Conversational Feedback in Scripted versus Spontaneous Dialogues: A Comparative Analysis

Ildikó Pilán[1], Laurent Prévot[2,3], Hendrik Buschmeier[4], and Pierre Lison[1]

[1] Norwegian Computing Center, Oslo, Norway
[2] CEFC, CNRS, MEAE, Taipei, Taiwan
[3] Aix Marseille Université & CNRS, LPL, Aix-en-Provence, France
[4] Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

{pilan,plison}@nr.no laurent.prevot@univ-amu.fr hbuschme@uni-bielefeld.de

Abstract

Scripted dialogues such as movie and TV subtitles constitute a widespread source of training data for conversational NLP models. However, there are notable linguistic differences between these dialogues and spontaneous interactions, especially regarding the occurrence of *communicative feedback* such as backchannels, acknowledgments, or clarification requests. This paper presents a quantitative analysis of such feedback phenomena in both subtitles and spontaneous conversations. Based on conversational data spanning eight languages and multiple genres, we extract lexical statistics, classifications from a dialogue act tagger, expert annotations and labels derived from a fine-tuned Large Language Model (LLM). Our main empirical findings are that (1) communicative feedback is markedly less frequent in subtitles than in spontaneous dialogues and (2) subtitles contain a higher proportion of negative feedback. We also show that dialogues generated by standard LLMs lie much closer to scripted dialogues than spontaneous interactions in terms of communicative feedback.

1 Introduction

While the amount of text data available for training or fine-tuning LLMs is large and growing steadily, spoken conversational data remains relatively scare. Although corpora of spontaneous spoken interactions have been collected for various languages (Dingemanse and Liesenfeld, 2022), those are generally of a modest size and limited to specific topics or tasks. Due to this scarcity of available data, a common approach for the development of conversational models is to rely on corpora of authored dialogues extracted from movie scripts (Danescu-Niculescu-Mizil and Lee, 2011) or movie and TV subtitles (Lison et al., 2018; Davies, 2021).

However, those dialogues are markedly different from spontaneous interactions. Most importantly, movie scripts and subtitles are explicitly written with the aim of *narrating a story*. Subtitles must also abide to strict length constraints, and thus tend to only transcribe the most salient part of each turn. As a consequence, many conversational phenomena such as disfluencies (Shriberg, 1996), overlapping talk (Schegloff, 2000), and backchannels (Yngve, 1970) are either absent or uncommon in those dialogues, unless their presence happens to contribute to the storyline (Berliner, 1999; Chepinchikj and Thompson, 2016).

This paper provides a quantitative analysis of how subtitles differ from spontaneous dialogues, focusing more specifically on *conversational feedback* (Allwood et al., 1992) and *grounding* (Clark and Schaefer, 1989) phenomena. To highlight differences in linguistic properties between subtitles and spontaneous conversation corpora, we first compile a range of lexical statistics and use a dialogue act tagger to estimate the relative frequencies of various feedback signals. To obtain more fine-grained estimates on three core feedback categories, respectively *Agreement / Acceptance*, *Acknowledgement / Backchannel* and *Negative Feedback*, we collect manual annotations on multiple dialogue samples and fine-tune a LLM on those annotations to automatically detect the presence of those feedback in our corpora. Finally, we apply the fine-tuned LLM on synthetic dialogues generated with standard autoregressive LLMs, and show that those dialogues are comparatively much closer to scripted dialogues than to spontaneous interactions when it comes to the frequency and type of conversational feedback. Those experiments are conducted for eight languages (English, Chinese, French, German, Hungarian, Italian, Japanese and Norwegian) for which corpora of spontaneous dialogues are readily available.

The paper is structured as follows. Section 2 reviews related work, and Section 3 presents the corpora employed in our experiments. Section 4 describes the observed lexical distributions of feed-

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 440–457
September 18–20, 2024.

back phenomena and Section 5 compares them to estimates derived with a dialogue act tagger. In Section 6, we describe the manual annotation of dialogue samples and the fine-tuning of an LLM to automate this process. Finally, Section 7 describes the results of applying this LLM-based method to synthetic dialogues, and Section 8 concludes.

2 Related Work

2.1 Conversational Feedback and Grounding

A key aspect of any communicative activity is the management of the common ground, a process often called *conversational grounding* (Clark and Schaefer, 1989). The study of grounding and related phenomena, such as conversational feedback (Allwood et al., 1992), has been instrumental to cognitive approaches to communication (Clark, 1996), and to dialogue system development (Traum, 1994; Paek and Horvitz, 2000; Yaghoubzadeh et al., 2015).

Feedback and grounding can happen at any of the *levels of communication* that includes simple contact, perception, understanding and higher-level evaluation of what had been said (Allwood et al., 1992; Clark, 1996). Conversational feedback may appear at different positions in a dialogue. However, a number of corpus studies found that they have a tendency to occur at specific places, mostly where they cause little interference (Kjellmer, 2009). These places of occurrence have also been referred to as *Feedback Relevant Spaces* (Heldner et al., 2013; Howes and Eshghi, 2021). Although, arguably, any utterance relates directly or indirectly to grounding (through implicit and high level pragmatic inference, Clark and Schaefer 1989), *acknowledgments* and other positive feedback signals (see Ex. (1)), along with *repair* (see Ex. (2)), have been identified as the most prominent grounding mechanisms (Jefferson, 1972; Bunt, 1994). Their frequency in human-human dialogue is known to be very high (e.g., Stolcke et al., 2000a) and universal across languages (Liesenfeld and Dingemanse, 2022; Dingemanse et al., 2015). These conversational signals, while they do not cover all grounding phenomena, can therefore be seen as a useful proxy to quantify feedback in a dialogue.

(1) **A:** and uh it really does irk me to see those guys out there uh you know making that ///much money///
B: ///yeah///[1]

Recent works have emphasized the role of feedback and grounding signals in their study of human-human conversations (Fusaroli et al., 2017; Dideriksen et al., 2022; Dingemanse and Liesenfeld, 2022) as well as human-agent interaction (Visser et al., 2014; Hough and Schlangen, 2016; Buschmeier and Kopp, 2018; Axelsson et al., 2022).

The literature tends to merge the two closely related concepts of *backchannels* and *acknowledgments*. Backchannels (Yngve, 1970), or *continuers* (Schegloff, 1982), are not positioned on the main channel, but uttered by the "listener", often as low intensity unobtrusive overlapping speech (Heldner et al., 2010) or non-verbally (Allwood et al., 2007; Truong et al., 2011). Acknowledgments, on the other hand, have a slightly broader, functional definition of minimal positive feedback (Jefferson, 1984; Allwood et al., 1992).

There is a large body of work on lexical markers, also called *cue phrases* or *discourse markers* (Jefferson, 1984; Allwood et al., 1992; Muller and Prévot, 2003), since they present interesting linguistic features and constitute convenient explicit cues for detecting feedback utterances automatically (Jurafsky et al., 1998; Gravano et al., 2012; Prévot et al., 2015). Gravano et al. (2012) developed a list of affirmative cue words made of *alright, mm-hm, okay, right, uh-huh, yeah*. Form-Function studies of similar lists have been made at least for Swedish (Allwood, 1988), U.S. English (Ward, 2006), and French (Prévot et al., 2015).

Few studies have, however, concentrated on direct negative feedback associated with rejection and corrective dialogue acts. Although Allwood et al. (1992) suggests a polarity dimension for characterizing feedback, most recent studies have focused on positive feedback. Indeed, in collaborative dialogue and everyday conversations, which are the two genres dominating available datasets, positive feedback constitutes the large majority of explicit feedback (e.g., Malisz et al., 2016). Negative feedback is instead often expressed constructively, using repair mechanisms, specifically *clarification requests* (Purver, 2004). These may rely on simple lexical cues (e.g., for English, *pardon?, huh?*), sluices (such as *what?, who?*), or on clarification ellipsis, as in the following example (Fernández et al., 2007):

[1] Notation: ///text/// produced in overlap with the speech of the other speaker. From Switchboard (Godfrey et al., 1992)

(2) **A:** and then we're going to turn east
B: mmhmm
A: not straight east slightly sort of northeast
B: slightly northeast?[2]

The occurrence of feedback signals in dialogue transcriptions can be detected using various types of sequence labeling models from classical hidden Markov models (Stolcke et al., 2000b) to modern neural architectures and large language models (Liu et al., 2017; Noble and Maraev, 2021).

2.2 Analysis of Subtitles

Subtitles are typically short written text snippets and they accompany audiovisual content on the screen. They are often subject to condensation and normalization, where non-standard verbal elements (repetitions, signs of hesitation etc.) are omitted or replaced by more standard alternatives (Gottlieb, 2012) due to constraints on the length, readability and writing conventions. As subtitles are displayed alongside audiovisual content, viewers can typically recover omitted dialogue-relevant cues from the accompanying images and sounds. *Interlingual subtitling* – where the original language of the audio is different from the subtitling language – differs somewhat from *intralingual subtitling*, which is meant for same-language audio and subtitles which also records non-verbal elements writing for the benefit of hearing impaired audiences or non-native speakers (Gottlieb, 2012).

Rühlemann (2020) compared real conversations and scripted ones and observed that continuers were absent from the latter. Prevot et al. (2019) compared data from the *Open Subtitles* corpus (Lison and Tiedemann, 2016; Lison et al., 2018) in English, French and Mandarin with both written and conversational corpora and found that OpenSubtitles occupied an intermediate position between written and conversational data in terms of lexical and syntactic features. This paper builds upon those earlier works but focuses specifically on communicative feedback, using a combination of lexical statistics, manual and automate annotations to quantify its frequency in various corpora.

3 Corpora

We rely on data from both OpenSubtitles and existing, publicly available corpora of real conversations covering eight different languages (see Table 1).

[2]From HCRC Map Task (Anderson et al., 1991).

3.1 Spontaneous Dialogues

German (de) We use the Hamburg MapTask corpus (HZSK, 2010), in which twelve dyads of (L2) speakers of German engage in dyadic task-oriented short dialogues.

English (en) For English, we use Switchboard (SWBD), consisting of dyadic topic oriented phone conversation (Godfrey et al., 1992) as well as Fisher (Cieri et al., 2004) for some experiments; AMI, with multi-party multimodal task-oriented dialogues (Carletta, 2007); HCRC MapTask (MT) comprising dyadic task-oriented short dialogues (Anderson et al., 1991); and STAC, a multi-party negotiation chat corpus (Asher et al., 2016).

French (fr) We include CID, consisting of dyadic, 1-hour long, loosely topic-oriented face-to-face conversations with 16 participants (Blache et al., 2017); French MapTask with 16 participants (Gorisch et al., 2014); and Aix-DVD, dyadic face-to-face conversations about movie preferences of 16 participants (Prévot et al., 2016).

Hungarian (hu) We employ BUSZI-2 corpus (Budapest Sociolinguistic Interview, Váradi, 2003), where 50 participants with different educational levels participated in a 30-minute directed conversation and then performed language tasks (e.g. grammaticality judgments).

Italian (it) We use the CLIPS corpus (Savy and Cutugno, 2009), consisting of both a map task and a difference spotting task between images. We exclude dialogues with a high proportion ($> 10\%$) of utterances with dialectal words.

Japanese (ja) This language is represented by the transcripts of the CallHome Japanese corpus (Den and Fry, 2000) consisting of 120 unscripted telephone conversations between native speakers, mostly family members or close friends.

Norwegian (no) We use the NoTa-Oslo corpus (Johannessen et al., 2007), containing interviews and conversations from 2004–2006 with 166 informants from the Oslo area. The dialogues consist of 10-minute semi-formal interviews and 30-min informal dialogues with other informants.

Mandarin Chinese (zh) The source of our Mandarin Chinese data was CALLHOME (Wheatley, 1996) consisting of unscripted telephone conversations between native speakers.

Language	de	en	fr	hu	it	ja	no	zh	total
# Spontaneous dialogues	24	2766	48	50	88	120	259	120	**3475**
# Utterances	4K	373K	27K	31K	24K	39K	86K	18K	**602K**
# Subtitles	98	100	100	68	95	74	87	93	**715**
# Utterances	131K	140K	126K	93K	138K	106K	98K	114K	**946K**

Table 1: Overview of dialogue data sources for both spontaneous conversations and subtitles employed in this paper.

3.2 Subtitles

The scripted dialogues are extracted from OpenSubtitles 2018 (Lison et al., 2018), a large collection of over 3.7 million subtitles (amounting to $\approx$ 22.1 billion words) extracted from the OpenSubtitles.org database and covering 60 languages. We include both (1) subtitles for the hearing impaired, where the subtitle language and the original audio language are identical and (2) subtitles for foreign audiences. The subtitles are then filtered according to several criteria. Only recent movies (year $\geq$ 1990) are included to reflect contemporary language use, as is the case for the corpora of spontaneous conversations. We also omit subtitles with less then 100 utterances and exclude genres that are less relevant for this study (Documentary, Reality-TV, Biography, Sport, Musical, Music, Adult, Animation, Short and Game-Show).

We sample up to ten movies per audience type (hearing impaired vs. foreign audience) from the five largest genres, namely drama, comedy, crime, action, and romance. Table 1 shows the number of movies and utterances per language for the selected subtitles. Note that subtitles are typically segmented by dialogue turns or sentences instead of utterances. The term "utterance" should therefore be understood broadly in this paper.

This paper focuses on the textual aspects of grounding phenomena. While speech and non-linguistic aspects of communicative feedback (such as timing, intonation, gestures or gaze) are both important and well-studied, in particular for acknowledgements and backchannels, those information are not available in subtitles corpora, which are intrinsically limited to text transcriptions.

4 Lexical Analysis

Lexical statistics of acknowledgment cues gives us a first picture of the feedback frequency. Acknowledgments tend to be produced by the addressee (not the main speaker) and are therefore often short productions uttered in overlap and potentially with a lower voice. Out of those three properties (brevity, overlap, lower volume), only the first is practically measurable in our experiments, as the subtitles are by construction text-based.

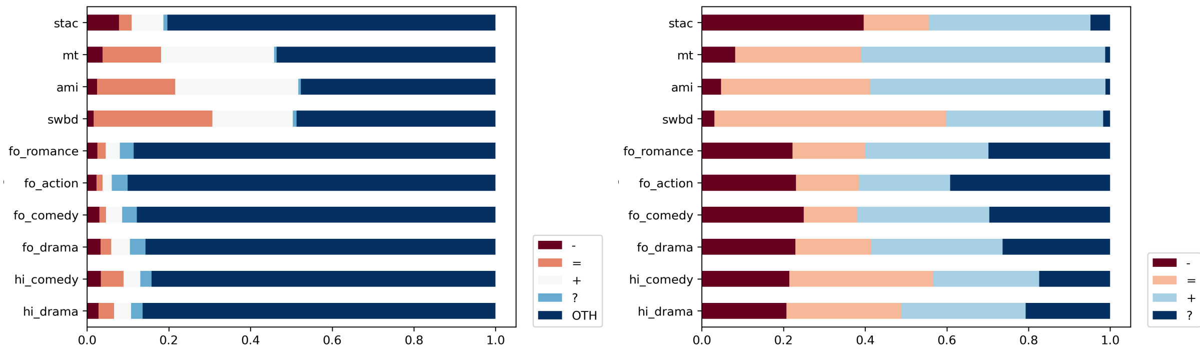
Given their relation to acknowledgments, we first analyse "very short utterances" (Edlund et al., 2009), defined here as three tokens or less. Feedback is also very well represented at initial positions of longer turns/contributions. We therefore targeted two locations: *very short utterances* (all tokens) and *initial positions* (one token) of all other utterances. Comparing term frequencies between these locations and the overall corpus allowed us to compile language-specific lists of *cue words*. Those lists of cue words (presented in Table 3 in the Appendix) are divided into four core classes of feedback:

- positive feedback/acknowledgment (**+**)
- neutral/continuer (**=**)
- negative feedback (**-**)
- clarification request (**?**).

We plot in Figure 1 the frequencies of those feedback classes in each corpus, either in terms of absolute frequency (left side) or by looking at the relative proportions of the feedback classes (right side). Figure 2 shows the lexical distribution of the most frequent lexical items observed in the utterances of plot (b) for English.

We observe that the statistics based on cue words differ substantially between subtitles and spontaneous dialogues. This difference is observed across all languages and sub-genres, (see Appendix A for other languages). We sought to identify and reduce other sources of variation between corpora. STAC, as a chat corpus, exhibits different patterns than other dialogue corpora, notably due to the presence of emojis. Similarly, for English and French, we explored the impact of politeness expression (highly frequent in OpenSubtitles). Those peculiarities did not, however, change the overall picture of our analysis (see Figure 13 in Appendix A).

One key difference between real dialogues and subtitles relates to the overall frequency of feedback cues, which is much higher in spontaneous



(a) Absolute frequency

(b) Relative frequency

Figure 1: Frequency of conversational feedback of various types among utterances in the English corpora (both spontaneous and subtitles) based on manually curated lists of cue words to detect. Fig. (a) shows the absolute frequency while Fig. (b) zooms in on utterances labelled with at least one feedback. **+** denotes positive feedback/acknowledgement, **=** neutral/continuer feedback, **-** negative feedback, **?** clarification requests and 'OTH' is for other utterances. `fo` and `hi` respectively stand for 'foreign audience' and 'hearing-impaired' subtitles. Corpora without these prefixes are spontaneous dialogues.

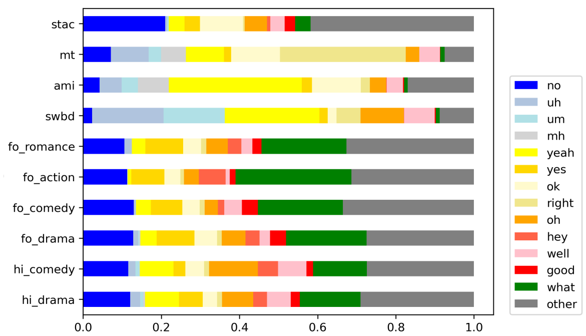


Figure 2: Most common lexical items associated with communicative feedback, as detected through manually curated lists of cue words in English, factored by corpus.

dialogues (40–50%) than in subtitles (10–20%), as observed in figure1(a). Furthermore, as shown in Figure 1(b), feedback in spontaneous dialogues consists mostly in positive or neutral (*continuers*) feedback, while subtitles have few neutral signals but seem to exhibit a much higher proportion of negative feedback and clarification requests.

We compared our English cue word lists against the annotations in Switchboard. After grouping feedback-related labels into a single *Feedback* category, we find that the cue word lists yield an F_1 score of 0.76.

5 Dialogue Act Tagging

Although lexical statistics do highlight substantial differences in subtitles and spontaneous dialogues, they remain imprecise estimates, as many cue words related to feedback tend to be ambiguous. In this section, we refine our analysis using a dialogue act tagging model trained on the DAMSL-Switchboard corpus.

5.1 Data

We map the original set of Switchboard (SWBD) tags, and their clustered DAMSL-SWBD equivalents, into five coarse dialogue act (DA) classes: *Forward looking*, *Yes/no answers*, *Assessment*, *Backchannel* and *Other*. The two classes most directly relevant for feedback, namely *Backchannel* and *Assessment*, are inspired, in part, by Mezza et al. (2018). Distinguishing between these two feedback-related classes is also motivated by Goodwin (1986), who outline a number of positional and functional differences between these. The *Backchannel* category consists of the SWBD-DAMSL labels[3] *Acknowledge (Backchannel)*, (SWBD tag `b`), *Backchannel in question form* (`bh`), *Response Acknowledgment* (`bk`), *Summarize/reformulate* (`bf`) and *Signal-non-understanding* (`br`). As this latter tag suggests, negative feedback signals are also part of the *Backchannel* category, since they are too few to reliably learn a separate class from. The *Assessment* category comprises not only the labels *Agree/Accept* (`aa`), but also *Appreciation* (`ba`) and *Exclamation* (`fe`). The forward looking category contains utterances expressing explanations, instructions and suggestions as well as questions. Table 4 in Appendix B shows the distribution of instances per label and their SWBD tag.

[3] web.stanford.edu/ jurafsky/ws97/manual.august1.html

DA group	Data	de	en	fr	hu	it	ja	no	zh
Assessment	**SpConv**	16.50	9.11	4.62	15.49	12.64	15.74	17.05	6.96
	Subs	9.08	7.07	7.72	9.29	8.34	6.48	6.53	5.00
Backchannel	**SpConv**	11.57	10.79	11.96	4.28	5.73	18.96	2.67	5.65
	Subs	3.49	3.72	3.44	3.48	3.45	3.74	3.47	3.00
Yes/no answer	**SpConv**	2.22	1.15	1.24	4.00	6.55	2.84	5.09	1.00
	Subs	1.97	1.37	1.68	1.47	1.38	1.15	2.32	0.76

Table 2: Proportions (%) of the relevant dialogue act groups detected by the BERT-based dialogue act tagger in the spontaneous conversation (SpConv) and in the subtitle (Subs) corpora.

5.2 Model Training

We fine-tune the monolingual `bert-base-cased` pretrained model (Devlin et al., 2019) using 80% of the Switchboard data as training and 20% for development and testing. We set up the task as a sequence classification problem, including the preceding utterance as context. We train the model with a batch size of 8, a learning rate of $4E-5$ and default values for the other parameters. We run and compare three different random seeds, yielding similar performance. To improve recall, we also adjust the probability thresholds for the feedback classes.

The model performs relatively well on the Switchboard test set, yielding an accuracy of 0.81. The F_1 scores for the *Assessment* and *Backchannel* classes are respectively 0.59 and 0.83. This score difference may be due to *Backchannel* instances being better represented in the training data, as well as some label confusion between the *Assessment* and the *Yes/No question* categories.

5.3 Empirical Results

We then use the trained dialogue act tagger to detect conversational feedback signals in both the spontaneous dialogue and subtitles. For non-English corpora, we machine translate the data using the Google Translate API. Feedback-annotated conversational corpora is non-existent for most languages and the quality of current MT systems is generally considered high enough to serve as a viable alternative (Isbister et al., 2021).

Table 2 presents the empirical results obtained with our dialogue act tagger on both spontaneous dialogues and subtitle corpora. We observe that backchannels are considerably more frequent (by a factor three) in spontaneous dialogues than in subtitles for half of the languages – which is in line with the results of our lexical analysis in Section 4. The number of utterances labeled as *Assessment* differs less, but subtitles still seem to contain less of this feedback type in almost all genres and languages except French (see Appendix B for details). Given that the tagger is only trained on a single corpus, some of the differences found may also be attributed to the generalization ability of the tagger to certain domains. We therefore also conduct some manual error analysis.

5.4 Error Analysis

In general, the proportion of the *Backchannel* category for the spontaneous conversations is lower for Hungarian, Italian, Norwegian and Mandarin than for the other languages. This is likely due to the use of infrequent spelling variants of backchannel signals such as *hmm, mh*. We have also found that the tagger has difficulties detecting feedback when they are part of longer utterances, whether they appear in an utterance-initial position or not. We also observe a general tendency to associate sentence-final question marks to feedback cues. When inspecting the most frequent utterances tagged as feedback, we also notice that short utterances pose some challenges for machine translation due to polysemy, e.g., Cosa? “Thing?”, also translatable as “What?”, in Italian.

6 Further Annotations

The results from the dialogue tagger do show some clear trends regarding the extent to which communicative feedback is expressed in subtitles compared to spontaneous interactions. However, the use of DAMSL-Switchboard as sole source of training data is a limiting factor in our analysis, in particular when it comes to non-English dialogues, which must be machine-translated prior to labelling. Furthermore, the tagger does not provide information about the frequency of negative feedback, although the lexical analysis from Section 4 does seem to point towards a higher frequency of those communicative signals in subtitles.

We therefore complement the analyses of the two previous sections with a manual annotation effort. To this end, we sample from each corpus a set of 300 utterances to annotate. However, as evidenced by the results of the previous sections, many utterances of our corpora do not seem to contain any communicative feedback. To ensure the annotation process can cover a sufficiently broad variety of feedback signals despite this class imbalance, we do not select the utterances purely at random, but select half among those marked as feedback-relevant by the cue words of Section 4, and the other half among those that do not.

Annotation Process

We recruited 6 annotators with prior expertise in linguistic annotation and proficient in the language corresponding to the corpus to annotate. Those annotators were provided each utterance in its context, and were tasked to decide whether the utterance in question contains one of the following three categories of communicative feedback: defined in the annotation guidelines as such:

AGREE_ACCEPT : indicates that the speaker agrees or accepts what has been said.

ACK_BACK : indicates that the speaker is listening to her interlocutor, or at least heard what has been said, without necessarily agreeing with it or committing to its content.

NEGATIVE_FEEDBACK : indicates that the speaker could not hear or understand her interlocutor, or even rejects or disagrees with what the other person has said.

Answers to explicit questions should not be considered as feedback. Each utterance can be tagged with zero, one, or multiple feedback labels. These categories specifically target and distinguish between different conversational feedback phenomena and are therefore somewhat more comprehensive than the categories employed by the tagger of the previous section. There, similar categories were derived by merging the available feedback-relevant dialogue act labels from the SWBD annotations.

A total of 24 corpus samples, each comprising 300 utterances, were annotated[4]. Three corpus samples (respectively for English, French and Chinese) were doubly annotated, and the Kappa's score of their agreement was found to be 0.59 for *AGREE_ACCEPT*, 0.42 for *ACK_BACK* and 0.54 for *NEGATIVE_FEEDBACK* across the 3 samples. This relatively low inter-annotator agreement illustrates the challenging nature of the annotation task, in particular due to the lack of explicit turn boundaries in subtitles, making it at times difficult to determine the context behind each utterance.

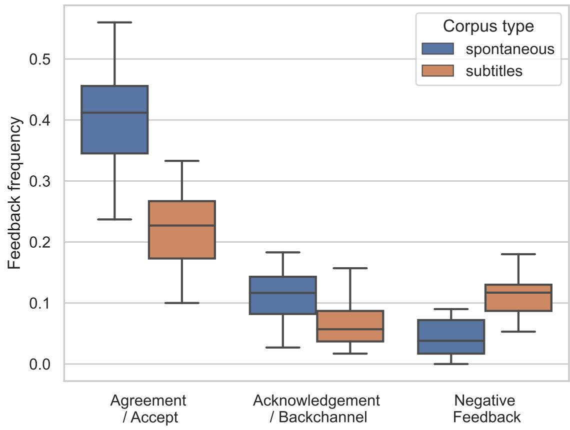


Figure 3: Frequency of communicative feedback depending on the source of the dialogue sample (spontaneous interactions or subtitles) and the category of feedback, based on annotations from human experts.

[4]The full set of annotated dialogue samples is available at https://github.com/NorskRegnesentral/conv_feedback.

6.1 Annotation Results

Figure 3 illustrates the frequencies of the three feedback categories across the 24 annotated samples. We observe again a lower proportion of both *Agree / Accept* and *Acknowledgement / Backchannel* feedbacks in the subtitles compared to real interactions. The proportion of *Negative feedback* is, however, higher for the subtitles. We hypothesise that this may stem from the fact that disagreements between interlocutors are more interesting from the storytelling perspective, and are therefore more common in subtitles than in real interactions.

We investigated whether subtitles for foreign audiences differed from subtitles written for the hearing impaired (as those often need to adhere more closely to the original on-screen conversation), but did not find any substantial disparity.

6.2 LLM-based Annotation

The frequencies of Figure 3 are obtained using the manually annotated dialogue samples. However, those samples only cover a small fraction of available corpora. Furthermore, as the sampling procedure relied on the use of cue-words to cover a

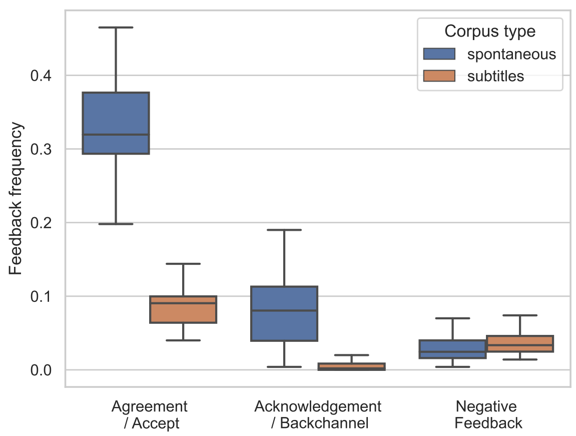


Figure 4: Frequency of communicative feedback depending on the corpus type and category of feedback, based on the predictions of the fine-tuned Gemma 2 model trained on human annotations.

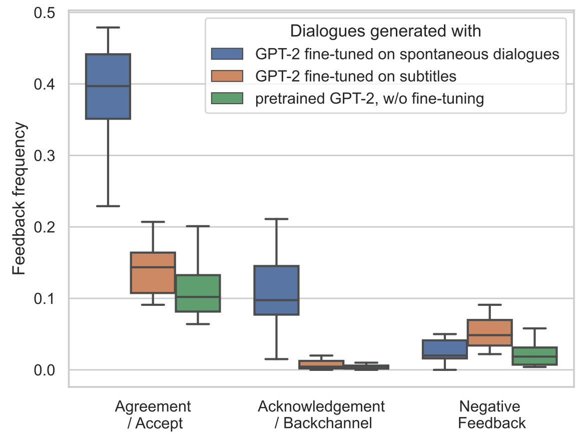


Figure 5: Frequency of communicative feedback in synthetic dialogues generated using GPT-2 models, either applied without fine-tuning or after fine-tuning on corpora of spontaneous interactions or subtitles.

sufficiently broad set of feedback types (see above), it is likely to overestimate the proportion of communicative feedbacks. To mitigate this bias, we fine-tune an instruction-tuned Gemma 2 model (Gemma Team et al., 2024) to predict the probability of an utterance including one of the three defined feedback categories. The fine-tuning relied on LoRA (Hu et al., 2021) and included as instructions the annotation guidelines also provided to the human experts. The full set of 24 dialogue samples was used for the fine-tuning, each utterance being provided in its local dialogue context. For non-English utterances, we also include in the prompt an English translation of the utterance and its context, obtained using Google Translate.

The fine-tuned Gemma2 LLM was then applied to all corpora to predict whether their utterances contained one of the three categories of feedback defined above. The results are shown in Figure 4. The proportions of communicative feedback are somewhat lower in the actual corpora than in the annotated samples (which is expected given how the dialogue samples were derived), but the overall trends remain similar to Figure 3.

7 Conversational Feedback in Synthetic Dialogues

We conclude by investigating the occurrence of communicative feedback in synthetic dialogues generated with autoregressive language models. More precisely, we wish to analyse whether the communicative feedback generated by those models are closer to the patterns found in real interactions or to scripted dialogues such as subtitles.

To this end, we use available GPT-2 models (Radford et al., 2019) for the eight covered languages [5]. The use of GPT-2 models is motivated by practical considerations and the need to obtain pre-trained models for each of the eight languages. For each corpus, we derive a fine-tuned version of its corresponding GPT-2 model by further training the model on the corpus dialogues. To account for the corpus size differences, the number of epochs is adjusted to ensure that the total number of gradient updates is similar across all corpora.

The GPT-2 models are then employed to produce synthetic dialogues (100 dialogues of about 50 turns per model. For the fine-tuned models, all turns are automatically generated, while for the base models, the following dialogue start is used as context: *Hi! – Hi, how are you? – Fine, and you?* to bias the model towards the generation of dialogues. Finally, the LLM annotator from the previous section is applied on those synthetic dialogues to estimate their frequency of communicative feedback.

The results are shown in Figure 5. We observe that the synthetic dialogues generated with the standard GPT-2 models without any further fine-tuning are much closer to the ones derived from subtitles than to those derived from spontaneous interactions when it comes to communicative feedback. This is

[5] The following pre-trained models are employed: `gpt2-base` (English), `gpt-fr-cased-small` (French), `german-gpt2` (German), `gpt2-small-italian` (Italian), `PULI-GPT-2` (Hungarian), `norwegian-gpt2` (Norwegian), `gpt2-chinese-cluecorpussmall` (Mandarin Chinese), and `japanese-gpt2-medium` (Japanese).

notably the case for positive and neutral feedback. The occurrence of negative feedback is, however, not as common as in subtitles. Although the above results were obtained here using only GPT-2 pre-trained models, we expect to find similar patterns for other (and more recent) LLMs.

8 Conclusion and Future Work

As evidenced in this paper, movie and TV subtitles exhibit notable linguistic differences to actual spontaneous dialogues in the amount and type of conversational feedback they include. Based on a collection of corpora of both spontaneous dialogues and subtitles across eight languages, we provide both lexical statistics and dialogue act estimates derived with a fine-tuned dialogue act tagger. We show that the proportion of conversational feedback is considerably lower in subtitles than in spontaneous dialogues across the corpora included. Furthermore, the type of conversational feedback also differs, as negative feedback is proportionally more frequent in subtitles. This is corroborated by manual annotations of 24 dialogue samples from the selected corpora, and the use of a fine-tuned LLM trained on those annotations. Finally, we also show that dialogues generated from language models are closer to scripted dialogue than real interactions in their use of communicative feedback. Beyond their linguistic interest, these results can provide useful insights for the development of conversational models, as those are often trained on scripted dialogues and might therefore struggle both to understand communicative feedback from the user and to produce such feedback themselves.

Acknowledgments

This work was carried out within the *GraphDial* project (`https://graphdial.nr.no/`), supported by the Research Council of Norway. The work was initiated during the bilateral PHC-Aurora project *French Norwegian Research Effort on Applied Dialogue Modelling*. Laurent Prévot would also like to acknowledge continuous support from the Institute for Language Communication and the Brain (ILCB ANR-16-CONV0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX AAP-ID-17-46-170301-11.1). Hendrik Buschmeier was supported by the German Research Foundation (DFG) in the Collaborative Research Center TRR 318/1 2021 'Constructing Explainability' (438445824). We would like to thank Hiro Yamazaki for helping with the Japanese lexical items.

References

Jens Allwood. 1988. Om det svenska systemet för språklig återkoppling. In Per Linell, Viveka Adelswärd, Torbjörn Nilsson, and Per A. Pettersson, editors, *Svenskans Beskrivning 16*, volume 1, pages 89–106. Linköping University, Tema Kommunikation, Linköping, Sweden.

Jens Allwood, Stefan Kopp, Karl Grammer, Elisabeh Ahlsén, Elisabeth Oberzaucher, and Markus Koppensteiner. 2007. The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behaviour simulation. *Language Resources and Evaluation*, 41:255–272.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Anne H. Anderson, Miles Bader, et al. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2721–2727.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modelling feedback in interaction with conversational agents – a review. *Frontiers in Computer Science*, 4:744574.

Todd Berliner. 1999. Hollywood movie dialogue and the "real realism" of john cassavetes. *Film Quaterly*, 52(3):2–16.

Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. 2017. The corpus of interactional data: A large multimodal annotated resource. *Handbook of Linguistic Annotation*, pages 1323–1356.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1213–1221, Stockholm, Sweden.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41:181–190.

Neda Chepinchikj and Celia Thompson. 2016. Analysing cinematic discourse using conversation analysis. *Discourse, Context & Media*, 14:40–53.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, OR, USA.

Mark Davies. 2021. The TV and movies corpora: Design, construction, and use. *International Journal of Corpus Linguistics*, 26:10–37.

Yasuharu Den and John Fry. 2000. CallHome Japanese corpus (in Japanese). *Journal of the Phonetic Society of Japan*, 4(2):24–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2022. Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General*.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland. Association for Computational Linguistics.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and Nick J. Enfield. 2015. Universal principles in the repair of communication problems. *PLoS ONE*, 10:e0136100.

Jens Edlund, Mattias Heldner, and Antoine Pelcé. 2009. Prosodic features of very short utterances in dialogue. In Vainio Martti, Aulanko Reijo, and Olli Aaltonen, editors, *Nordic Prosody. Proceedings of the Xth Conference, Helsinki 2008*, pages 57–68. Peter Lang, Frankfurt am Main, Germany.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach. *Computational Linguistics*, 33(3):397–427.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, pages 2055–2060. Cognitive Science Society.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-Hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2-3):205–217.

Jan Gorisch, Corine Astésano, Ellen Gurman Bard, Brigitte Bigi, and Laurent Prévot. 2014. Aix map task corpus: The French multimodal corpus of task-oriented dialogue. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2648–2652, Reykjavik, Iceland. European Language Resources Association (ELRA).

Henrik Gottlieb. 2012. Subtitles: readable dialogue? *Eye tracking in audiovisual translation*, pages 37–82.

Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39.

Mattias Heldner, Jens Edlund, and Julia Hirschberg. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of INTERSPEECH 2010*, pages 3054–3057, Makuhari, Japan.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Proceedings of Nordic Prosody XI*, pages 137–146, Tartu, Estonia.

Julian Hough and David Schlangen. 2016. Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 288–298, Los Angeles, CA, USA.

Christine Howes and Arash Eshghi. 2021. Feedback relevance spaces: Interactional constraints on processing contexts in Dynamic Syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank adaptation of large language models.

HZSK. 2010. HAMATAC. The Hamburg MapTask Corpus.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 385–390, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Gail Jefferson. 1972. Side Sequences. *Studies in social interaction*.

Gail Jefferson. 1984. Notes on a systematic deployment of the acknowledgement tokens "Yeah"; and "Mm Hm";. *Paper in Linguistics*, 17(2):197–216.

Janne Bondi Johannessen, Kristin Hagen, Joel James Priestley, and Lars Nygaard. 2007. An advanced speech corpus for Norwegian. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 29–36, Tartu, Estonia. University of Tartu, Estonia.

Dan Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Discourse Relations and Discourse Markers*.

Göran Kjellmer. 2009. Where do we backchannel?: On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14(1):81–112.

Andreas Liesenfeld and Mark Dingemanse. 2022. Bottom-up discovery of structure and variation in response tokens ('backchannels') across diverse languages. In *Proceedings of Interspeech 2022*, pages 1126–1130, Incheon, Korea.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark. Association for Computational Linguistics.

Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp, and Petra Wagner. 2016. The alico corpus: Analysing the active listener. *Language Resources and Evaluation*, 50:411–442.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Philippe Muller and Laurent Prévot. 2003. An empirical study of acknowledgment structures. In *7th workshop on the semantics and pragmatics of dialogue*.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 166–172, Groningen, The Netherlands (online). Association for Computational Linguistics.

Tim Paek and Eric Horvitz. 2000. Grounding criterion: Toward a formal theory of grounding. Technical Report MSR-TR-2000-40, Microsoft Research, Redmond, WA, USA.

Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A CUP of CoFee: A large collection of feedback utterances provided with communicative function annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3180–3185, Portorož, Slovenia. European Language Resources Association (ELRA).

Laurent Prévot, Jan Gorisch, and Sankar Mukherjee. 2015. Annotation and classification of French feedback communicative functions. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 298–306, Shanghai, China.

Laurent Prevot, Pierre Magistry, and Pierre Lison. 2019. Should we use movie subtitles to study linguistic patterns of conversational speech? a study based on French, English and Taiwan Mandarin. In *Third International Symposium on Linguitic Patters of Spontaneous Speech*.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London, London, UK.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Christoph Rühlemann. 2020. What dialog is absent from constructed dialog? *English Text Construction*, 13:132–157.

Renata Savy and Francesco Cutugno. 2009. CLIPS diatopic, diamesic and diaphasic variations in spoken Italian. In *Proceedings of the 5th Corpus Linguistics Conference: CL2009*.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of the International Conference on Spoken Language Processing*, volume 96, pages 11–14.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000a. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000b. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester, Rochester, NY, USA.

Khiet P. Truong, Ronald Poppe, Iwan de Kok, and Dirk Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proceedings of INTERSPEECH 2011*, pages 2973–2976, Florence, Italy.

Thomas Visser, David R. Traum, David DeVault, and Rieks op den Akker. 2014. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8:61–73.

Tamás Váradi. 2003. A Budapesti Szociolingvisztikai Interjú. In Ferenc Kiefer, editor, *A magyar nyelv kézikönyve*, pages 339–360. Akadémiai Kiadó, Budapest.

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182.

Barbara Wheatley. 1996. CALLHOME Mandarin Chinese Transcripts. *FTP FILE. Philadelphia, USA: Linguistic Data Consortium*.

Ramin Yaghoubzadeh, Karola Pitsch, and Stefan Kopp. 2015. Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users. In *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, pages 28–38, Delft, The Netherlands.

Victor H. Yngve. 1970. On getting a word in edgewise. In Mary Ann Campbell et al., editors, *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577. Chicago Linguistic Society, Chicago, IL, USA.

A Conversational Feedback Lexical Statistics

Cue Word Lists

In Table 3, we present the list of cue words used for computing the lexical statistics in Section 4. **Content warning**: the lists contain potentially offensive language.

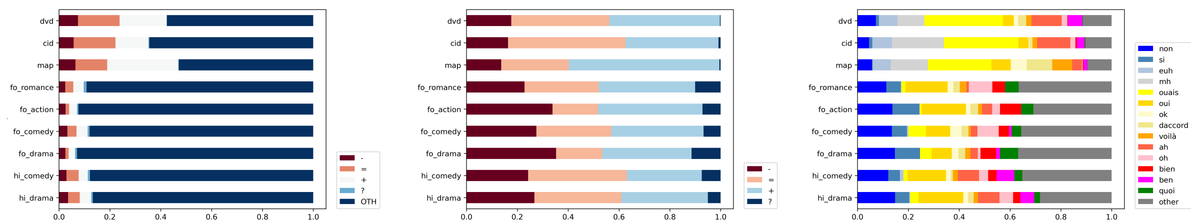
Language	FB	Lexical cues
de	+	ja, jaa, jaha, jap, jep, jo, joa, aha, hey, ach, achso, okay, ok, richtig, sicher, verstehe, cool, wow, klar, gut, definitiv, absolut, genau, natürlich, ja ja, jaja, ja okay, okay ja, ja genau, ja klar, ja gut, gut okay, ah ja, ja richtig, aber sicher, aber klar, na klar, ich weiß, weiß ich, das stimmt, du hast recht, sie haben recht, ja genau richtig, vermutlich, ja vermutlich, aber wirklich
	-	nein, nee, nö, niemals, stimmt nicht, das glaube ich nicht,glaube nicht, das glaub ich nicht,glaub nicht, vermutlich nicht
	?	wirklich, bitte, entschuldige, häh, was, wo, warum welchen, welcher, welche, welches, echt, bist du sicher,sind sie sicher
	=	mhm, m, mm, hm,ähm,mh,oh,äh
en	+	yes, yeah, yep, okay, oh, right, alright, good, ok, sure, ah, nice, cool, exactly, absolutely, true, great, oh wow, right right, oh okay, oh yeah, yeah right, um-hum yeah, that's great, yes yes, yeah yeah, uh-huh yeah, that's right, right yeah, oh yes, i see, i know, that right, that's true, that's good, all right, of course, got it, is he, oh that's nice, oh that's good, well that's nice, oh i see, oh that's great, yeah that's true, well that's good, well that's great, right that's right, oh yeah yeah, that sounds good, yeah that's right, yeah yeah yeah, yeah oh yeah, oh yeah oh, well that's true, i guess so, yeah i agree, yeah it is, i think so, oh i know, yeah i know, it really is, it is, i agree, definitely, i do too, you bet, you're right, it does, i think so too, that's it, i think you're right, i know it, i agree with you, it was, i agree with that, they are, deal, indeed, obviously, clearly, precisely, certainly, no doubt, so do I, I guess so, they really are, it did, they were, they did, me too, to me too, for me too
	-	no, wait, gosh, nope, my goodness, oh no, but um, but uh, stop it, oh my goodness, oh my gosh, wait a minute, oh my god, not really, not much, no way, shit, fuck, oh no
	?	what, really, oh really, why not, you sure, is that right
	=	um-hum, uh-huh, huh-uh, uh, hum, hm, hey, well, wow, um, huh, mh, mmhmm, m, um-hum um-hum, oh uh-huh, uh-huh uh-huh, um-hum um-hum um-hum, oh, ooh, hmm, mm, mmm
fr	+	oui, ouais, ok, ah, voilà, bien, daccord, super, parfait, exactement, ah ouais, ouais ouais, et ouais, d'accord, ah oui, oui oui, c'est ça, eh ouais, ah ouais, je sais, très bien, je comprends, bien sûr, ouais ouais ouais, ah ouais ouais, c'est vrai, ah ouais d'accord, ah d'accord, ah ouais OK, ah ouais ok, ah oui oui, ah ben oui, tu m'étonnes, c est bien, sans doute, tout à fait, absolument, vachement, je suis d'accord, moi aussi, c'est vrai, c'est juste, c'est exactement ça
	-	non, putain, pff, si, merde, oh putain, non non, mon dieu, oh mon dieu, je sais pas, non non non, pas trop, pas vraiment, pas possible
	?	hein, quoi, vraiment, comment ça
	=	ah, mh, euh, oh, han, ben, bon, hm, hum, peut-être, m, mh mh, mh ouais, ah bon, mh mh mh, eh, hé, hey
no	+	ja, jo, å ja, ok, oi, greit, presis, wow, riktig, sant, nettopp, absolutt, jepp, definitivt, åpenbart, deal, selvfølgelig, sikkert, akkurat, god, bra, helt sikkert, jeg vet, jeg skjønner, helt riktig, det stemmer, klart det, uten tvil, det er riktig, det er greit, det er sant, det er det, jeg er enig, du har rett, det gjør det, jeg tror det, jeg vet det, det var det, det gjør jeg, jeg antar det, det gjorde det, det gjør jeg også, det tror jeg også, jeg tror du har rett, jeg er enig med deg, jeg er enig i det,de er det, de var, det gjorde de, meg også, til meg også, for meg også
	-	nei, faen, javel, herregud, ikke helt, ikke mulig, ikke i det hele tatt
	?	virkelig, hva, hæ
	=	m, mhm, mh, hmm, mm, mmm, mmhmm, hm, uh-huh, ikke sant
hu	+	igen, tényleg, úgy van, helyes, jogos, igaz, valóban, pontosan, tudom, rendben, ok, oké, oksi, okés, okszi, igen az, de az, bizony, természetesen, határozottan, feltétlenül, mindenképp, egyetértek, szerintem is, ó igen, hogyne, tényleg az, én is, nekem is, engem is, tőlem is, bennem is, igazad van, naná, mi az hogy, meghiszem azt, biztosra veheted, biztos lehetsz benne, jó, ja, szerintem igen, szerintem is, én is így gondolom, én is úgy gondolom, ennyi, ez az, így van, úgy van, szerintem igazad van, szerintem igazatok van, tudom, jól tudom, egyetértek, az volt, ez volt, de, azok, igen, azok, megegyeztünk, egyértelműen, azt hiszem, kétségtelenül, biztosan, persze, értem, tudod, stimmel, valóban, hát igen, hát dehogynem
	-	nem, nem igazán, nem létezik, a francba, a fenét, ne, a csodát, hogy a csodába, hát nem
	?	ó tényleg, micsoda, tényleg, miért ne, biztos
	=	aha, hú, ú, ó, óh, hű, ja, mhm, mm, mmm, hmm, hmmm, wow, azta, ejha, nahát, ühüm

it	+	ehi, okay, okay, ok, sì, si, vabbè, ecco, perfetto, wow, esatto, certamente, esattamente, assolutamente, sicuramente, decisamente, ovviamente, precisamente, di sicuro, sono d'accordo, concordo, eccellente, grandioso, ottimo, certo, infatti, fantastico, magnifico, naturalmente, giusto, bene, già,lo ben so, ah ah, ah ha, vero, é vero, lo so, lo è, davvero, vero, oh sì, lo è veramente, anch'io, anche io, hai ragione, d'accordo, va bene, benissimo, bello, buono,penso di sì, credo di sì, mi sa di sì, mi pare di sì,anche secondo me, lo penso anch'io, è così, penso che tu abbia ragione, penso tu abbia ragione, credo che tu abbia ragione, credo tu abbia ragione, mi sa che hai ragione, sono d'accordo con te, sono d'accordo con voi, lo era, lo è stato, lo è stata, sono d'accordo con ciò, lo sono, senza dubbio, a posto, ci sto, lo sono stati, lo erano, anche a me
	-	oddio, merda, no, non proprio, non molto, non è possibile, cazzo, oh no, macché
	?	come, davvero, cosa
	=	eh, Mm-hmm, hmm, mmm, mh, eh, mhmh, eh, m, hm, ah, oh, beh, uh-huh, mmh, eeh
ja	+	そう, はい, ええ, そうか,はあ, どうぞ, 本当, は, あっ, ああ, あ, ね
	=	うん, ふーん,えっ, へえ, うーん, ふん, え,う
	-	ううん, いいえ,いや, いえ, ない, 全くない,ちょっと
	?	何
zh	+	okay, yeah, yes, ok, 对, 哦, 好, 是, 有, 真的, 还行, 然, , 太好了, 耶, 行, 一定, 没错, 那好, 对了, 真好, 好啊, 好吧, 可以, 太棒了, 太棒了, 好极了, 说得对, 没问题, 我同意, 懂了, 一樣, 我也是, 不错, 是啊, 就是这样, 当然可以
	-	不, 沒有, 不起, 不是
	?	啊, 是吗, 什麼, 什么, 为什么
	=	hey, oh, 嘿, 嗯, 呃, 哼, 哈, 嘘, 喔, 呵呵, 噢, 哇, 哦, 哟, 咦

Table 3: Lists of cue phrases employed in the lexical overview of Section 4. We distinguish between four core categories of feedback, namely positive feedback/acknowledgment (**+**), neutral/continuer (**=**), negative (**-**), and clarification request (**?**).

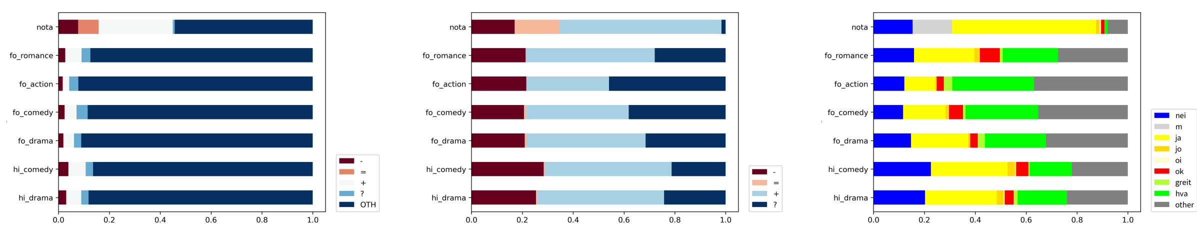
Lexical Statistics plots

Figures 6 – 12 present statistics for utterance and feedback types as well as common feedback-related lexical items for different languages. Figure 13 shows politeness keywords and emojis in our English and French corpora.



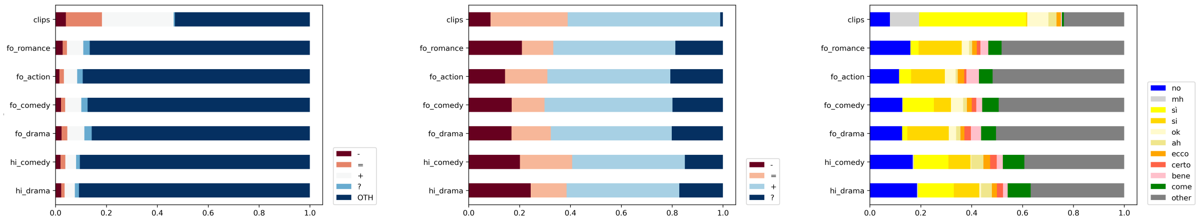
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 6: French across genres (rule-based, based on cue word lists).



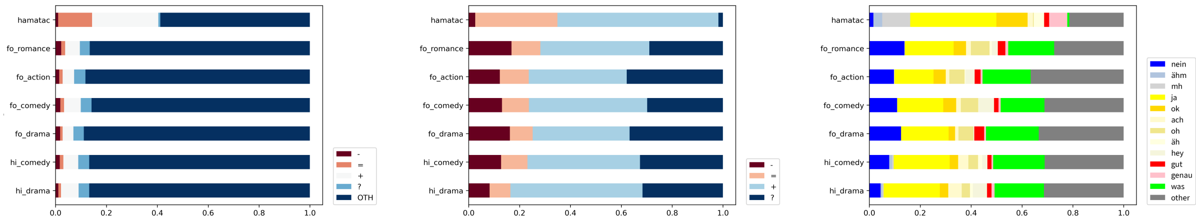
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 7: Norwegian across genres (rule-based, based on cue word lists).



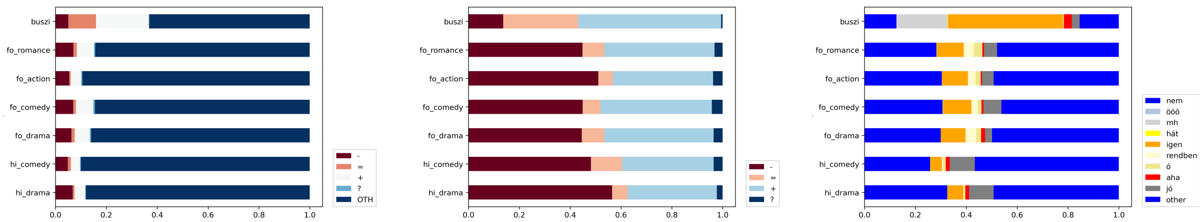
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 8: Italian across genres (rule-based, based on cue word lists).



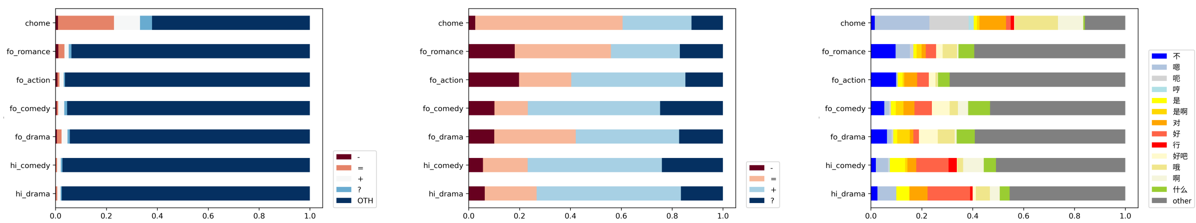
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 9: German across genres (rule-based, based on cue word lists).



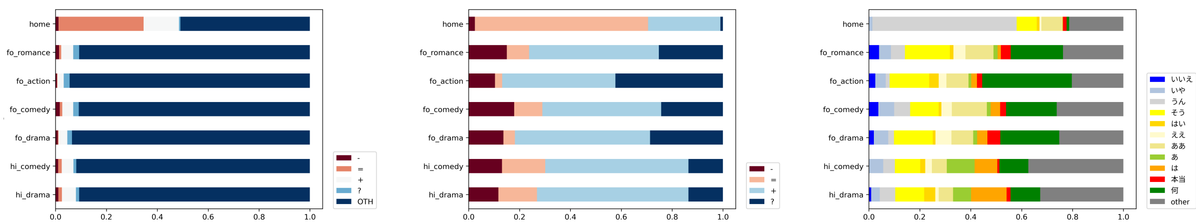
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 10: Hungarian across genres (rule-based, based on cue word lists).



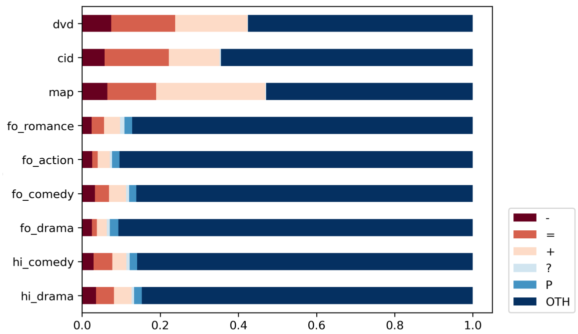
(a) Utterances type (b) Feedback types (c) Lexical items

Figure 11: Mandarin Chinese across genres (rule-based, based on cue word lists).

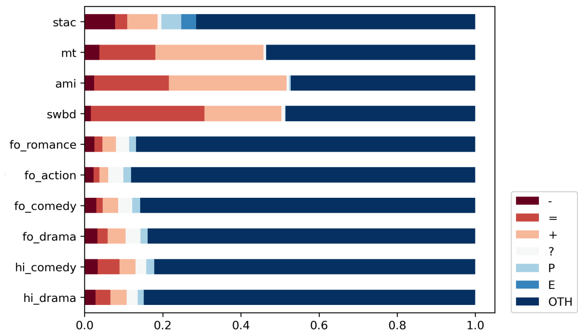


(a) Utterances type (b) Feedback types (c) Lexical items

Figure 12: Japanese across genres (rule-based, based on cue word lists).



(a) French including politeness (P) keywords



(b) English with politeness (P) and Emojis (E) keywords

Figure 13: Short utterance distribution including politeness and emojis.

B Detailed Dialogue Act Tagging Results

Dialogue Act Grouping

Table 4 shows the distribution of instances per mapped dialogue act group in the DAMSL-Switchboard (SWBD) corpus.

DA group	# inst.	SWBD labels
Forward looking	109,382	sd, fx/sv, bf, na, ny^e, arp, nd, no, cc, co, oo, ad, qr/qy, qw, qw^d, qh, qo
Backchannel	41,017	b, bk, bh, bf, br
Assessment	15,727	aa, fe/ba
Yes/no answer	4,324	ny, nn
Other	40,124	*all other categories*
Total	**210,574**	

Table 4: Instances created from the DAMSL-SWBD corpus with labels mapped to coarse-grained dialogue act groups.

Results per Corpus

Tables 5 and 6 present the results of our dialogue act tagger per (sub)corpus used. Here, we only make a binary distinction by grouping the feedback-relevant classes *Backchannel* and *Assessment* into a single *Feedback* category. The number of utterances refers to the final version of the data after pre-processing with meta-linguistic information removed.

Lang	Corpus	# utt	# feedback	% feedback
de	action_foreign	12,760	1,703	13.35
	action	12,134	1,637	13.49
	comedy_foreign	12,627	1,849	14.64
	comedy	16,152	2,369	14.67
	crime_foreign	12,589	1,245	9.89
	crime	11,817	1,581	13.38
	drama_foreign	14,669	1,350	9.2
	drama	11,460	1,452	12.67
	romance_foreign	13,499	1,500	11.11
	romance	11,809	1,596	13.52
en	action	11,094	1,437	12.95
	action_foreign	12,908	1,448	11.22
	comedy	13,948	1,665	11.94
	comedy_foreign	13,533	1,677	12.39
	crime	14,990	1,700	11.34

	crime_foreign	13,911	1,267	9.11
	drama	14,944	1,729	11.57
	drama_foreign	10,243	1,041	10.16
	romance	16,132	2,166	13.43
	romance_foreign	15,521	1,698	10.94
fr	action_foreign	11,236	1,119	9.96
	action	12,406	1,453	11.71
	comedy_foreign	17,239	1,788	10.37
	comedy	13,932	1,913	13.73
	crime_foreign	12,159	1,017	8.36
	crime	10,821	1,003	9.27
	drama_foreign	10,002	804	8.04
	drama	11,094	1,313	11.84
	romance_foreign	12,043	1,360	11.29
	romance	13,959	1,604	11.49
hu	action_foreign	12,781	1,377	10.77
	comedy_foreign	15,031	1,998	13.29
	comedy	14,692	2,462	16.76
	crime_foreign	13,620	1,655	12.15
	drama_foreign	13,138	1,400	10.66
	drama	7,872	1,103	14.01
	romance_foreign	13,771	1,611	11.7
it	action_foreign	12,010	1,585	13.2
	action	7,703	826	10.72
	comedy_foreign	15,055	2,058	13.67
	comedy	15,363	1,777	11.57
	crime_foreign	12,454	1,320	10.6
	crime	13,885	1,479	10.65
	drama_foreign	17,444	2,289	13.12
	drama	12,838	1,467	11.43
	romance_foreign	14,702	1,696	11.54
	romance	14,573	1,549	10.63
ja	action_foreign	11,245	967	8.6
	action	3,007	443	14.73
	comedy_foreign	16,173	1,777	10.99
	comedy	15,675	2,555	16.3
	crime_foreign	16,296	1,311	8.04
	drama_foreign	14,201	997	7.02
	drama	11,410	1,204	10.55
	romance_foreign	14,042	1,210	8.62
	romance	1,780	145	8.15
no	action_foreign	10,480	892	8.51
	action	1,855	290	15.63
	comedy_foreign	14,406	1,834	12.73
	comedy	11,957	1,199	10.03
	crime_foreign	12,788	1,137	8.89
	crime	9,863	853	8.65
	drama_foreign	12,031	1,202	9.99
	drama	6,688	589	8.81
	romance_foreign	12,830	1,313	10.23
	romance	4,197	399	9.51
zh	action_foreign	11,570	967	8.36
	action	2,722	159	5.84
	comedy_foreign	14,692	1,564	10.65
	comedy	13,587	1,034	7.61
	crime_foreign	10,778	795	7.38
	crime	11,182	697	6.23
	drama_foreign	14,527	1,330	9.16
	drama	9,567	743	7.77
	romance_foreign	13,362	1,079	8.08
	romance	11,440	700	6.12

Table 6: Number and frequency of communicative feedback phenomena predicted by the BERT-based dialogue act tagger on our subtitle corpora. Non-English datasets were automatically translated into English before inference.

Lang	Corpus	# utt	# feedback	% feedback
de	Hamburg MapTask	4,012	1,126	28.07
en	AMI	83,085	20,044	24.12
	Fisher	2,117,748	421,069	19.88
	HCRC MapTask	26,949	8,366	31.04
	STAC	5,841	514	8.8
fr	CID	12,326	1,754	14.23
	Aix-DVD	7,578	1,323	17.46
	French MapTask	6,046	1,226	20.28
hu	BUSZI-2	30,979	6,125	19.77
it	CLIPS	24,289	4,461	18.37
ja	Japanese CallHome	38,701	13,432	34.71
no	NoTa-Oslo	85,506	16,861	19.72
zh	Chinese CallHome	17,853	2,251	12.61

Table 5: Number and frequency of communicative feedback phenomena predicted by the BERT-based dialogue act tagger on spontaneous dialogue corpora. Non-English datasets were automatically translated into English with the Google Translate API before inference.