Filled pauses in User-generated Content are Words with Extra-propositional Meaning

Ines Rehbein
SFB 632 “Information Structure”
Potsdam University
irehbein@uni-potsdam.de

Abstract

In this paper, we present a corpus study investigating the use of the fillers *äh* (uh) and *ähm* (uhm) in informal spoken German youth language and in written text from social media. Our study shows that filled pauses occur in both corpora as markers of hesitations, corrections, repetitions and unfinished sentences, and that the form as well as the type of the fillers are distributed similarly in both registers. We present an analysis of fillers in written microblogs, illustrating that *äh* and *ähm* are used intentionally and can add a subtext to the message that is understandable to both author and reader. We thus argue that filled pauses in user-generated content from social media are words with extra-propositional meaning.

1 Introduction

In spoken communication, we can find a high number of utterences that are disfluent, i.e. that include hesitations, repairs, repetitions etc. Shriberg (1994) estimates the ratio of disfluent sentences in spontaneous human-human communication to be in the range of 5-6%.

One particular type of disfluencies are filled pauses (FP) like *äh* (uh) and *ähm* (uhm). FP are a frequent phenomenon in human communication and can have multiple functions. They can be put at any position in an utterance and are used when a speaker encounters planning and word-finding problems (Maclay and Osgood, 1959; Arnold et al., 2003; Goffman, 1981; Levelt, 1983; Clark, 1996; Barr, 2001; Clark and Fox Tree, 2002), or as strategic devices, e.g. as floor-holders or turn-taking signals (Maclay and Osgood, 1959; Rochester, 1973; Beattie, 1983). Filled pauses can function as discourse-structuring devices, but they can also express extra-propositional aspects of meaning beyond the propositional content of the utterance, e.g. as markers of uncertainty or politeness (Fischer, 2000; Barr, 2001; Arnold et al., 2003).

Examples (1)-(6) illustrate the use of FP to mark repetitions (1), repairs (2), breaks (3) and hesitations (4) (the last one often used to bridge word finding problems). FPs can also express astonishment (5), excitement or negative sentiment (6). Extralinguistic reasons also come into play, such as the lack of concentration due to fatigue or distraction, which might lead to a higher ratio of FP in the discourse.

(1) I will *uh* I will come tomorrow.
(2) I will leave on Sat *uh* on Sunday.
(3) I think I *uh* have you seen my wallet?
(4) I have met Sarah and Peter and *uhm* Lara.
(5) Sarah is Michael’s sister. *Uh*? Really?
(6) A: He cheated on her. B: *Ugh*! That’s bad!

The role of fillers in spoken language has been discussed in the literature (for an overview, see Corley and Stewart (2008)). Despite this, work on processing disfluencies in NLP has mostly considered them as mere performance phenomena and focused on disfluency detection to improve automatic processing (Charniak and Johnson, 2001; Johnson and Charniak, 2004; Qian and Liu, 2013; Rasooli and

Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015), pages 12–21,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

Tetreault, 2013; Rasooli and Tetreault, 2014). Far fewer studies have focused on the information that disfluencies contribute to the overall meaning of the utterance. An exception are Womack et al. (2012) who consider disfluencies as extra-propositional indicators of cognitive processing.

In this paper, we take a similar stand and present a study that investigates the use of filled pauses in informal spoken German youth language and in written, but conceptually oral text from social media, namely Twitter microblogs.[1] We compare the use of FP in computer-mediated communication (CMC) to that in spoken language, and present quantitative and qualitative results from a corpus study showing similarities as well as differences between FP in both the spoken and written register. Based on our findings, we argue that filled pauses in CMC are words with extra-propositional meaning.

The paper is structured as follows. Section 2 gives an overview on the different properties of spoken language and written microblogs. In section 3 we present the data used in our study and describe the annotation scheme. Section 4 reports our quantitative results which we discuss in section 5. We complement our results with a qualitative analysis in section 6, and conclude in section 7.

2 Filled Pauses in Spoken and Written Registers

Clark and Fox Tree (2002) propose that FP are *words with meaning*, but so far there is no conclusive evidence to prove this. While experimental results have shown that disfluencies do affect the comprehension process (Brennan and Schober, 2001; Arnold et al., 2003), this is no proof that listeners have access to the *meaning* of a FP during language comprehension but could also mean that FP are produced "unintentionally [...], but at predictable junctures, and listeners are sensitive to these accidental patterns of occurrence." (Corley and Stewart, 2008), p.12.

To show that fillers are words in a linguistic sense, i.e. lexical units that have a specific semantics that is understandable to both speaker and hearer, one would have to show that speakers are able to produce them intentionally and that recipients are able to interpret the intended meaning of a filler.

Assuming that fillers are not linguistic words but simply noise in the signal, caused by the high demands on cognitive processing in spoken online communication, we would not expect to find them in medially written communication such as user-generated content from social media, where the production setting does not put the same time pressure on the user as there is in oral face-to-face communication. However, a search for fillers on Twitter[2] easily proves this wrong, yielding many examples for the use of FP in medially written text (7).

(7) Oh **uh**.. I got into the evolve beta.. yet I have no idea what this game is.. **uhm**..

Both, informal spoken dialogues and microblogs can be described as *conceptually oral*, meaning that both display a high degree of interactivity, signalled by the use of backchannel signals and question tags, and are highly informal with grammatical features that deviate from the ones in the written standard variety (e.g. violations of word order constraints, case marking, etc.). Both registers show a high degree of expressivity, e.g. interjections and exclamatives, and make use of extra-linguistic features (spoken language: gestures, mimics, voice modulation; microtext: emoticons, hashtags, use of uppercased words for emphasis, and more).

Differences between the two registers concern the spatio-temporal setting of the interaction. While spoken language is synchronous and takes place in a face-to-face setting, microblogging usually involves a spatial distance between users and is typically asynchronous, but also allows users to have a *quasi-synchronous* conversation.[3] Quasi-synchronous here means that it is possible to communicate in real time where both (or all) communicating partners are online at the same time, tweeting and re-tweeting in quick succession, but without the need for turn-taking devices as there is a strict first-come-first-serve order for the transmission of the dialogue turns. As a result, microblogging does not put the same time pressure on the user but permits them to monitor and edit the text. This should rule out the use of FP as markers of disfluencies such

[1] See the model of medial and conceptual orality and literacy by Koch & Oesterreicher (1985).

[2] `https://twitter.com/search-home`

[3] See (Dürscheid, 2003; Jucker and Dürscheid, 2012) for an account of *quasi-synchronicity* in online chatrooms.

as repairs, repetitions or word finding problems, and also the use of FP as strategic devices to negotiate who takes the next turn. Accordingly, we would not expect to observe any fillers in written microblogs if their only functions were the ones specified above.

However, regardless of the limited space for tweets,[4] microbloggers make use FP in microtext. This suggests that FP do indeed serve an important communicative function, with a semantics that must be accessible to both the blogger and the recipient.

3 Annotation Experiment

This section describes the data and setup used in our annotation experiment.

3.1 Data

The data we use in our study comes from two different sources. For spoken language, we use the KiezDeutsch-Korpus (KiDKo) (Wiese et al., 2012), a corpus of self-recordings of every-day conversations between adolescents from urban areas. All informants are native speakers of German. The corpus contains spontaneous, highly informal peer group dialogues of adolescents from multiethnic Berlin-Kreuzberg (around 266,000 tokens excluding punctuation) and a supplementary corpus with adolescent speakers from monoethnic Berlin-Hellersdorf (around 111,000 tokens). On the normalisation layer where punctuation is included, the token counts add up to around 359,000 tokens (main corpus) and 149,000 tokens (supplementary corpus).

The first release of KiDKo (Rehbein et al., 2014) includes the transcriptions (aligned with the audio files), a normalisation layer, and a layer with part-of-speech (POS) annotations as well as non-verbal descriptions and the translation of Turkish code-switching.

The data was transcribed using an adapted version of the transcription inventory GAT 2 (Selting et al., 1998), also called GAT minimal transcript, which uses uppercased letters to encode the primary accent and hyphens in round brackets to mark silent pauses of varying length.

The microblogging data consists of German-language Twitter messages from different regions of Germany, and includes 7,311,960 tweets with 105,074,399 tokens. For retrieving the tweets we used the Twitter Search API[5] which allows one to specify the user's location by giving a latitude and a longitude pair as parameters for the search. Over a time period of 6 months we collected tweets from 48 different locations.[6] The corpus was automatically augmented with a tokenisation layer and POS tags.[7]

	KiDKo		Twitter	
äh	646	35.8	6403	0.6
ähm	360	19.9	4182	0.4
both	1,006	55.7	10,585	1.0
# tokens	180,558	10,000	105,074,399	10,000

Table 1: Distribution of *äh* and *ähm* in KiDKo and Twitter microtext (raw counts (grey column) and normalised numbers (white column) per 10,000 tokens).

A string search in both corpora, looking for variants of *äh* and *ähm* (including upper- and lowercased spelling variants with multiple *ä*, with and without a *h*, and with one or more *m*) shows the following distribution (Table 1). Filled pauses are far less frequent in microblogs compared to spoken language, but due to the large amount of data we can easily extract more than 10,000 instances from the Twitter corpus. Note that the tweets in our corpus come from different registers like news, ads, public announcements, sports, and more, with only a small portion of private communication. When constraining the corpus search to the subsample of private tweets, we will most likely find a higher proportion of FP in the social media data.

In summary, we observe a higher amount of FP in spoken language than in Twitter microblogs. However, in both corpora variants of *äh* outnumber *ähm* by roughly the same factor. This observation is compatible with the results of (Womack et al., 2012) who report that around 60% of the FP in their corpus of English diagnostic medical narratives are nasal filled pauses (*uhm, hm*) and around 40% are non-nasal (*uh, er, ah*).

[4] The maximum length of a tweet is limited to 140 characters.

[5] https://dev.twitter.com/docs/api

[6] Note that the Twitter geoposition parameter can only approximate the regional origin of the speakers as the location where a tweet has been sent is not necessarily the residence or place of birth of the tweet author.

[7] Unfortunately, for legal reasons we are not allowed to distribute the data.

	Categories	Position
1	Repetition	**B/I**
2	Repair	**B/I**
3	Break	**B/I**
4	Hesitation	**B/I**
5	Question	**B/I**
6	Interjection	**B/I**
7	Unknown	

Table 2: Labels used for annotating the fillers (B: between utterances; I: integrated in the utterance).

	Twitter		KiDKo	
Sample	*äh*	*ähm*	*äh*	*ähm*
1	n.a.	n.a.	0.79	0.75
2	n.a.	0.84	0.73	0.64
3	0.80	0.83	0.78	0.84
4	0.87	0.87	0.78	0.75
5	0.86	0.86	0.74	n.a.
avg. κ	**0.84**	**0.85**	**0.76**	**0.75**

Table 3: Inter-annotator agreement (κ) for 3 annotators.

3.2 Annotating Fillers in Spoken Language and in Microtext

To be able to compare the use of fillers in spoken language with the one in Twitter microtext, we extract samples from the two corpora including 500 utterances/tweets with at least one use of *äh* and 500 tweets with at least one instance of *ähm*. At the time of the investigation, the transcription of KiDKo was not yet completed, and we only found 360 utterances including an *ähm* in the finished transcripts.

For annotation, we used the BRAT rapid annotation tool (Stenetorp et al., 2012). Our annotation scheme is shown in Table 2. We distinguish between different categories of fillers, namely between FP that mark repetitions, repairs, hesitations, or that occur at the end of an unfinished utterance/tweet (breaks). We also annotated variants of *äh* and *ähm* which were used as question tags or interjections, but do not consider them as part of the disfluency markers we are interested in. The Unknown label was used for instances which either do not belong to the filler class and shouldn't have been extracted, such as example (8), or which couldn't be disambiguated, usually due to missing context.

(8) Hääähähh !!!

Each filler is labelled with its *category* and *position*. By *position* we mean the position of the filler in the utterance or tweet. Here we distinguish between fillers which occur between (B) utterances/at the beginning or end of tweets (example 9b) and those which are integrated (I) in the utterance/tweet (9a). The numbers in the first column of Table 2 correspond to examples (1)-(6).

(9) a. das 's irgend so 'n **äh** (-) RAPper der ...
this 's some such a **uh** rapper who ...

this is some **uh** rapper who ... (Hesitation-I)

b. **äh** weiß ich nich
uh know I not

uh I don't know (Hesitation-B)

3.3 Inter-Annotator Agreement

The data was divided into subsamples of 100 utterances/tweets. Each sample was annotated by three annotators. Table 3 shows the inter-annotator agreement (Fleiss' κ) on the KiDKo and Twitter samples. We report agreement for all but three samples which we used to train the annotators, refine the guidelines and to discuss problems with the annotaton scheme. As we had only 360 instances of *ähm* from KiDKo, we divided them into three samples with 100 utterances and a fourth sample with 60 utterances.

Table 3 shows that the annotation of fillers is not an easy task. The disagreements in the annotations concern both the category and the position of the FP. In some cases the annotators agree on the label but disagree on the position of the filler (10a). This can be explained by the fact that spoken language (and sometimes also tweets) does not come with sentence boundaries, and it is often not clear where we should segment the utterance. In example (10a) two annotators interpreted the reparandum as part of the utterance and thus assigned REPAIR-I, while the third annotator analysed *am Samstag* (on Saturday) as a new utterance, resulting in the label REPAIR-B.

(10) a. SPK39 trifft sich am SONNtag mit
SPK39 meets REFL on-the Sunday with
den SPK23 **ÄH** am SAMStag
the SPK23 uh on-the Saturday
"SPK39 meets SPK23 on Sunday uh on Saturday"

b. wir HAM dann **ÄH** wir ham halbe stunde
we have then uh we have half hour
UNterricht
class
"then we have uh we have class for half an hour"

More often, however, the disagreements concern the category of the filler, as in (10b) where two annotators analysed the utterance as a repair while the third annotator interpreted it as a break followed by a new start. The results show that the annotation of fillers in KiDKo seems to be much harder, with average κ scores around 0.1 lower than for the tweets.

4 Quantitative Results

Table 4 shows that the ranking for the different categories of *äh* and *ähm* is the same in both corpora (11). Hesitations are the most frequent category marked by *äh* and *ähm*, followed by repairs and breaks. Repetitions are less frequent, especially in the written microblogs, as are *äh* and *ähm* as question tags and interjections.

(11) Hesitation > Repairs > Breaks > Repetitions > Questions/Interjections

äh/ähm	KiDKo		Twitter	
	#	%	#	%
Hesitations	557	64.78	759	72.91
Repairs	105	12.21	191	18.35
Breaks	88	10.23	52	0.05
Repetitions	53	6.16	9	0.01
Questions	10	1.16	6	0.01
Interjections	11	1.28	5	0.00
total	860=100%		1041=100%	

Table 4: Frequencies of *äh/ähm* in KiDKo and in Twitter (note: numbers don't add up to 100% because of *Unknown* cases).

However, we can also observe a substantial difference between the spoken and the written register. In the latter one, the two most frequent categories, hesitations and repairs, make up for more than 90% of all instances of *äh* and *ähm*, while in spoken language these two categories only account for 76-77% of all occurrences of the two fillers. A possible explanation is that breaks and repetitions in spoken language are either performance phenomena or caused by discourse strategies (e.g. floor-holding) which are both superfluous in asynchronous written communication. This still leaves us with the question why hesitations and repairs do occur in written text at all. We will come back to this question in section 6.

The next question we ask is whether the two forms, *äh* and *ähm*, are used interchangeably or whether the use of each form is correlated with its function. As shown in Table 5, hesitations and breaks are more often marked by *ähm* while *äh* occurs more frequently as a marker of repairs and repetitions. This observation holds for both the spoken and the written register. 72.8% and 80.0% of all instances of *ähm* occur in the context of a hesitation in KiDKo and Twitter, while only 59.0% (KiDKo) and 65.8% (Twitter) of the non-nasal fillers *äh* are used to mark a hesitation. A Fisher's exact test shows that for hesitations and repairs, the differences are statistically significant with $p < 0.01$ and $p < 0.05$, while for breaks and repetitions, the differences might be due to chance.

Next we look at the syntactic position where those fillers occur in the text. We would like to know how often FP are integrated in the utterance and how often they occur between utterances.

	KiDKo %		Twitter %	
	äh	*ähm*	*äh*	*ähm*
Hesitation	59.0	72.8	65.8	80.0
Break	9.0	11.9	4.5	5.5
Repair	16.1	5.8	25.4	11.5
Repetition	7.4	4.2	1.2	0.6

Table 5: Distribution of *äh* and *ähm* between different types of disfluencies.

		KiDKo %		Twitter %	
		B	I	B	I
äh	Hesitations	24.6	34.4	42.6	23.2
	Repairs	0.1	16.0	0.6	24.8
	Repetitions	0.0	7.4	0.2	1.0
ähm	Hesitations	31.4	41.4	62.4	17.6
	Repairs	0.0	5.8	0.4	11.1
	Repetitions	0.0	4.2	0.0	0.6

Table 6: Position of *äh* and *ähm* in correlation to their category.

Fox et al. (2010) present a cross-linguistic study on self-repair in English, German and Hebrew, and observe that self-corrections in English often include the repetition of whole clauses, i.e. English speakers "recycle" back to the subject pronouns (Fox et al. 2010:2491). In their German data this pattern was less frequent. Fox et al. (2010) conclude that morpho-syntactic differences between the languages have an influence on the self-repair practices in the speakers.

Our findings are consistent with Fox et al. (2010) in that we mostly observe the repetition of words, not of clauses (Table 6). Nearly all fillers which mark repetitions are integrated in the utterance or tweet, only a few occur between utterances/tweets. Fillers as markers of repairs are also mostly integrated.

For hesitations, the most frequent category, we get a more diverse picture. In our spoken language data, *äh* and *ähm* are more often integrated in the utterance, while for tweets FP as hesitation markers mostly appear at the beginning or end of the tweet.

So far, our quantitative investigation showed some striking similarities in the use of filled pauses in the two corpora. In both registers, the ranking of the different disfluency types marked by the FP were the same. Furthermore, we showed that speakers/users are sensitive to the surface form of a FP and prefer to use *äh* in repairs and *ähm* in hesitations, regardless of the medium they use for communication.

5 Discussion

In this section we will look at related work on FP and try to put our findings into context. Previous work on the difference between nasal and non-nasal fillers (Barr, 2001; Clark and Fox Tree, 2002) has described nasal fillers such as *uhm, hm* as indicators of a high cognitive load, while their non-nasal variants indicate a lower cognitive load during speech production. Clark and Fox Tree (2002) have proposed the *filler-as-word hypothesis*, stating that FP like *uh* and *uhm* are words in a linguistic sense with the basic meaning that a minor (*uh*) or major (*uhm*) delay in speaking is about to follow. This analysis is based on a corpus study showing that silent pauses following a nasal filler are longer than silent pauses after a non-nasal filler. Beyond the basic meaning, FP can have different implicatures, depending on the context they are used in, such as indicating that the speaker wants to keep the floor, is planning the next (part of the) utterance, or wants to cede the floor. To illustrate this, Clark and Fox Tree (2002) use *goodbye* which has the basic meaning "express farewell" but, when uttered while someone is approaching the speaker, can have the implicature "Go away".

We take the *filler-as-word hypothesis* of Clark and Fox Tree (2002) as our starting point and see how adequate it is to describe the use of FP in written microblogs (section 6). However, we try to avoid the term *implicature* which seems problematic in this context, as we are not dealing with implicatures built on regular lexical meanings but rather with implicatures on top of non-propositional meaning. As a side-effect, the implicatures based on filled pauses are not cancellable.

The analysis of Clark and Fox Tree (2002) is not uncontroversial (see, e.g., Womack et al. (2012) for a short discussion on that matter). O'Connell and Kowal (2005) criticise that the corpus study of Clark and Fox Tree (2002) is based on pause length as perceived by the annotators (instead of being analysed by means of acoustic measurements).

Furthermore, it might be possible that the semantics of FP to indicate the length of a following delay only applies to English. Belz and Klapi (2013) have measured pause lengths after nasal and non-nasal fillers in German L1 and L2 dialogues from a MAP task and could not find a similar correlation between filler type and pause length.

In summary, it is not clear whether the different findings are due to methodological issues, or might be particular to certain languages and text types. Shriberg (1994), p.130 suggests that for English, models of disfluencies based on the ATIS corpus, a corpus of task-oriented dialogues about air travel planning, might not be able to predict the behaviour of disfluencies in spoken language corpora with data recorded in a less restricted setting.

The MAP task corpora used in Belz and Klapi (2013), for example, includes dialogues where one speaker instructs another speaker to reproduce a route on a map. Due to the functional design, the content of the dialogues is constrained to solving the task at hand and thus the language is expected to differ from the one used in the London–Lund corpus (Svartvik, 1990), a corpus of personal communication, that was used by Clark and Fox Tree (2002).

Fox Tree (2001) presents a perception experiment showing that *uh* helps recognizing upcoming words, while the nasal *um* doesn't. In our study we found a strong correlation between the category of the filler and its form (nasal vs. non-nasal). Nasal fillers were mostly used in the context of hesitations, which is consistent with their ascribed basic function as indicators of longer pauses (Clark and Fox Tree, 2002). The tendency to use *äh* within repairs might be explained by Fox Tree (2001)'s findings that non-nasal fillers help to recognise the next word. Thus, we would expect a preference for non-nasal FP to be used as an interregnum before the repair.

Other evidence comes from Brennan and Schober (2001) who present experiments where the subjects had to follow instructions and select objects on a graphical display. They showed that insertions of *uh* after a mid-word interruption in the instruction helped the subjects to correctly identify the target object, as compared to the same instruction where the filler was replaced by a silent pause. They conclude that fillers help to recover from false information in repairs.[8]

So far, our findings are consistent with previous work outlined above, but do not rule out other explanations. A major argument against the analysis of FP as linguistic words is that so far there is no conclusive evidence that speakers do produce them intentionally (Corley and Stewart, 2008).

Our corpus study provides this evidence by showing that FP in CMC are produced deliberately and intentionally. Furthermore, we observed a statistically significant correlation between filler form (nasal or non-nasal) and filler category, which also points at *äh* and *ähm* being separate words with distinguishable meanings.

In the next section, we show that FP in CMC can add a subtext to the original message that can be understood by the recipients, and that the information they add goes beyond the contribution made by nonverbal channels such as facial expressions or gestures. We illustrate this, based on a qualitative analysis of our Twitter data.

6 Extra-propositional Meaning of FP in Social Media Text

New text from social media provides us with a good test case to investigate whether filled pauses are words with (extra-propositional) meaning, as the production of written text is to a far greater extent subject to self-monitoring processes. This means that we can confidently rule out that the use of fillers in tweets is due to performance problems caused by the time pressure of online communication. Another important point is that communication on Twitter is not synchronous but can be time-delayed and works on a first-come-first-serve basis. This is quite important, as it means that we can also exclude the discourse-strategic functions of FP (e.g. floor-holding and turn-taking) as possible explanations for the use of fillers in user-generated microtext.

We conclude that there have to be other explanations for the use of filled pauses as markers of hesitations and repairs in microblogs. Consider the following examples (12)-(14).

(12) Mein ... **ääh** Glückwunsch! RT
My ... **uh** congratulation! RT
@germanpsycho: Ich bin nun verheiratet.
@germanpsycho: I am now married.
"My ... **uh** congratulations! RT @germanpsycho: I'm married now."

(13) Die hat aber schöne **ähm** Augen.
This one has PTCL beautiful **uhm** eyes.
"This one has really beautiful **uhm** eyes."

(14) Ich frage für, **ähm**, einen Freund.
I ask for, **uhm**, a friend.
"I'm asking for **uhm** a friend."

[8] Unfortunately, they did not compare the effect of *uh* in repairs to the one obtained by a nasal filler like *um*.

The fillers in the examples above add a new layer of meaning to the tweet which results in an interpretation different from the one we get without the filler. While a simple "Congratulations!" as answer to the message "I'm married now" would be interpreted as a polite phrase, the mere addition of the filler implies that this tweet should not be taken at face value and has a subtext along the lines "Actually, I really feel sorry for you". The same is true for (13) where the subtext can be read as "In fact, we're talking about some other bodyparts here". In example (14), the subtext added by the filler will most probably be interpretated as "I'm really asking for myself but won't admit it".[9]

In the next examples (15)-(17), also hesitations, the filler is used to express the author's uncertainty about the proposition.

(15) 30000 € für die 2h db Show für regiotv...
30000 € for the 2h db show for regiotv...
ähm...? Ich weiss grad auch nich..
uhm...? I know just now also not..
"30000 € for the 2h db show for regiotv... **uhm**...? I don't know right now, either.."

(16) Tor für #Arminia durch, **ääh**, wir glauben
Goal for #Arminia by **uuh**, we believe
Schütz.
Schütz.
"Goal for #Arminia by **uuh**, we believe Schütz."

(17) @zinken **äh**.. so 98%
@zinken **uh**.. around 98%
"@zinken **uh**.. around 98%"

Thus, the most general commonality between the examples above is that the speaker does not make a commitment concerning the truth content of the message.

The following examples (18)-(21) show instances of *äh* and *ähm* in repairs where the FP occur as *interregnum* between *reparandum* and *repair*.[10]

$$\textit{I will leave you } \underbrace{\textit{on Sat}}_{\text{REPARANDUM}} \ \underbrace{\textit{uh}}_{\text{INTERREGNUM}} \ \underbrace{\textit{on Sunday}}_{\text{REPAIR}}$$

The tweet author enacts a slip of the tongue, either by using homonymous or near-homonymous words (Diskus (discus) – Discos (discos), hängst (hang) – Hengst (stallion)) or by using analogies and conventionalised expressions (off – on, resist – contradict). The "mistake" was made with humorous intention and is then corrected. The filler takes again the slot of the interregnum and serves as a marker of the intended pun.

(18) Ob Diskuswerfer früher immer in
Whether discus-throwers in the past always in
Diskus *äh* **Discos** geübt haben, etwa als
discus *uh* **discos** trained have, perhaps as
Rauswerfer am Eingang?
bouncers at the entrance?
"In the past, have discus-throwers always trained in **discus** *uh* **discos**, maybe as bouncers at the entrance?"

(19) Du **Hengst**! *äh*, **hängst**.
You **stallion**! *uh*, **hang**.
"You **stallion**! *uh*, **hang**."

(20) MacBook aus, Handy aus, TV aus. Buch **an**,
MacBook off, mobile off, TV off. Book **on**,
ähh, **aufgeklappt**.
uhh, **open**.
"MacBook off, mobile off, TV off. Book **on**, *uhh*, **open**."

(21) wer könnte Dir schon **widerstehen**, *ähm*, ich
who could you PTCL **resist**, *uhm*, I
meine **widersprechen**.
mean **contradict**.
"who could **resist** you, *uhm*, I mean **contradict**."

In the next set of examples, (22)-(24), a taboo word or word with a strong negative connotation is reformulated into something more socially acceptable (minister of propaganda → district mayor; madness → spirit; tantalise → educate). Often, this is done with a humorous intention, but also to express negative sentiment (e.g. in (22) towards Buschkowsky, or in (23) towards Apple).

(22) Exakt. Wie es das Buch von eurem
Exactly. How it the book of your
RMVP Minister Goebbels *äh*
RMVP minister Goebbels *uh*
Bezirksbürgermeister Buschkowsky so
district mayor Buschkowsky so
beschrieben hat. :-)
described has :-)

[9] In fact, this adds an interesting meta-level to the utterance, as by inserting the filler the author draws attention to the fact that there is something she seemingly wants to hide.

[10] We follow the terminology of Shriberg (1994).

“Exactly. Just as the book of your **minister of propaganda Goebbels** *uh* **district mayor Buschkowsky** has described :-)”

(23) Du hast den Apple **Wahnsinn**... *äh*, **Spirit**
You have the Apple **madness**... *uh*, **spirit**
einfach noch nicht verstanden ;)
simply still not understood ;)

“You haven’t yet understood the Apple **madness**... *uh* **spirit** ;)”

(24) ... ein bisserl Nachwuchs **quäl**... *ähm*
... a little bit new blood **tant**... *uhm*
ausbilden
educating

“... **tant[alising]** the new blood *uhm* **educating**”

These examples show that the use of *äh* and *ähm* in tweets is intentional and highly edited. The two forms are used to express the speaker’s uncertainty about the propositional content of the message, or as a signal that the speaker does not warrant the truth of the message. Other functions include the use of fillers as markers of humorous intentions and of negative sentiment (see Table 7). Note that the meanings are not necessarily distinct but often overlap.

We thus argue that FP in user-generated content from social media are linguistic words that are produced intentionally and have an extra-propositional meaning that can be understood by the recipients.

Meaning	**Description**
UNCERTAINTY	Speaker is uncertain about the propositional content
TRUTH CONTENT	Speaker does not warrant the truth content of the proposition
HUMOR	Marker of humorous intention
EVALUATION	Marker of negative sentiment

Table 7: Extra-propositional meaning of fillers in CMC.

7 Conclusions

The results from our corpus study show that fillers in user-generated text from social media are linguistic words that are produced intentionally and function as carriers of extra-propositional meaning.

This finding has consequences for work on Sentiment Analysis and Opinion Mining in social media text, as it shows that FP are used as a marker of irony and humour in Twitter, and also indicate uncertainty and negative sentiment. Thus, filled pauses might be useful features for irony detection, sentiment analysis, or to assess the strength of an opinion in online debates.

Acknowledgments

This research was supported by a grant from the German Research Association (DFG) awarded to SFB 632 “Information Structure”. I am grateful to Sören Schalowski and Josef Ruppenhofer for valuable comments and discussions that contributed to this work. I would also like to thank our annotators, Nadine Reinhold and Emiel Visser, and the anonymous reviewers for their thoughtful comments.

References

Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Dale J. Barr, 2001. *Trouble in mind: paralinguistic indices of effort and uncertainty in communication*, pages 597–600. Paris: L’Harmattan.

Geoff W. Beattie. 1983. *Talk: an analysis of speech and non-verbal behaviour in conversation*. Milton Keynes: Open University Press.

Malte Belz and Myriam Klapi. 2013. Pauses following fillers in L1 and L2 German Map Task dialogues. In *The 6th Workshop on Disfluency in Spontaneous Speech*, DiSS.

Susan E. Brennan and Michael F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44:274–296.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL.

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

Herbert H. Clark. 1996. *Using language*. Cambridge: Cambridge University Press.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2:589–602.

Christa Dürscheid. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für angewandte Linguistik*, 38:3756.

Kerstin Fischer. 2000. *From cognitive semantics to lexical pragmatics: the functional polysemy of discourse particles*. Mouton de Gruyter: Berlin, New York.

Barbara Fox, Yael Maschler, and Susanne Uhmann. 2010. A cross-linguistic study of self-repair: evidence from English, German and Hebrew. *Journal of Pragmatics*, 42:2487–2505.

Jean E. Fox Tree. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, 2(29):320–326.

Erving Goffman, 1981. *Radio talk*, pages 197–327. Philadelphia, PA: University of Pennsylvania Press.

Mark Johnson and Eugene Charniak. 2004. A tag-based noisy channel model of speech repairs. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL.

Andreas H. Jucker and Christa Dürscheid. 2012. The linguistics of keyboard-to-screen communication. A new terminological framework. *Linguistik Online*, 6(56):39–64.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Howard Maclay and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15:19–44.

Daniel C. O'Connell and Sabine Kowal. 2005. Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34(6):555–576.

Xian Qian and Yang Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT.

Sadegh Mohammad Rasooli and Joel Tetreault. 2013. Joint parsing and disfluency detection in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Sadegh Mohammad Rasooli and Joel Tetreault. 2014. Non-monotonic parsing of fluent umm i mean disfluent sentences. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Ines Rehbein, Sören Schalowski, and Heike Wiese. 2014. The KiezDeutsch Korpus (KiDKo) release 1.0. In *The 9th International Conference on Language Resources and Evaluation*, LREC.

Sherry R. Rochester. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1):51–81.

Margret Selting, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Uta Quasthoff, Christoph Meier, Peter Schlobinski, and Susanne Uhmannet. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California at Berkeley.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Jan Svartvik. 1990. *The London Corpus of Spoken English: Description and Research*. Lund: Lund University Press.

Heike Wiese, Ulrike Freywald, Sören Schalowski, and Katharina Mayr. 2012. Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 2(40):797–123.

Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Ex-ProM '12.