Annotating Conversational Grunts in Video Data

Richard Brutti

updated: March 17, 2025

1 Introduction

In this document, we introduce an annotation scheme for the classification of conversational grunts. A conversational grunt is an intentional non-lexical vocalization (NLV) that occurs as part of the stream of speech. This includes filled pauses, interjections, backchannels, and similar, but excludes other audible noises like sniffing, crying, swallowing, or laughing. *Oh* and *uh-huh* are emblematic examples. For this project, we exclude semi-lexicalized utterances like *okay* and *yeah*.

The aim of the annotation is to find NLVs, provide a loose transcription, and describe their role in a multi-party interaction, in terms of Dialogue Acts.

1.1 Purpose of Annotations

This annotation scheme will help us analyze the functional role of non-lexical vocalizations in conversation, understand their distribution across different dialogue contexts, with a goal of improving computational models of conversational dynamics in multimodal settings.

1.2 Scope of the Project

This project analyzes conversational dynamics within the Weights Task Dataset (WTD) (Khebour et al., 2024a). The WTD consists of 10 recorded collaborative problem-solving sessions, each featuring 3 participants working together. In each session, participants use a scale to determine the weights of five blocks. They are provided with the weight of one reference block (10 g) and must collaboratively figure out the weights of the remaining blocks. This dataset offers a rich environment for studying non-lexical vocalizations in the context of collaborative reasoning and problem-solving interactions.

2 ELAN Software

For this annotation project, we use ELAN. ELAN is an annotation tool for the creation of complex annotations on video and audio resources (Brugman, Russel, and Nijmegen, 2004).

2.1 ELAN Resources

- Official ELAN website
- User Manual

3 Project Overview

3.1 Research Questions

We argue that events significantly contribute to the multimodal context for speech, and can aid in automatic understanding and interpreting of dialogues, from the perspective of the Common Ground (Grice, 1975; Clark and Brennan, 1991).

Additionally, grunts, or non-lexical utterances, play a key role in conversations. In this presentation, I propose an annotation scheme for the classification of conversational grunts (intentional non-lexical conversational sounds that occur as part of the stream of speech (Ward, 2006). The annotation scheme is designed to classify the pragmatic functions of conversational grunts, considering their phonetic and prosodic form. The approach is inspired by work on laughter (Ginzburg, Mazzocconi, and Tian, 2020; Mazzocconi, Tian, and Ginzburg, 2020), which we extend by identifying the *gruntable* as the event or utterance that proceeds a conversational grunt.

4 Technical Setup

4.1 File Organization

- All video files are stored in [location]
- Each annotator will work on files assigned in [spreadsheet/system]
- Save annotation files (.eaf) using the naming convention: [convention, e.g., "VideoID_AnnotatorInitials_YYYYMMDD.eaf"]

5 Annotation Tiers and Structure

5.1 Tier Hierarchy

The annotation file includes pre-existing tiers containing automatic speech recognition (ASR) transcripts for each participant in the Weights Task Dataset.

5.2 Provided Tiers

The ELAN files come pre-annotated with the following provided tiers:

- **researcher**: ASR transcript for the researcher/facilitator

- **1**: ASR transcript for Participant 1
- **2**: ASR transcript for Participant 2
- **3**: ASR transcript for Participant 3

All provided tiers are independent (no parent tier), time-aligned, and contain automatically generated speech transcriptions for each speaker. These tiers should be used as reference for identifying where non-lexical vocalizations occur in the interaction.

5.3 Annotation Tiers

1. **Grunts**
    - Purpose: For annotating non-lexical vocalizations (NLVs) or conversational grunts. Annotations should last the duration of the spoken grunt.
    - Annotation format: For each identified grunt, annotators should include the following information in the annotation text field, separated by commas:
        - Speaker: Indicate who produced the grunt (researcher, p1, p2, p3)
        - Transcription: Provide a phonetic transcription of the NLV (e.g., "oh", "mm", "uh")
        - Dialogue Act: Specify the communicative function of the grunt. Conventions for each Dialogue Act in section 6.
        - Expectedness: Note whether the gruntable was expected in the context (Yes/No) [omit for FILLED_PAUSE]
        - Polarity: Indicate (+) for positive or (-) for negative [omit for FILLED_PAUSE]
        - Standalone: Specify whether the grunt occurs on its own (Yes) or as part of a larger utterance (No) [omit for FILLED_PAUSE]
    - Example: "p1, um, STATEMENT, yes, +, no"
2. **Gruntables**
    - Purpose: For annotating utterances or events that grunts are responding to
    - Annotation format: For each gruntable, annotators should include the following information in the annotation text field, separated by commas:
        - Speaker: Indicate who produced the speech or initiated the event (researcher, p1, p2, p3)

  - Type: Specify whether the gruntable is: event, speech, speech describing an event (DE), speech describing a gesture (DG), or other
  - Content: Transcribe the gruntable utterance or provide a brief description of the event (e.g., "scale does not balance")
  - Dialogue Act: Specify the communicative function of the gruntable using the list of Dialogue Acts below
- Example: "p2, speech, that one is 50, STATEMENT"

6 Annotation Conventions

6.1 Time Alignment

- Start annotations exactly when the phenomenon begins
- End annotations exactly when the phenomenon ends

6.2 Definitions

The following definitions should be used when annotating:

6.2.1 Dialogue Acts

Dialogue Acts are adopted from (Khebour et al., 2024b), which are adopted from the Dialogue Game Board (Ginzburg, 2012).

- **STATEMENT**: A factual assertion about a block, the world, or the procedure. (Note: Events should be interpreted as STATEMENTs if they provide new information about the scenario.)
- **ACCEPT**: Agreement with a previous STATEMENT or OBSERVATION
- **DOUBT**: Disagreement or uncertainty regarding a previous STATEMENT
- **QUESTION**: Yes/no questions or wh-questions seeking information
- **ANSWER**: A response to a QUESTION
- **RECOMMENDATION**: High-level planning about the block weights (e.g., "Let's try yellow next")
- **OBSERVATION**: An utterance that does not reflect a known fact (e.g., "it seems pretty balanced"). This can also apply to off-task utterances.
- **INFERENCE**: Bridges between OBSERVATION and STATEMENT/DOUBT, reflecting the deduction process of participants from what they observed to what they conclude. (Note: It is rare for a grunt to fill this role).

- **FILLED_PAUSE**: Used for filled pauses, turn starters, and disfluencies (Note: when this dialogue act is selected, annotators do not need to include Expectedness, Polarity, or Standalone information). Filled pauses do not require a preceding gruntable.
- **OTHER**: Any dialogue act that doesn't fit into the categories above

6.2.2 Polarity

- **Positive (+)**: The non-lexical vocalization affirms or agrees with the gruntable (speech or event) it relates to
- **Negative (-)**: The non-lexical vocalization negates or disagrees with the gruntable (speech or event) it relates to

6.2.3 Expectedness

- **Yes**: The information contained in the gruntable (speech or event) is reasonably expected in the context
- **No**: The information contained in the gruntable (speech or event) is unexpected or surprising

References

Brugman, Hennie, Albert Russel, and Xd Nijmegen (2004). "Annotating Multimedia/Multi-modal Resources with ELAN". In: *LREC*. Lisbon, pp. 2065–2068.

Clark, Herbert H and Susan E Brennan (1991). "Grounding in communication." In.

Ginzburg, Jonathan (2012). *The interactive stance*. Oxford University Press, USA.

Ginzburg, Jonathan, Chiara Mazzocconi, and Ye Tian (2020). "Laughter as language". In: *Glossa: a journal of general linguistics (2021-...)* 5.1.

Grice, Herbert P (1975). "Logic and conversation". In: *Speech acts*. Brill, pp. 41–58.

Khebour, Ibrahim et al. (2024a). "When text and speech are not enough: A multimodal dataset of collaboration in a situated task". In: *Journal of open humanities data* 10.

Khebour, Ibrahim Khalil et al. (2024b). "Common Ground Tracking in Multimodal Dialogue". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3587–3602.

Mazzocconi, Chiara, Ye Tian, and Jonathan Ginzburg (2020). "What's your laughter doing there? a taxonomy of the pragmatic functions of laughter". In: *IEEE Transactions on Affective Computing* 13.3, pp. 1302–1321.

Ward, Nigel (2006). “Non-lexical conversational sounds in American English”. In: *Pragmatics & Cognition* 14.1, pp. 129–182.