The Distribution of Editing Phrases in German, French and Chinese Dialogues

Ye Tian[1], **Julian Hough**[2],
David Schlangen[2], **Jonathan Ginzburg**[1]
[1]Laboratoire de linguistique formelle, Université Paris-Diderot,
[2]Dialogue Systems Group, Bielefeld University,

Abstract content

1. Introduction

In natural conversations, speakers frequently produce lexical and non-lexical filled pauses, both during hesitations, and within self-repairs. (Ginzburg et al., 2014) categorize disfluencies into ones that are *backwards looking* and ones that are *forwards looking*. Forward looking disfluencies are cases when an utterance is interrupted by a filled or silent pause, but are continued without an alteration. Backwards looking disfluencies are cases when an utterance is interrupted and replaced with an alteration that refers back to an already uttered reparandum, and an editing phrase (EP) is often inserted. we define EPs not by their lexical meaning, but by its structural context. Any ”words” used between a reparandum and its repair is considered an EP.

This paper introduces a multi-lingual natural dialogue corpus annotated for disfluency, and presents a preliminary results on the repertoire of filled pauses and EPs in three languages: French, Chinese and German.

2. Data and transcription

We use the DUEL corpus (Hough et al., 2016), consisting of 24 hours of natural, face-to-face, loosely task-directed dialogue in German, French and Mandarin Chinese. The corpus is uniquely positioned as a cross-linguistic, multimodal dialogue resource controlled for domain. DUEL includes audio, video and body tracking data and is transcribed and annotated for disfluency, laughter and exclamations. The data consists of 10 dyads per language.

Transcription was done from the WAV audio files using Praat (Boersma and Weenink, 2010), following the instructions of the DUEL transcription and annotation manual (Hough et al., 2015), which specifies language general practices such as segmentation, disfluency annotation and laughter annotation, as well as language specific instructions regarding filled pauses, exclamations, and non-standard orthography.

2.1. Editing Phrase and Repair Annotation

Our annotations follow the light-weight inline method of dialogue annotation described by Hough et al. (2015). We utilize the disfluencies marked up as EPs (a class which includes filled pauses).

The filled pauses are annotated by a {`F` }, bracketing other fillers and editing terms simply with { } - e.g. `I { you know } like her`.

The inventory of EPs and filled pauses differ depending on the language. For example, in German, the common filled pauses are {F äh}, {F ähm} and {F hm}; in French they are {F euh}, {F mmh} and {F euhm}; in Chinese, they are {F en}, {F eh}, as well as demonstratives {F nage} (literally “that”) and {F zhege} (literally “this”).

For repairs, restarts and abandoned utterances, we mark the structure according to this scheme, consistent with the Switchboard repair mark-up (Meteer et al., 1995):

(*reparandum* + { *EP* } *repair*)

3. The distribution of editing phrases across languages

3.1. Filled pauses

The distribution of filled pauses were summarized in tables 1,2 and 3. They are the most frequent in French, at 0.29 filled pauses per utterance, compared to 0.17 per utterance in Chinese and 0.13 per utterance in German. ”Non-lexical” vowel based filled pauses such as ”euh”, ”eh” and ”äh” are the most frequent filled pauses in all three languages. Certain ”discourse markers” have similar distributions as those ”non-lexical” filled pauses, e.g. ”bah” (an interjection) in French, ”ranhou” (”then”) in Chinese, and ”also” (”so”) in German.

Table 1: French filled pauses

Filler	Occurrences	Percentage
euh	4089	60%
bah	651	9%
hein	291	4%
genre	279	4%
tuvois	260	4%
Rmmh	255	4%
ah	248	4%
’fin	199	3%
euhm	192	3%
enfait	138	2%
bon	135	2%
ouais	128	2%

Table 2: Chinese filled pauses

Filler	Occurrences	Percentage
eh	1066	28%
ranhou (then)	578	15%
en	541	14%
jiushi (it is)	514	14%
nage (that)	483	13%
ah	304	8%
em	284	7%
zhege (this)	50	1%

Table 3: German filled pauses

Filler	Ocurrences	Percentage
äh	698	48%
ähm	413	29%
also	124	9%
Hm/mh/uhm	56	4%
Fäh/Fähm	38	3%
Oh	32	2%
ach	15	1%
achso	11	1%
ja	8	1%

3.2. Editing phrases

We extracted instances of disfluencies marked in the form of (reparandum + optional EP repair). They include repetitions and repairs. French and Chinese were similar in the rates of repetitions/ repairs. In French, 19% of utterances contain repetitions/repairs. There was a total of 3684 occurrences, on average 1.2 per utterance. In Chinese, 17% of utterances contain repetitions/ repairs. There was a total of 4476 occurrences, on average 1.15 per utterance. In contrast, only 8%of utterances in German contain repetitions/repairs. There was a total of 1125 occurrences, on average 1.2 per utterance.

In terms of EPs, French uses them more frequently than Chinese and German. 25% (French), 14% (Chinese) and 13% (German) of repetitions/ repairs used an EP. Both filled pauses and lexical items can be used as EPs. Few of the frequent lexical EPs contain in their meaning the sense of ”editing” or ”correction”. Tables 4, 5 and 6 summarize the distributions of EPs in three languages.

Table 4: French editing phrases

French	Occurrences	Percentage
euh	655	72%
ouais (yeah)	44	5%
genre (like)	42	5%
bah	29	3%
’fin (lastly)	23	3%
euhm	21	2%
enfin (lastly)	16	2%
voila	15	2%
tu vois (you see)	13	1%
bon (good)	8	1%
donc (so)	8	1%

Table 5: Chinese editing phrases

Editing phrase	Occurrences	Percentage
e	192	32%
jiushi (is)	130	21%
nage (that)	97	16%
ranhou (then)	77	13%
dui (correct)	30	5%
em	22	4%
non-verbal teeh noise ”tze”	16	3%
oh	12	2%
zhege (this)	11	2%
shenme (what)	9	1%
bushi/budui (no)	9	1%

Table 6: German editing phrases

Editing phrase	Occurrences	Percentage
äh	82	56%
ähm	30	20%
also	19	13%
ja	4	3%

4. Conclusion and future work

We analyzed filled pauses and editing phrases in a multilinguistic dialogue corpus DUEL. We found that different languages use filled pauses and EPs are different rates (more frequent in French than in Chinese and German). A repertoire of both ”non-lexical” and lexical filled pauses were used in all three languages, and most of these filled pause can also function as EPs.

References

Boersma, P. and Weenink, D. (2010). Praat: doing phonetics by computer.

Ginzburg, J., Fernández, R., and Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):64.

Hough, J., de Ruiter, L., Betz, S., and Schlangen, D. (2015). Disfluency and laughter annotation in a light-weight dialogue mark-up protocol. In *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*.

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Schlangen, D., and Ginzburg, J. (2016). DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In *10th edition of the Language Resources and Evaluation Conference*.

Meteer, M. W., Taylor, A. A., MacIntyre, R., and Iyer, R. (1995). *Disfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania.