Published in: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente et al. (Hg.) (2016): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 3180-3185.

A CUP of CoFEE: A Large Collection of Feedback Utterances Provided with Communicative Function Annotations

Laurent Prévot, Jan Gorisch, Roxane Bertrand
LPL: Aix-Marseille Université, Institut für Deutsche Sprache, LPL: Aix-Marseille Université
Aix-en-Provence (France), Mannheim (Germany), Aix-en-Provence (France)
laurent.prevot@lpl-aix.fr, gorisch@ids-mannheim.de, roxane.bertrand@lpl-aix.fr

Abstract

There have been several attempts to annotate communicative functions to utterances of verbal feedback in English previously. Here, we suggest an annotation scheme for verbal and non-verbal feedback utterances in French including the categories *base*, *attitude*, *previous* and *visual*. The data comprises conversations, maptasks and negotiations from which we extracted ca. 13,000 candidate feedback utterances and gestures. 12 students were recruited for the annotation campaign of ca. 9,500 instances. Each instance was annotated by between 2 and 7 raters. The evaluation of the annotation agreement resulted in an average best-pair kappa of 0.6. While the *base* category with the values *acknowledgement*, *evaluation*, *answer*, *elicit* and *other* achieves good agreement, this is not the case for the other main categories. The data sets, which also include automatic extractions of lexical, positional and acoustic features, are freely available and will further be used for machine learning classification experiments to analyse the form-function relationship of feedback.

Keywords: Conversational Feedback, Communicative Functions, Annotator Agreement

1. Introduction

Feedback utterances are among the most frequent in dialogue. Feedback is also a crucial aspect of linguistic theories taking interaction into account. For example it is often claimed that linguistic resources, such as the use of prosody, is used by interactional participants for specific communicative ends. In order to systematically test this relationship between communicative functions and the linguistic form of feedback, two tasks are required. One task is the extraction of linguistic features, including prosodic, lexical, positional, etc. features. The other task involves the annotation of communicative functions related to feedback utterances. In this paper we address the annotation task.

A range of communicative functions have been identified in previous research, however the annotation of feedback communicative functions is still a notoriously difficult task, as the inter-annotator agreement measures show (if the annotations are evaluated at all in such papers). The annotation involves an interpretative process that integrates various sources of information, such as auditory/acoustic and visual information.

The study reported in this paper takes place in the project CoFEE[1] (Prévot and Bertrand, 2012) that aims to use, among other methodologies, quantitative clues to decipher the form-function relationship within feedback utterances. More precisely, we are interested in the creation of (large) datasets composed of feedback utterances annotated with communicative functions. From these datasets, we conduce quantitative (statistical) linguistics tests as well as machine learning classification experiments.

We analyzed material from three corpora where participants (i) have a free conversation, (ii) do a Map Task, (iii) do a conversation that also involves some negotiation. First, we extracted all potential feedback: the verbal feedback utterances via the orthographic transcription, and the non-verbal feedback via a visual pre-segmentation of the videos. Second, we performed an annotation campaign with 12 raters who annotated 9500 instances. They were given annotation guidelines that are based on the annotation framework that we introduce in this paper. Third, we evaluated the inter-annotator agreement by measuring the best-pair kappa. The agreement ranges from good (for our base category) to weak for more complex peripheral annotations.

The outcome of the annotation task, i.e. the communicative function annotation, will be used subsequently for the analysis of the form-function relationship of conversational feedback.

The remainder of this paper is structured as follows. Previous work on feedback utterances is reviewed in Section 2. The material is introduced in Section 3. This is followed by the introduction to the annotation schema and guidelines in Section 4. The annotation infrastructure and campaign is presented in Section 5. The results and its evaluation of the annotation campaign are shown in Section 6..

2. Previous work

Concerning the definition of the term *feedback utterance*, we follow Bunt (1994, p.27): *"Feedback is the phenomenon that a dialogue participant provides information about his processing of the partner's previous utterances. This includes information about perceptual processing, about interpretation, about evaluation (agreement, disbelief, surprise,...) and about dispatch (fulfilment of a request,...)."*

The study of feedback is generally associated with *backchannels* (Yngve, 1970), the utterances that are not produced on the *main* communication channel in a way not to interfere with the flow of the main speaker, but on the *back* channel. In the seminal work by Schegloff (1982), back-channels have been divided between *continuers* and *assessments*. While *continuers* are employed by conversational participants to make a prior speaker continue with an ongoing activity, e.g. the telling of a story, *assessments* are employed to evaluate the prior speaker's utterance.

[1] http://cofee.hypotheses.org/

Grounded on Allwood et al. (1992) but more concerned with annotation constraints, especially in the context of multi-modal annotations, Allwood et al. (2007) use a simpler framework in which feedback analysis is split into three dimensions: (i) basic *(contact, perception, understanding)*; (ii) acceptance; (iii) emotion / attitudes that do not receive an exhaustive list of values.

More recent frameworks include work by Gravano et al. (2012) who propose a flat typology of affirmative cue word functions. This typology mixes *grounding* functions with *discourse sequencing* and other unrelated functions. It includes for example *Agreement, Backchannel*, discourse segment *Cue-Beginning* and *Cue-Ending* but also functions such as *Literal modifier*.

Neiberg et al. (2013) also adopt a form-driven approach but it is an approach that combines automatic data selection with lexical and acoustic cues. As for the function annotation, they identify five scalar attributes related to feedback: *non-understanding – understanding*, *disagreement – agreement*, *uninterested – interested*, *expectation – surprise*, *uncertainty – certainty*. This scalar approach is appealing because many of these values seem to have indeed a scalar nature.

Finally, the work around ISO-TC37 linguistic annotation standard (Bunt et al., 2012) provides a fine grained annotation schema for communicative functions. The framework identifies two dimensions for feedback: *Auto-feedback* concerns information processing by the feedback producer (*I have understood*), while *Allo-feedback* deals with information processing by interlocutor (*You have understood.*). The standard created distinguishes between *positive, negative* and *elicit* values for both *Auto-* and *Allo-feedback.*

This overview on related work on conversational feedback shows the large variety of approaches towards communicative function annotation. It might seem unnecessary to introduce a new modified schema, but as the interactional situations tend to differ across studies and in our case even within a study (cf. corpus section 3.1.), the annotation scheme has to be adjusted accordingly. Additionally, it is required to measure inter-annotator agreement in order to evaluate the validity of that schema, especially because the annotated categories are to be used in subsequent quantitative analyses involving further parameters in classification experiments.

3. Dataset

Our annotation schema should be able to cope with all feedback phenomena that appear in different interactional environments, including face-to-face interaction. Therefore we chose material from corpora that involve different recording situations and have high quality audio recordings and videos. This section introduces these corpora and the selection of verbal and non-verbal feedback instances, based on the orthographic transcription and a gesture pre-segmentation respectively.

3.1. Corpora

Our collection of feedback instances comes four different corpora: an 8 hour conversational data corpus (Bertrand and Priego-Valverde, 2008), a 2.5 hours MapTask corpus (Bard et al., 2010), a 2.5 hours face-to-face MapTask corpus (Gorisch and Prévot, 2014) and a 4 hours DVD negotiation corpus (Gorisch and Prévot, 2014). All these corpora are accessible as a collection of resources through the Ortolang-SLDR platform (Prévot et al., 2015). All recordings include headset microphone channels that were transcribed on IPU (Inter-Pausal Unit) level and automatically aligned on word and phone level. The first two corpora (CID and MTR) already existed before our current project, while the other two (MTX and DVD) were specifically recorded and transcribed for this project. All the details about the primary data can be found in Prévot et al. (2015a).

CID The Corpus of Interactional Data (CID) includes participants having a chat about "unusual situations" or "conflicts at work" (Bertrand et al., 2008). Each interaction took 60 minutes. Three of them were additionally recorded on video.

MTR The remote condition of the French MapTask corpus (MTR) (Bard et al., 2013) follows the original MapTask protocol (Anderson et al., 1991), where the role of map giver and follower change through the 8 maps per session.

MTX The face-to-face condition of the French MapTask corpus (MTX) (Gorisch et al., 2014) includes additional video recordings for both participants individually as they could see each other during the dialogue. Similar to the remote condition, 4 maps were "given" by one participant and "followed" by the other and the other way around.

DVD The DVD corpus is made of 8 dialogues of 30 minutes in which two participants negotiate and argue about a set of DVDs placed in front of them. The recordings comprise headset microphone channels and videos.

Post-processing Due to clocking differences in the audio and video recording devices and random image loss in the video, both signals ran out of synchronisation over time. For multimodal analyses, such desynchronisation is not acceptable. The videos of the CID corpus have been corrected by hand in order to match the audio channels. A more precise and less time-consuming procedure was developed for the newer recordings of MTX and DVD, as it is described by Gorisch and Prévot (2015).

3.2. Gesture pre-segmentation

As our project aims to describe conversational feedback in general, the visible part of that feedback should receive sufficient attention, too. Three of the four corpora include participants' visibility and video recordings[2]. An entire labeling of all gestures of the corpus was however impossible. Therefore, we employed two students who performed a pre-segmentation task. Those sections of a video that involve feedback in the domain of gestures or facial expressions were segmented using the ELAN tool in its segmentation mode (Wittenburg et al., 2006). The focus on this pass was on recall rather than precision since all the marked units

[2]In the CID-corpus, only 3 out of 8 sessions were video recorded and merely on a single camera, which made the identification of gestures difficult. Therefore, the results (cf. Figure 3) do not include gesture annotations for CID

were annotated later on for precise gestures and potentially discarded if it turned out that they are not feedback. This means that we were able to provide very broad instruction to catch any facial expression or head movement possibly involved in providing feedback.

3.3. Instance selection

We start from the observation that the vast majority of feedback utterances are Inter-Pausal Units composed of only a few tokens. We first identified the small set of most frequent lexical items (*ouais, oui, mh, ...*) composing feedback utterances by building the lexical tokens distribution for Inter-Pausal Units made of three tokens or less. The nine most frequent lexical forms are : *ouais / yeah (2781), mh (2321), d'accord / agree-right (1082), laughter (920), oui / yes (888), euh / uh (669), ok (632), ah (433), voilà / that's it-right (360)*. The next ones are *et / and (360), non / no (319), tu / you (287), alors / then (151), bon / well (150)*. We excluded *tu, et, alors* as we considered their presence in these short isolated IPUs were not related to feedback.
We then selected all isolated utterances in which these lexical items were included and treated each IPU as a data set instance. We also included in the data set sequences of lexical items presented above that were located at the beginning of an IPU. This yielded us a total of about 13000 candidate feedback utterances.

4. Annotation schema and guidelines

During the design process of the annotation schema, the most desirable aspects of the ones detailed in Section 2. should be preserved, while keeping the annotation process manageable. On the one hand, given the objective of the study, a fine grained study of the form - function relationship requires a fine grained functional analysis, too. On the other hand, a part of the literature clearly gave up annotating levels of communication in feedback, as for example Bunt et al. (2012). Moreover, an additional challenge (but worth pursuing in our opinion) is that our schema needs to be applied to various communicative situations. Finally, in addition to the functional annotation, several contextual features, such as pausing and overlap are extracted automatically in the project this study is based in. Therefore, the intention is to restrict the manual annotation to categories that require a deep understanding and interpretation of the ongoing interaction between the participants. The resulting schema is illustrated in the Type Feature Structure of Figure 1. In order to facilitate the annotation process for 'naive' raters, annotation guidelines provide paraphrases in colloquial language for each value to be used.

Base Despite the difficulty to annotate them reliably, levels of grounding are considered to be the central dimension of communication management analysis. This is first materialized by the BASE category and its values: *acknowledgement* and *evaluation*. These values follow the traditional entailment scale $acknowledgement > evaluation$. The *perception / understanding* difference is not considered here because in our communicative situations the environment is not noisy. We assume that participants were able to perceive all non-overlapping talk from the interlocutor.

Attitude This framework attempts to capture attitudinal reaction / evaluation aspects without multiplying the categories. The issue is that most of these attitudes are sparse in the data. We try to be as generic as possible and to propose pairs, i.e. elements that are strictly speaking not opposite but that are incompatible with each other.

Previous This category is the function of the previous utterance of the interlocutor. It is considered as being an approximation of the function of the feedback utterance's target. Using this feature for classification is circular but it is an interesting way to observe the dependency of feedback function on this specific contextual feature.

Visual For all visual feedback rough categories of facial expression and head movements were proposed. Examples were shown during the training to learn the rough categories.

5. Annotation Campaign

The aim to investigate conversational feedback from a qualitative and quantitative perspective demands for an annotation campaign that can take both into account: the detailed annotation schema and the relatively large collection of feedback instances.

Annotation infrastructure The annotation schema results with several tiers involving some dependencies and controlled vocabularies. The number of annotation files to be generated is above 200, which makes the ELAN GUI manipulation impractical for all the creation operations. This process is performed thanks to an open source Python ELAN API (Pympi, 2015). ELAN (Wittenburg et al., 2006) has been chosen for several reasons: controlled-vocabulary capabilities, possibility to extend the study to video data, familiarity of our students with the tool.

Annotation campaign The annotation campaign was conduced in two periods and involved two different groups of students (made of respectively 7 and 5 students). The campaign was realised on a duration of 2 months for most raters. Annotating one feedback instance took on average 1 minute. The annotated data set consisted of more that 9500 feedback items annotated by between 2 and 7 raters (most often 3).

6. Annotated Dataset

6.1. Results

The diagrams in Figure 2 show the distribution of the annotated categories in the data set composed of 9428 instances. The data set is available through the corpus resource page. It includes the annotation values (obtained through a vote among the raters) together with a set of lexical, positional and acoustic values that were automatically extracted in a parallel study.
The box-whiskers plots in Figure 3 show for each corpus the κ values for the best pair of raters according to the annotation categories. Each data point for these plots is a best-kappa score for one sample (or collection of samples annotated by the same raters in case of MTR and MTX). We can see that the value for the `base` category is always within reasonable boundaries and centered around 0.6. The agreement on this `base` but also on `amusement, attente`

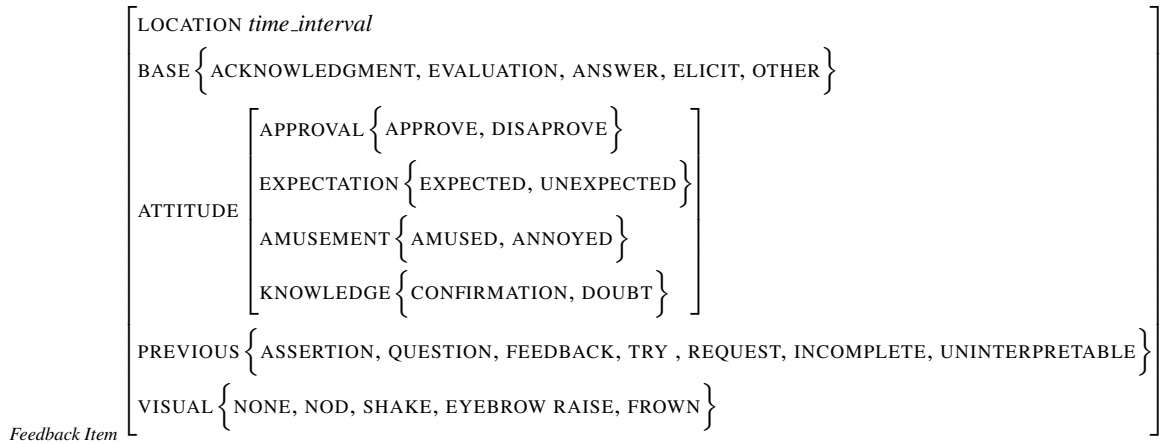


Figure 1: Type Feature Structure for the annotation of feedback items.

(a) BASE function

(b) EVALUATION function

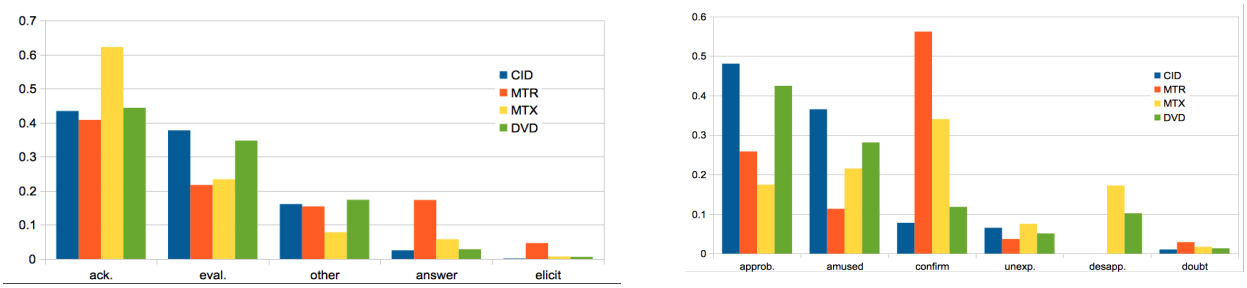


Figure 2: Distribution of annotated communicative functions across corpora.

and gest categories are satisfying. All the categories that exhibit a mean value below 0.4 should not be able to constitute a reference data set. The sub-optimal results were kept in the data set but tagged with a confidence score corresponding to the ratio of rater agreeing on the final decision (e.g 0.67 if 2 voters agreed out of three). .

6.2. Discussion

While best pair's κ seems to be a very favourable metric, most of our samples received only 3 concurrent annotations. Moreover, aside couple exceptions, it is always the same two raters that are excluded. As a result, what we call 'best-pair kappa' is actually simply the removal from the data set of the annotation of the worse two raters (which is a relatively standard practice). There is a reason for one of the rater to behave differently from others: Because of timing issues, he could not follow the training sessions with the others and had to catch up later.

7. Conclusion and on-going work

In this work, we ran a complete annotation campaign involving a dozen raters annotating about 9500 utterances with communicative functions. We presented our annotation schema, guidelines and campaign. The results of the annotation campaign in terms of reliability were rather satisfactory for the main functional category studied (average κ-scores around 0.6 for the best rater pair).

The annotation of communicative functions for feedback remains a difficult task. However, with the suggested annotation schema, a good number of annotators and high quality recordings, it is possible to achieve good inter-annotator agreement on large collections of instances.

The corpora and the data sets are freely available for the community through the ORTOLANG-SLDR platform. On our side, we will use the annotated data set for machine learning classification experiments of the kind we performed only on a partial data set so far (Prévot et al., 2015b) as well as for more specific and linguistic studies. In the future we plan to release the data set in XML format fitting the ISO-DiaML standard (Bunt et al., 2012).

8. Acknowledgements

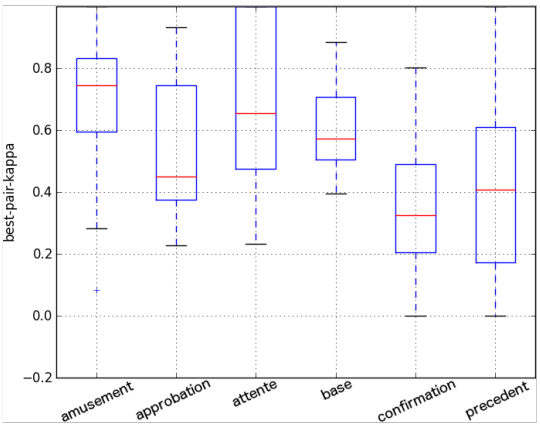
We would like to thank all our annotators for doing this difficult task. Without the support from the Agence National de la Recherche (ANR) for funding the project "Conversational Feedback" (Grant Number: ANR-12-JCJC-JSH2-006-01), this study would not have been possible.

9. Bibliographical References

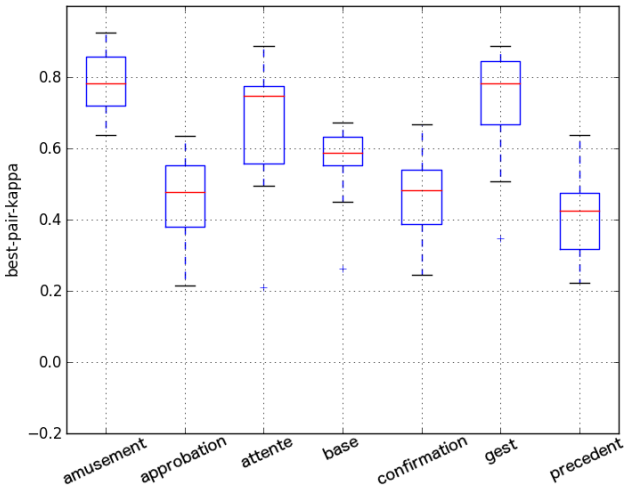
Allwood, J., Nivre, J., and Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequenc-

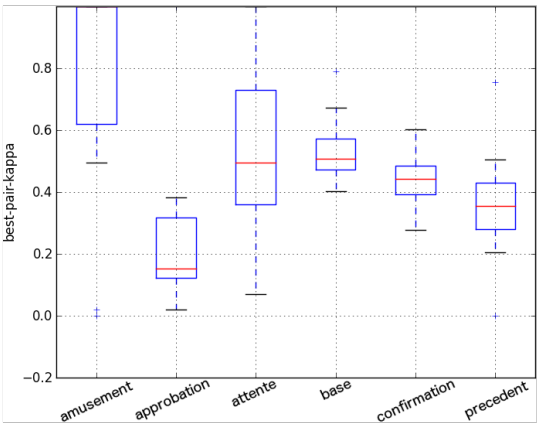
(a) Best pair of raters on CID



(b) Best pair of raters on DVD



(c) Best pair of raters on MTR



(d) Best pair of raters on MTX

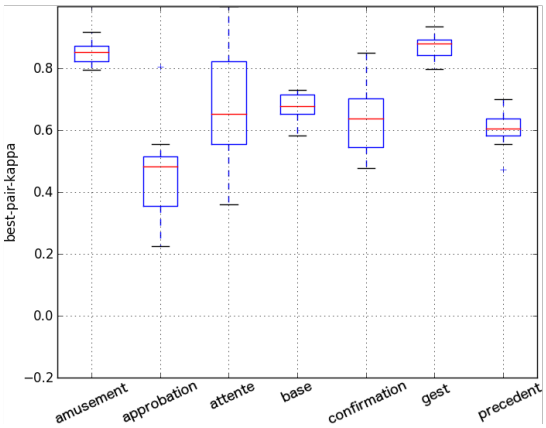


Figure 3: κ for the best rater pair for all corpora (*approbation = approval ; attente = expectation ; precedent = previous*)

ing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34:351–366.

Bard, E. G., Astésano, C., D'Imperio, M., Turk, A., Nguyen, N., Prévot, L., and Bigi, B. (2013). Aix Map-Task: A new French resource for prosodic and discourse studies. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, France.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le CID-Corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. R. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey.

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

Gorisch, J. and Prévot, L. (2015). Audio synchronisation with a tunnel matrix for time series and dynamic programming. In *Proceedings of ICASSP 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3846–3850, Brisbane, Australia.

Gorisch, J., Astésano, C., Bard, E., Bigi, B., and Prévot, L. (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)*, Reykjavik, Iceland.

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

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

Prévot, L. and Bertrand, R. (2012). Cofee-toward a multidimensional analysis of conversational feedback, the

case of french language. In *Proceedings of the Workshop on Feedback Behaviors*. (poster).

Prévot, L., Gorisch, J., Bertrand, R., Gorene, E., and Bigi, B. (2015a). A SIP of CoFee: A Sample of Interesting Productions of Conversational Feedback. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 149–153.

Prévot, L., Gorisch, J., and Mukherjee, S. (2015b). Annotation and classification of french feedback communicative functions. In *Proceedings of the The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*, Shangai.

Pympi, D. (2015). A python module for processing ELAN and Praat annotation files. `https://github.com/dopefishh/pympi`. Accessed: 2015-02-25.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some use of uh-huh and other things that come between sentences. *Georgetown University Round Table on Languages and Linguistics, Analyzing discourse: Text and talk*, pages 71–93.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC)*. Citeseer.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578.

10. Language Resource References

Ellen Bard and Corine Astésano and Cheryl Frenck-Mestre and Mariapaola D'imperio and Alice Turk and Noël Nguyen. (2010). *MAPTASK-AIX*. Laboratoire Parole et Langage - UMR 7309 (LPL, Aix-en-Provence FR), Speech and Language Data Repository (SLDR/ORTOLANG), ISLRN oai:sldr.org:sldr000732.

Roxane Bertrand and Béatrice Priego-Valverde. (2008). *Transcriptions du corpus CID*. Laboratoire Parole et Langage - UMR 7309 (LPL, Aix-en-Provence FR), Speech and Language Data Repository (SLDR/ORTOLANG), ISLRN oai:sldr.org:sldr000720.

Jan Gorisch and Laurent Prévot. (2014). *Audio-visual condition of Aix Map Task*. Laboratoire Parole et Langage - UMR 7309 (LPL, Aix-en-Provence FR), Speech and Language Data Repository (SLDR/ORTOLANG), ISLRN oai:sldr.org:sldr000875.

Laurent Prévot and Jan Gorisch and Roxane Bertrand. (2015). *CoFee project Corpus Collection*. Laboratoire Parole et Langage - UMR 7309 (LPL, Aix-en-Provence FR), Speech and Language Data Repository (SLDR/ORTOLANG), ISLRN oai:sldr.org:sldr000911.