunnel Tottie. 2014. On the use of uh and um in american english. *Functions of Language*, 21(1):6–29.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Ioana Vasilescu, Sophie Rosset, and Martine Adda-Decker. 2010. On the functions of the vocalic hesitation euh in interactive man-machine question answering dialogs in french. In *DiSS-LPSS Joint Workshop 2010*.

Charlotte Wollermann, Eva Lasarcyk, Ulrich Schade, and Bernhard Schröder. 2013. Disfluencies and Uncertainty Perception-Evidence from a Human-Machine Scenario. In *Sixth Workshop on Disfluency in Spontaneous Speech (DISS)*.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *arXiv preprint arXiv:1904.04398*.