Annotation of Communicative Functions of Short Feedback Tokens in Switchboard

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, Gabriel Skantze

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1849–1859
Marseille, 20-25 June 2022

Annotation of Communicative Functions of Short Feedback Tokens in Switchboard

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, Gabriel Skantze
Furhat Robotics, ReadSpeaker, Aix-Marseille University, KTH Royal Institute of Technology
carol@furhatrobotics.com, adaeze.adigwe@readspeaker.com, magalie.ochs@lis-lab.fr, skantze@kth.se

Abstract

There has been a lot of work on predicting the timing of feedback in conversational systems. However, there has been less focus on predicting the prosody and lexical form of feedback given their communicative function. Therefore, in this paper we present our preliminary annotations of the communicative functions of 1627 short feedback tokens from the Switchboard corpus and an analysis of their lexical realizations and prosodic characteristics. Since there is no standard scheme for annotating the communicative function of feedback we propose our own annotation scheme. Although our work is ongoing, our preliminary analysis revealed lexical tokens such as *yeah* are ambiguous and therefore lexical forms alone are not indicative of the function. Both the lexical form and prosodic characteristics need to be taken into account in order to predict the communicative function. We also found that feedback functions have distinguishable prosodic characteristics in terms of duration, mean pitch, pitch slope, and pitch range.

Keywords: feedback, prosody, function

1. Introduction

According to Clark (1996), conversations are a joint activity. When one person is speaking the listener may produce vocalizations such as *mhmm*, *yeah*, *wow* which help establish common ground. Clark (1996) defines common ground as "the sum of mutual, common, joint knowledge, beliefs, and suppositions". These vocalizations serve as linguistic feedback to "give information about the basic communicative functions, i.e., continued contact, perception, understanding, and emotional/attitudinal reaction, in a sufficiently unobtrusive way to allow communication to serve as an instrument for pursuing various human activities" (Allwood et al., 1992).

Short vocalizations such as *mhmm*, *yeah*, *wow* have been described in different terms. Yngve (1970) referred to these short messages, either vocal or gestural signals, as *backchannels*. The term was used to distinguish the main channel (where the current speaker is sending messages) from the backchannel (where the interlocutor is sending unobtrusive messages).

Schegloff (1982) described these vocalizations as *continuers*, which function as a way of telling the interlocutor that they would like them to continue speaking and are passing the opportunity to interrupt the ongoing utterance. Following the work of Schegloff (1982), Goodwin (1986) further categorizes these vocalizations into *continuers* and *assessments*. Goodwin (1986) distinguished continuers and assessments by their position, whether they occur between or within talking units of the interlocutor, as well as their function. Unlike continuers, assessments are a response to something specific that the interlocutor just said.

The terms *generic* and *specific* listener responses, introduced by Bavelas et al. (2000), describe how these listening responses contribute to the narration of a narrator. These listener responses have the same function as continuers and assessments. The generic listener responses correspond to continuers; "they are not specifically connected to what the narrator is saying" (Bavelas et al., 2000). Specific listener responses correspond to assessments; they are "tightly connected to what the narrator is saying at the moment" and therefore the listener becomes a co-narrator (Bavelas et al., 2000).

When it comes to incorporating feedback in conversational systems (Axelsson et al., 2022), there has been a lot of work on generating feedback, with the main focus on predicting the timing of feedback (Ward and Tsukahara, 2000; Ruede et al., 2019; Adiba et al., 2021; Skantze, 2021). While there has also been work on predicting the functions of feedback (Gravano et al., 2007; Prévot et al., 2015; Ortega et al., 2020; Boudin et al., 2021; Jang et al., 2021), there has been less recent work on generating feedback (i.e., predicting their lexical and prosodic realization), given their communicative function (Wallers et al., 2006; Edlund et al., 2005). To develop such models, conversational data annotated with communicative functions of feedback is needed.

The aim of this paper is to provide annotations of the communicative functions of short feedback tokens in the Switchboard corpus (Godfrey et al., 1992). The reason we have chosen to annotate the Switchboard corpus is due to the fact that it is a relatively large manually transcribed dataset with separate audio channels and word alignments. While the corpus has been annotated with dialog acts related to backchannels (Jurafsky et al., 1998), their coding scheme is fairly limited, which we will discuss further in Section 4.

The distinction between feedback and other short vocalizations is not clear-cut. Edlund et al. (2009) introduced the auxiliary unit *very short utterance* (VSU) to encompass a broad range of phenomena such as

backchannels, feedback, and continuers in dialogue. In a similar way, we use the term *short feedback tokens* to refer to the broad set of brief vocalizations that serve some sort of feedback function. This term might include a wider range of feedback functions than what is sometimes referred to as backchannels. For example, we include brief acknowledgements (e.g. "okay"), even if they could be regarded as a separate turn (and thus not necessarily as backchannels). Following the definition of feedback by Allwood et al. (1992), we also include brief yes/no responses. Although yes/no responses are often not considered as feedback in other schemes, we think that from the perspective of a conversational system, they are similar in form and must be properly classified. We thus aim to explore the wider range of communicative functions that these short feedback tokens have.

The structure of this paper is as follows. In Section 2 we review existing schemes for annotating feedback communicative functions. In Section 3 we describe our scheme for annotating feedback communicative functions. Section 4 describes the Switchboard corpus and the process used to select our short feedback tokens. In Section 5 we report the inter-annotator agreement. In Section 6 we discuss the prosodic characteristics and lexical forms[1] of these short feedback tokens, followed by our conclusion.

2. Related Work

In this section, we present an overview of the coding schemes that have been proposed for the annotation of feedback functions for corpora in various different languages (English, Swedish, French, German). Some have been developed specifically for vocal feedback, whereas others consider multimodal feedback (vocal and visual).

Several coding schemes (Allwood et al., 2007; Buschmeier et al., 2011; Malisz et al., 2016) are based on the four feedback functions introduced by Allwood et al. (1992): *Contact*, *Perception*, *Understanding*, and *Attitudinal reactions*. The definitions are as follows:

- *Contact*: whether the interlocutor is willing and able to continue the interaction
- *Perception*: whether the interlocutor is willing and able to perceive the message
- *Understanding*: whether the interlocutor is willing and able to understand the message
- *Attitudinal reactions*: whether the interlocutor is willing and able to react and (adequately) respond to the message, specifically whether he/she accepts or rejects it.

[1]There is no consensus whether vocalizations such as *mhm* and *uh-huh* are lexical or non-lexical.

These four feedback functions are respectively related to the four levels of joint actions of an addressee proposed by Clark (1994). Clark (1996) argues that conversations are joint activities made up of joint actions by the speaker and addressee. These joint activities occur in four levels (Clark, 1994):

- *Level 1 - Vocalization and Attention*: The speaker vocalizes an utterance and the addressee attends to the utterance.
- *Level 2 - Presentation and Identification*: The speaker presents an utterance and the addressee identifies the utterance.
- *Level 3 - Meaning and Understanding*: The addressee understands the meaning of the speaker's utterance.
- *Level 4 - Proposal and Uptake*: The speaker proposes a project and the addressee takes up the project.

Although they use different terms, notice how *Contact* and *Attention*, *Perception* and *Identification* have similar definitions. Note how *Uptake* doesn't fully correspond to *Attitudinal reactions* because Clark (1994) doesn't specify attitudes. However, *Uptake* partially corresponds to *Attitudinal reaction* because *Attitudinal reaction* is defined as the acceptance of the interlocutor which can be thought of as uptake.

The four functions proposed by Allwood et al. (1992) have been used to create the MUMIN coding scheme (Allwood et al., 2007) which contains three functions:

- *Basic* - has two features: (i) Continuation/contact and perception (CP) and (ii) Continuation/contact, perception and understanding (CPU).
- *Acceptance* - has two features: (i) either accept or non-accept, meaning the interlocutor has both perceived and understood the message and (ii) either agrees or disagrees with the message. Note that 'Acceptance' was part of the definition of *Attitudinal reactions* in (Allwood et al., 1992).
- *Additional emotion/Attitude* - This category lists six emotions: Angry, Disgusted, Frightened, Happy, Sad, Surprised. However, 'other' is also added so as to not limit the list.

The MUMIN scheme has inspired the function annotation scheme of Neiberg et al. (2013). What is unique about this annotation scheme is that the functions are represented on a scale: *Non-understanding – Understanding, Disagreement – Agreement, Uninterest – Interest, Expectation – Surprise, and Uncertainty – Certainty*. When the feedback tokens were judged in the center of the scale, they were assigned neutral function, which corresponds to *contact* and *perception*. Although this scheme doesn't have an *Acceptance* function, it does have an *understanding* function and *agreement* function which were part of MUMIN's definition

of *Acceptance*. It also has three attitude functions (*Interest, Surprise, and Certainty*).
Allwood et al. (1992) also inspired the annotation scheme by Buschmeier et al. (2011) and Malisz et al. (2016). They define three positive categories:

- P1: The partner indicates her/his perception of the signal: "I hear you and please continue."
- P2: The partner indicates her/his perception and understanding of the message content: "I understand what you mean."
- P3: The partner indicates her/his perception, understanding and acceptance of the message or agreement with the message: "I accept/agree/believe what you say."

The three negative categories (N1, N2, N3) are the opposites to the definitions of P1, P2, and P3. Note that this scheme excludes *Contact*. There is also a function for attitudinal information, category A which can also be a modifier. Annotation A is used for when "the partner expresses an attitude towards the message, e.g. surprise, excitement, admiration, anger, disgust" (Buschmeier et al., 2011; Malisz et al., 2016). The feedback token would be annotated solely as A when used as a category and P3A when used as a modifier. Although this scheme does not have a category for *Contact*, P1 and P2 corresponds to MUMIN's *Basic* function and P3 corresponds to MUMIN's *Acceptance* function.
Although not explicitly inspired by Allwood et al. (1992) feedback functions or Clark (1994) levels of grounding, the DIT++ taxamony of communicative functions for dialogue by Bunt (2009) contain two feedback functions *allo-feedback* and *auto-feedback* which provide information about attention, perception, interpretation, evaluation, and execution.
The discourse functions defined by Gravano et al. (2007) and Benus et al. (2007) are used to annotate single affirmative words such as *alright*, *mmhmm*, *okay*, *right*, *uhhuh*, *yeah* etc. Yet, these affirmative words can be regarded as feedback. The two discourse functions that can be feedback functions are:

- A1: *Acknowledgement/agreement*. Indicates "I believe what you said," and/or "I agree with what you say."
- A2: *Backchannel*. Only indicates "I hear you and please continue" in response to another speaker's utterance.

There are other function schemes that annotate feedback into two levels and subcategories for each level as proposed in (Prévot et al., 2015):

- *Base Function*
  - *Contact*: I am still here listening.
  - *Acknowledgment*: I have heard / recorded what you said but nothing more.
  - *Evaluation-base*: I express something more than mere acknowledgement (approval, expression of an attitude,...)
  - *Answer*: I answer to your question / request.
  - *Elicit*: Please provide some feedback.
  - *Other*: This item is not related to feedback.
- *Evaluation*
  - *Approval*: I approve vs. disapprove / agree vs. disagree with what you said.
  - *Expectation*: I expected vs. did not expect what you said.
  - *Amusement*: I am amused vs. annoyed by what you said.
  - *Confirmation/doubt*: I confirm what you said vs. I still doubt about what you said.

Note that the definition of acknowledgment in this scheme differs from the one defined in Gravano et al. (2007). Prévot et al. (2016) used the same functions in this scheme, except they remove the *Contact* category and rename the *Evaluation* category to *Attitude*.
In the scheme by Boudin et al. (2021), feedback tokens are also categorized by their two functions, *generic* and *specific*. Within the *specific* category, feedback tokens are further classified as *positive-expected, positive-unexpected, negative-expected, negative-unexpected*. In total there are 5 categories of feedback functions.
To summarize, feedback function annotation schemes can be thought of as two groups, those inspired by Allwood et al. (1992), and those that categorize feedback into two levels *Generic/Base* and *Specific/Evaluation/Attitude*.

3. Proposed Annotation Scheme

Given that there is no standard annotation scheme for annotating feedback communicative functions and existing schemes use the same label name but different descriptions or use different label names for the same description, we propose a new scheme. Our annotation scheme is an attempt at bringing the ideas of the existing schemes discussed in Section 2 into a single scheme. Table 1 lists the description of the feedback functions with corresponding labels that we have identified in the Switchboard corpus[2].
In our scheme, we refer to the levels of joint actions of the addressee proposed by Clark (1994) and the four feedback functions proposed by Allwood et al. (1992) as *grounding levels*: *Contact, Perception, Understanding, Acceptance/Attitude* where the level below is implied by the level above. Since we regard *Contact, Perception, Understanding, Acceptance/Attitude*

[2]For examples of the annotation guideline see https://carolfigphd.github.io/FeedbackAnnotationScheme/.

Label	Function Description
(C) Continue	Continue speaking. I hear you and I'm listening but not necessarily agreeing/disagreeing.
(U) Non-Understanding	I'm uncertain I understood/heard what you said.
(A) Agree	I agree with what you said.
(D) Disagree	I doubt what you said is true. I disagree with what you said.
(Y) Yes	I am giving a positive response/answer to your yes/no question.
(N) No	I am giving a negative response/answer to your yes/no question.
(S) Sympathy	I'm expressing sympathy/pity/sorrow/concern/compassion to a negative statement.
(Ds) Disapproval	I am showing disapproval/disgust.
(MS) Mild Surprise	I am showing mild surprise, showing slight interest.
(SS) Strong Surprise	I am showing strong surprise; I am impressed.
(O) Other	Not a feedback. Filler, listener trying to take turn, or not enough context to determine function.
(?) Unknown feedback	A feedback that doesn't fit into the current labels.

Table 1: Labels and function description.

Grounding Level	Positive	Negative
Acceptance/Attitude	A, Y, S, SS, MS	D, N, Ds
Understanding	C	U
Perception		
Contact		

Table 2: Labels in Grounding Framework.

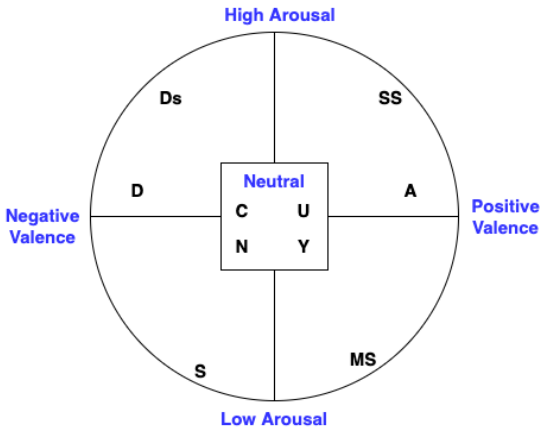

Figure 1: Labels in Valence-Arousal Space.

as grounding levels and not as functions, we therefore do not use them as function labels.

In their confusion matrix on inter-agreement Malisz et al. (2016) reported a lot of disagreement; they reported that annotators tended to favor the label representing the middle grounding level, P2 (perception and understanding). They explain that the reason why the annotators gravitated to the middle grounding level is because "feedback functions form a hierarchy of grounding strength" (Malisz et al., 2016). This is the reason why the grounding levels *Contact, Perception, Understanding, Acceptance/Attitude* in our scheme are not function labels themselves; we rather think of the feedback functions as being in these grounding levels. Table 2 shows the function labels in the corresponding grounding levels.

As can be seen, we do not make a more fine-grained distinction between feedback on the levels of *Contact, Perception, and Understanding*, as we believe these distinctions are hard to discern based on the form of short vocalizations, and that they would have to rely on context to a large extent. Schegloff (1982) described tokens such as *uh-huh*, *mhmm*, and *yeah* as claiming attention (contact) and/or understanding. Neiberg et al. (2013) represented contact and perception with a neutral function. Allwood et al. (2007) *Basic* function had two features CP (contact and perception) and CPU (contact, perception, and understanding). Our function *(C) Continue* can be interpreted as both having contact and perception, or contact, perception and understanding, depending on the context of the feedback.

Allwood et al. (2007) separate *Acceptance* and *Attitude* into two categories. In Buschmeier et al. (2011; Malisz et al. (2016) *Attitude* can either be a modifier to a category or a category by itself. For example, in their scheme, a feedback can be labeled as A which expresses an attitude or it can be used as a modifier P3A, where A is a modifier to P3 (perception, understanding, and acceptance).

In our scheme, the top grounding level is *Acceptance/Attitude*, where we collapsed *Acceptance* and *Attitude* into one single grounding level. Similarly, Prévot et al. (2015) and Prévot et al. (2016) list *Approval/non-approval*, which can be thought of as *Acceptance*, in their *Evaluation/Attitude* level. We did not want to separate *Acceptance* and *Attitude* into two levels because in our scheme the level below is implied in the level above. Therefore, if a feedback is annotated with function *(MS) Mild Surprise*, it does not necessarily mean

that the message was accepted (agree) or rejected (disagree), but we can assume that the levels of understanding, perception and contact are fulfilled.
Our scheme is also flexible in the sense that functions such as *(C) Continue* can be thought of as corresponding to the *generic* listener responses category by (Bavelas et al., 2000), while functions such as *(U) Non-understanding* and functions listed in the *Acceptance/Attitude* level can correspond to the *specific* listener responses.
Another way to visualize the labels is in a valence-arousal space similar to a circumplex model of affect (Russell, 1980), see Figure 1. In this space, feedback functions with attitude can be mapped to one of the four quadrants, while those that do not have attitude/emotion information can be mapped to the neutral section. The current labels in the *Acceptance/Attitude* grounding level are not exhaustive, they currently represent those annotated so far in the Switchboard corpus.

4. Corpus description and Feedback Selection

The Switchboard corpus consists of about 2500 telephone conversations lasting between 3-10 minutes long (Godfrey et al., 1992). It is a multi-speaker corpus consisting of 500 speakers from the U.S. The participants were given topic prompts, however, the speech is spontaneous. Each speaker was recorded in a separate channel and the recordings were transcribed and time-aligned at the word level for each speaker.
In previous work, the Switchboard corpus has been annotated with dialog acts which are backchannels or are related to backchannels (Jurafsky et al., 1998). They annotated four sub-types of backchannels with the following dialog act tags: continuer (the other speaker should continue speaking), incipient speakership (speaker trying to take the floor), yes-answer (response to a yes-no question), and agree/accept (agreement with opinion or acceptance of proposal). These four dialog acts correspond to functions in our scheme, however, attitudinal information is missing. For example, *wow* would not be annotated in their four sub-types of backchannels. In our annotation, we include such attitudinal information (e.g. surprised, disgust, pity).
Although Jurafsky et al. (1998) annotated the Switchboard corpus, we chose to identify short feedback tokens differently. First, we used the word alignments to identify possible feedback tokens that were preceded and followed by silences of at least 5 seconds. From these, we selected only lexical tokens that had a count of at least 10. From the full conversations, clips of length 20 seconds or more with at least 2 feedback tokens were selected. In total we had 7545 clips of length 20 seconds or more. From these 7545 clips we identified 26981 possible instances of feedback. So far we have annotated 1627 instances of feedback by one annotator; see Table 4 for the count per function.

5. Preliminary Inter-annotator agreement

In total 1124 feedback tokens were annotated by two raters given the scheme proposed in Table 1. Prior to the annotation exercise the two raters met to build a shared understanding of the labeling scheme and discussed example instances for each function. Afterwards the annotation was carried out individually by each of the annotators. The annotators were required to listen to the entire conversation clip, with multiple listens if necessary and infer the category of the feedback based on the context and the prosody of the token. However, in situations where the annotator was split between two labels they could assign two categories to the feedback token. For example, a feedback token could be given the label *'A/C'* indicating that the annotator could either view that feedback token as an *(A) Agree* or a *(C) Continue*. The distribution of the labels across each annotator is illustrated Figure 2.

Currently, the annotations are skewed, some function categories have less instances than others. For example, in our current annotations we have few instances of *(D) Disagree, (Y) Yes, (N) No, (S) Sympathy* compared to *(C) Continue* and *(A) Agree*. Only one annotator used the ? label, but we include this label in our annotation scheme so that new feedback functions that we have not identified can be discovered in future work. Although we do not have proportional distribution of all function labels and are aware that this can affect our inter-agreement calculations, we present our inter-annotator agreement calculations.
Similar to Kousidis et al. (2013), our exercise is mainly to validate our proposed feedback annotation scheme for which Cohen Kappa is a stricter inter-annotator reliability measurement for our evaluations. Therefore, in order to calculate the inter-agreement score, we employ two approaches - *percentage agreement* and *Cohen Kappa Coefficient*. The first approach calculates the percentage agreement over the samples labeled by both annotators. Within this approach, we report three different types of percentage-agreement metrics:

- *Absolute.* If only exact matches between the annotators are counted, we reach an agreement of 69.3%. However, note that this way of counting is very strict. For example, if an annotator assigns *'A'* and the other annotator assigns *'A/C'*, this approach does not consider the partial agreement for *'A'*.
- *Weighted.* To better report this partial agreement between both annotators, we also use a method that assigns a weight of 0.5 between instances of partial agreement as opposed to 1.0 between the absolute agreement. This weighted percentage approach yielded a 78.1% agreement.
- *Liberal.* Lastly, a more liberal approach considers the situation where either of the elements from

Approach	Agreement Percentage
Absolute	69.3%
Weighted	78.1%
Liberal	86.8%

Table 3: Inter-Annotator Percent Agreement Score

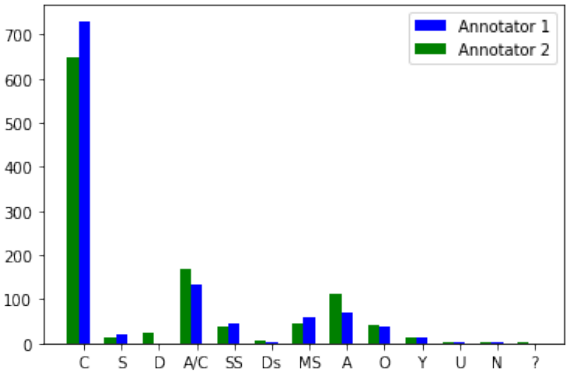

Figure 2: Feedback Token Labels distribution across annotators.

both annotators multi-category approach generates a match. The percentage agreement between both annotators yields a 86.8% agreement.

For our second approach we calculate the widely used Cohen Kappa Coefficient, which indicates the inter-reliability between two raters. To account for the fact that the annotators were allowed to assign up to two labels to a feedback token, we modified the Cohen Kappa Coefficient Calculation somewhat, so that cells in the confusion matrix can be assigned 0.5 in case of partial agreement. For example, if for a given feedback token, *annotator 1* assigns the label *'A'* and *annotator 2* assigns the label *'A/C'* we populate the cell $M_{A,A} = 0.5$ for partial agreement and $M_{A,C} = 0.5$ for partial disagreement. We report these values in a confusion matrix in Figure 3. The proposed calculation method gives a Cohen Kappa coefficient score of **0.51**. This score indicates moderate agreement between both annotators, slightly higher than the unmodified Cohen Kappa Coefficient calculation which results in a score of 0.484.

The Confusion matrix (Figure 3) indicates areas of agreement and disagreement between annotation classes of feedback tokens. We observe relatively strong agreement for most function categories between both annotators. The only instance where we observe strong disagreement is in function category *(D) Disagree*. This could possibly be explained by the low number of instances of this category in the corpus (as seen in Figure 2).

6. Preliminary Analysis

As a preliminary analysis, we studied the distribution of the lexical tokens per function. We also explored the differences in duration, mean pitch, pitch slope, and pitch range depending on the function using one-way ANOVA tests and Tukey Kramer post-hoc tests. The number of annotations per function, their duration and pitch characteristics are described in Table 4.

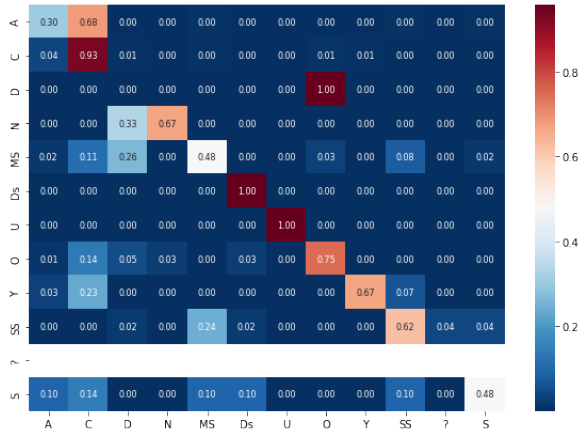

Figure 3: Normalized Confusion matrix between the annotators.

Lexical. In total, we found 59 unique lexical tokens which we identified as short feedback tokens. Figure 4 shows the distribution of lexical tokens which have a count of 8 or greater. For a detailed breakdown of the counts for all 59 unique lexical tokens and distribution within the function categories, see Table 5 in the Appendix.

As highlighted by Figure 2, function *(C) Continue* is the most frequent feedback function. This is in line with Jurafsky et al. (1998), who found that continuers were the most common type of backchannel in their annotation of Switchboard. Function *(C) Continue* also has the most variation in lexical tokens; *mhm* has the highest count followed by *yeah*, *uh-huh*, *right*, *sure*, *mmm*, *hm*, *okay* etc. In their analysis of the Columbia Games Corpus, Benus et al. (2007) also found that *mhm* was the most common backchannel. In this study, backchannels (their continuers) were *mhm*, *uh-huh*, *okay*, and *yeah* which overlap with our lexical tokens for *(C) Continue*.

Similar to Jurafsky et al. (1998), *yeah* is the most ambiguous lexical token. It is annotated in *(C) Continue, A/C, (A) Agree, (D) Disagree, (Y) Yes, (S) Sympathy* and *(SS) Strong Surprise* function categories. There are other lexical tokens that overlap in different functions such as *hm*.

Duration. Since there were errors in the start and end times of the word level alignments, we first revised them in order to calculate the duration for each feedback token. A one-way ANOVA test showed that there was a statistically significant difference in mean duration: $F(10,1616)=44.73$, $p<0.05$. Table 6 (see Appendix) shows the results of the Tukey test for

Function	Duration (s)	Mean Pitch (SD)	Pitch Slope (SD)	Pitch Range (SD)	Count
(C) Continue	0.329	-0.130	-0.092	1.674	740
(U) Non-Understanding	0.288	0.865	0.919	2.460	41
(A) Agree	0.389	-0.181	-0.307	1.795	201
A/C	0.320	-0.192	-0.172	1.595	182
(D) Disagree	0.360	0.240	-0.582	1.989	42
(Y) Yes	0.313	-0.110	-0.211	1.768	20
(N) No	0.357	0.043	-0.314	1.594	69
(S) Sympathy	0.580	-0.226	-0.572	2.061	18
(Ds) Disapproval	0.511	-0.160	-0.303	2.283	52
(MS) Mild Surprise	0.341	0.514	-0.817	2.675	87
(SS) Strong Surprise	0.511	0.065	-0.483	2.403	52

Table 4: Summary of duration, mean pitch, pitch slope, pitch range, and current count of annotations per function.

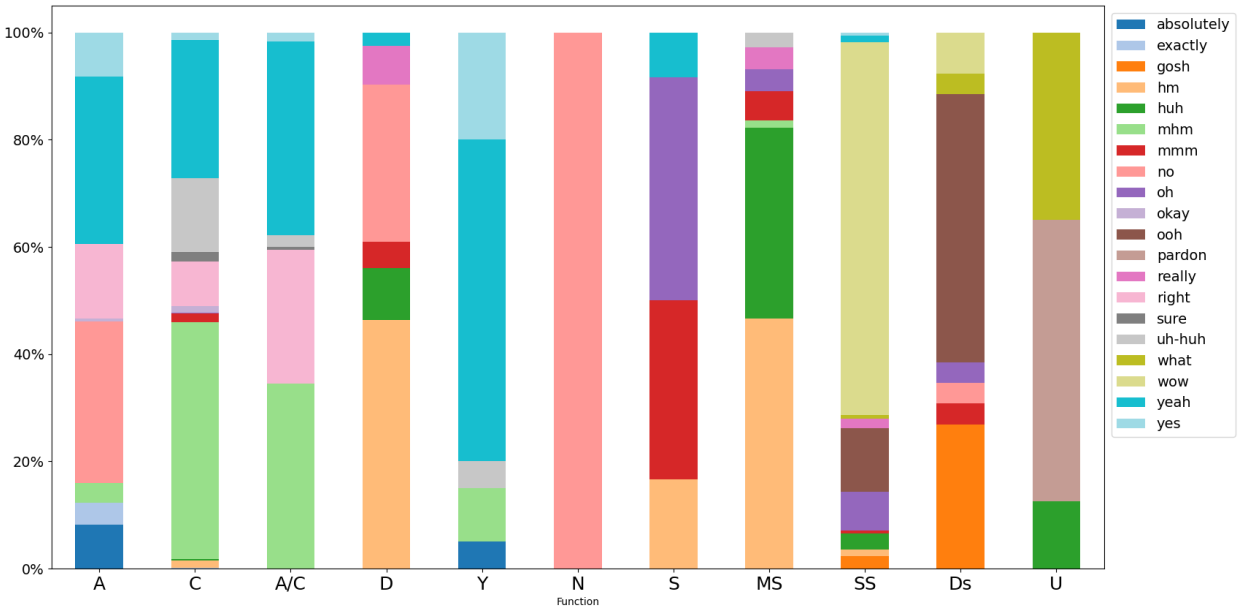

Figure 4: Distribution of lexical tokens per function.

duration; functions *(S) Sympathy, (SS) Strong Surprise, (Ds) Disapproval* have significantly longer duration compared to the other functions. When comparing the duration of agreements and backchannels, Benus et al. (2007) found that they were similar in duration. Jurafsky et al. (1998) found that continuers had shorter duration than agreement. We also find that *(A) Agree* has significantly longer duration than *(C) Continue*.

Mean Pitch. Pitch values were extracted using the World vocoder (Dinh et al., 2019), transformed to log scale, and z-score normalized per speaker, based on their entire speech. We eliminated 33 instances of C, 10 of A, 10 of A/C, 1 of Y, 1 of N, 1 of S, 2 of Ds, and 2 of SS due to technical problems during pitch extraction.

A one-way ANOVA test showed that there was a statistically significant difference in mean pitch: $F(10,1556)=19.13$, $p<0.05$. Table 7 (see Appendix) shows the results of the Tukey test for mean pitch. The results show that functions *(U) Non-understanding* and *(MS) Mild Surprise* have significantly higher mean pitch than *(SS) Strong Surprise, (N) No, (Y) Yes, (Ds) Disapproval, (A) Agreement,* and *(S) Sympathy*. We also observe that *(D) Disagree* and *(SS) Strong Surprise* have a significantly higher mean pitch than *(C) Continue* and *(A) Agreement*.

In their analysis of the Columbia Games Corpus, Benus et al. (2007) found that backchannels have a higher pitch than agreements. This is the opposite finding of (Jurafsky et al., 1998), who found that continuers have lower pitch compared to agreement. One reason why Benus et al. (2007) and Jurafsky et al. (1998) observed opposite effects might be due to the fact that Benus et al. (2007) only annotated affirmative words; *no*, which can be used to agree to negative statements, would not have been annotated in the Columbia corpus. In our results, we find no significant difference between *(C)*

Continue and *(A) Agree* in terms of mean pitch.

Pitch Slope. A one-way ANOVA revealed that there was a statistically significant difference in pitch slope; $F(10, 1556)=23.63$, $p<0.05$. Table 8 (see Appendix) shows the results of the Tukey test for pitch slope; function *(U) Non-Understanding* is the only function with a rising slope and it is significantly different from all the other functions. We also see that functions *(C) Continue, (A) Agree, (Y) Yes, (N) No, (Ds) Disapproval*, and *(SS) Strong Surprise* have significantly higher pitch slope than *(MS) Mild Surprise*.
Jurafsky et al. (1998) found that continuers are flatter than agreements. This is the opposite of what Benus et al. (2007) reported, when comparing agreement and backchannels, they found that backchannels have higher pitch slope than agreement. Again, the difference in observations may be due to the differences in the corpora. From our results we see that function *(C) Continue* is significantly different from *(A) Agree* and we also observe that *(C) Continue* has a flat pitch slope.

Pitch Range. A one-way ANOVA test showed that there was a statistically significant difference in pitch range; $F(10,1556)=13.23$, $p<0.05$. Table 9 (see Appendix) shows the results of the Tukey test for pitch range. The results show that functions *(MS) Mild Surprise*, *(U) Non-understanding*, and *(SS) Strong Surprise* have a significantly higher pitch range than *(A) Agree, (C) Continue* and *(N) No*. We also observe that function *(Ds) Disapproval* has a significantly higher pitch range than *(C) Continue*.

From our analysis we can see that it is important to consider both the lexical form and the prosodic characteristics in order to determine the function of the feedback tokens. When comparing *(C) Continue* and *(A) Agree* in terms of duration, mean pitch, and pitch slope to the findings of Jurafsky et al. (1998) and Benus et al. (2007), we observe both similarities and differences. The reasons why we observe differences might be due to the fact that Benus et al. (2007) only annotated affirmative words in the Columbia Games Corpus and even though Jurafsky et al. (1998) annotated the Switchboard corpus, the majority of their annotations were done without listening to the audio. Our proposed coding scheme encompasses functions with attitude, which have not been annotated and analyzed before. As more short feedback tokens are annotated with our scheme we hope to see clearer prosodic characteristics and lexical form distributions per function category.

7. Conclusion

In this paper, we present our annotation scheme which is an attempt at bringing the ideas of existing schemes into one single scheme, as there is no standard annotation scheme for annotating communicative functions of feedback. Our proposed annotation scheme encompasses functions with attitude information which provide us with a wider range of feedback functions. We also present our preliminary annotations of the communicative functions of short feedback tokens in the Switchboard corpus. So far, we have annotated 1627 instances of feedback. Our work is ongoing and we plan to release a complete dataset of the annotated short feedback tokens.
We report the distribution of lexical tokens per function and observe that some lexical tokens such as *mhm*, *yeah*, *hm* are ambiguous. Therefore, it is important to consider both the lexical form and prosodic features of feedback in order to determine the function. In our analysis, ANOVA tests revealed that there are significant differences in duration, mean pitch, pitch slope, and pitch range between functions. These prosodic characteristics that distinguish functions can be helpful for a feedback function classification task and for feedback generation.
For our future work, we plan to use our annotation scheme to annotate feedback in dyadic face-to-face spontaneous conversations in order to determine if there are differences in the prosodic realizations compared to those in Switchboard which are not face-to-face. From these feedback annotations, we plan to train a model to predict the prosody and lexical form of feedback given the communicative function. We will evaluate our generated feedback by the model in a conversational system.

8. Acknowledgements

The first and second authors have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588.

9. Bibliographical References

Adiba, A. I., Homma, T., and Miyoshi, T. (2021). Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412. IEEE.

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Axelsson, A., Buschmeier, H., and Skantze, G. (2022). Modeling feedback in interaction with conversational agents—a review. *Frontiers in Computer Science*, 4.

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Benus, S., Gravano, A., and Hirschberg, J. B. (2007). The prosody of backchannels in American English.

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., and Blache, P. (2021). A multimodal model for predicting conversational feedbacks. In *International Conference on Text, Speech, and Dialogue*, pages 537–549. Springer.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Buschmeier, H., Malisz, Z., Włodarczak, M., Kopp, S., and Wagner, P. (2011). Are you sure you're paying attention?'-Uh-Huh'communicating understanding as a marker of attentiveness. In *Twelfth Annual Conference of the International Speech Communication Association*.

Clark, H. H. (1994). Managing problems in speaking. *Speech communication*, 15(3-4):243–250.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Dinh, T., Kain, A., and Tjaden, K. (2019). Using a manifold vocoder for spectral voice and style conversion. In *INTERSPEECH*, pages 1388–1392.

Edlund, J., House, D., and Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *9th European Conference on Speech Communication and Technology*.

Edlund, J., Heldner, M., and Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In *Nordic Prosody-Proceedings of the Xth Conference*, pages 57–68. Citeseer.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (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.

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

Gravano, A., Benus, S., Hirschberg, J., Mitchell, S., and Vovsha, I. (2007). Classification of discourse functions of affirmative words in spoken dialogue. In *Eighth Annual Conference of the International Speech Communication Association*.

Jang, J. Y., Kim, S., Jung, M., Shin, S., and Gweon, G. (2021). BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452.

Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Discourse Relations and Discourse Markers*.

Kousidis, S., Malisz, Z., Wagner, P., and Schlangen, D. (2013). Exploring annotation of head gesture forms in spontaneous human interaction. In *Proceedings of the Tilburg Gesture Meeting (TiGeR 2013)*.

Malisz, Z., Włodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., and Wagner, P. (2016). The ALICO corpus: Analysing the active listener. *Language resources and evaluation*, 50(2):411–442.

Neiberg, D., Salvi, G., and Gustafson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55(3):451–469.

Ortega, D., Li, C.-Y., and Vu, N. T. (2020). Oh, Jeez! or uh-huh? A listener-aware backchannel predictor on ASR transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Prévot, L., Gorisch, J., and Mukherjee, S. (2015). Annotation and classification of French feedback communicative functions.

Prévot, L., Gorisch, J., and Bertrand, R. (2016). A CUP of COFEE: A large collection of feedback utterances provided with communicative function annotations.

Ruede, R., Müller, M., Stüker, S., and Waibel, A. (2019). Yeah, right, uh-huh: A deep learning backchannel predictor. In *Advanced Social Interaction with Agents*, pages 247–258. Springer.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

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

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Wallers, Å., Edlund, J., and Skantze, G. (2006). The effect of prosodic features on the interpretation of synthesised backchannels. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 183–187. Springer Verlag.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics*, 32(8):1177–1207.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

10. Appendix

Function	Lexical Tokens and Count
(C) Continue	mhm (321), yeah (188), uh-huh (100), right (60), sure (13), mmm (12), hm (10), yes (10), okay (9), huh-huh (4), huh (2), ah-hah (1), exactly (1), huh-hm (1), m-kay (1), mhm-hm (1), oh (1), uh-hah (1), uh-hm (1), well (1), yah (1), yuh (1)
A/C	yeah (65), mhm (62), right (45), uh-huh (4), yes (3), sure (1), yah (1), yup (1)
(U) Non-understanding	pardon (21), what (14), huh (5), sorry (1)
(A) Agree	yeah (61), no (59), right (27), absolutely (16), yes (16), exactly (8), mhm (7), indeed (1), nah (1), nuh (1), okay (1), true (1), yea-m (1), yep (1)
(D) Disagree	hm (19), no (12), huh (4), really (3), mmm (2), nah (1), yeah (1)
(Y) Yes	yeah (12), yes (4), mhm (2), absolutely (1), uh-huh (1)
(N) No	no (66), m-m (1), nuh (1), nuh-uh (1)
(S) Sympathy	oh (5), mmm (4), hm (2), uh-oh (2), awe (1), jeez (1), oof (1), uh (1), yeah (1)
(Ds) Disapproval	ooh (13), ugh (8), gosh (7), jeez (5), goodness (4), golly (2), wow (2), gee (1), jee (1), mmm (1), no (1), oh (1), oof (1), uh (1), uh-oh (1), what (1), wooh (1), yuck (1)
(MS) Mild Surprise	hm (34), huh (26), mmm (4), ah (3), goodness (3), oh (3), really (3), uh (3), interesting (2), uh-huh (2), eh (1), huh-uh (1), mhm (1), uh-ho (1)
(SS) Strong Surprise	wow (117), ooh (20), oh (12), huh (5), gosh (4), really (3), hm (2), yeah (2), ah (1), aha (1), goodness (1), jeez (1), mmm (1), uh (1), uh-oh (1), wah (1), what (1), yes (1)

Table 5: Distribution of lexical tokens and corresponding count per function.

Function	Duration (s)	Tukey, $p < 0.05$
(S) Sympathy	0.580	> A, D, N, MS, C, A/C, Y, U
(SS) Strong Surprise	0.511	> A, D, N, MS, C, A/C, Y, U
(Ds) Disapproval	0.511	> A, D, N, MS, C, A/C, Y, U
(A) Agree	0.389	> C, A/C, U
(D) Disagree	0.360	
(N) No	0.357	
(MS) Mild Surprise	0.341	
(C) Continue	0.329	
A/C	0.320	
(Y) Yes	0.313	
(U)	0.288	

Table 6: Tukey pair-wise test for duration. Empty cells indicate that the duration of the function was not significantly greater than other functions.

Function	Mean Pitch (SD)	Tukey, $p < 0.05$
(U) Non-understanding	0.865	> D, SS, N, Y, C, Ds, A, A/C, S
(MS) Mild Surprise	0.514	> SS, N, Y, C, Ds, A, A/C, S
(D) Disagree	0.240	> C, A, A/C
(SS) Strong Surprise	0.065	> C, A, A/C
(N) No	0.043	
(Y) Yes	-0.110	
(C) Continue	-0.130	
(Ds) Disapproval	-0.160	
(A) Agree	-0.181	
A/C	-0.192	
(S) Sympathy	-0.226	

Table 7: Tukey pair-wise test for mean pitch. Empty cells indicate that the mean pitch of the function was not significantly greater than other functions.

Function	Pitch Slope (SD)	Tukey, $p < 0.05$
(U) Non-understanding	0.919	> C, A/C, Y, Ds, A, N, SS, S, D, MS
(C) Continue	-0.092	> A, SS, D, MS
A/C	-0.172	> SS, D, MS
(Y) Yes	-0.211	> MS
(Ds) Disapproval	-0.303	> MS
(A) Agree	-0.307	> MS
(N) No	-0.314	> MS
(SS) Strong Surprise	-0.483	> MS
(S) Sympathy	-0.572	
(D) Disagree	-0.582	
(MS) Mild Surprise	-0.817	

Table 8: Tukey pair-wise test for pitch slope. Empty cells indicate that the pitch slope of the function was not significantly greater than other functions.

Function	Pitch Range (SD)	Tukey, $p < 0.05$
(MS) Mild Surprise	2.675	> A, C, A/C, N
(U) Non-understanding	2.460	> A, C, A/C, N
(SS) Strong Surprise	2.403	> A, C, A/C, N
(Ds) Disapproval	2.283	> C, A/C
(S) Sympathy	2.061	
(D) Disagree	1.989	
(A) Agree	1.795	
(Y) Yes	1.768	
(C) Continue	1.674	
A/C	1.595	
(N) No	1.594	

Table 9: Tukey pair-wise test for pitch range. Empty cells indicate that the pitch range of the function was not significantly greater than other functions.