Uh, Actually: Grunt Acts as Grounding Contributions in Multimodal Dialogue

A Dissertation

Presented to

The Faculty of Brandeis University

Department of Computer Science

James Pustejovsky, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

Richard Brutti

August 2026

This dissertation, directed and approved by Richard Brutti's Committee, has been accepted and approved by the Faculty of Brandeis University in partial fulfillment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Charles Golden, Vice Provost for Graduate Affairs

Dissertation Committee:

James Pustejovsky, Computer Science, Chair
Nianwen Bert Xue, Computer Science
Sophia Malamud, Linguistics
Candace Sidner, Computer Science, Worcester Polytechnic Institute

Acknowledgements

First, sincere thanks to my advisor James Pustejovsky, who welcomed me back to the department eight years after my master's. Though I started remotely and mid-pandemic, he let me take grunts seriously.

Thanks to my committee - Candace Sidner, Nianwen Xue, and Sophia Malamud - for the thoughtful questions and suggestions along the way.

So much of this dissertation grew out of collaborative projects. Many thanks to my co-authors and annotators who played with blocks and debated the finer points of non-verbal communication with me.

All my love to Rachel, Noah, Ben, and Jacob. Your grunts and nods mean everything to me.

ABSTRACT

Uh, Actually: Grunt Acts as Grounding Contributions in Multimodal Dialogue

A dissertation presented to the Faculty of Brandeis University
Waltham, Massachusetts

By Richard Brutti

This dissertation argues that non-lexical communicative acts and actions are first-class contributions to the common ground in multimodal dialogue. There is more than *oh* going on. The acknowledging minimal vocalizations, conversational grunts like *oh*, *mm-hm*, and *uh-huh*, are among the primary signals through which grounding occurs in co-situated interaction. Despite their frequent appearance, these phenomena are systematically excluded from computational models of dialogue, which assume that the propositional content of an interaction is carried by its lexical content. This dissertation develops the formal and empirical steps to update that assumption.

The dominant frameworks in dialogue modeling reflect this assumption in concrete ways. Dialogue state tracking systems represent information state as slot-value pairs populated by utterances. Meaning representation frameworks were designed for written sentences, and while their extensions can handle speech acts, they have no principled account of non-lexical modalities. Automatic speech recognition (ASR) systems explicitly filter out the vocalizations that are most communicatively relevant: OpenAI's Whisper removes *hmm*, *mm*, *mhm*, *uh*, and *um* as a matter of policy. BERT has been shown to use similar representations for sentences that have similar meaning with or without so-called disfluencies. Large language models are pre-trained predominantly on written text, and written language conventions actively suppress the non-lexical vocalizations that are ubiquitous in speech. The result is that the signals that most directly encode grounding in situated dialogue are invisible to the systems designed to model it.

The first contribution is Gesture AMR (GAMR), an extension of Abstract Meaning

Representation that captures the semantic content of gesture through a taxonomy of gesture act relations. GAMR is realized in an annotated corpus of multimodal speech and gesture. The theoretical result established by this work is foundational: non-lexical communicative acts have representable propositional content, and the formal machinery for capturing that content can be extended beyond gesture to other non-lexical modalities.

The scaffolding for the grunt analysis is developed through collaborative work. The Weights Task Dataset, a multimodal corpus of collaborative problem-solving, provides the primary empirical resource for the dissertation and was designed specifically to elicit genuine multimodal communication in a co-situated, shared physical environment. The Common Ground Tracking (CGT) framework operationalizes multimodal common ground formally as a dynamic structure of questions under discussion, evidenced propositions, and accepted facts, updated by dialogue moves whose formal semantics are specified using Evidence-Based Dynamic Epistemic Logic (EB-DEL).

The second major contribution is a systematic annotation and formal model of grunts as grounding contributions. Drawing on approximately three hours of collaborative task interaction, an annotation scheme is presented that captures grunt tokens, their conversational triggers, and their dialogue act functions. The central finding is that grunts respond to speech and observable events at nearly equal rates, demonstrating that non-verbal events function as conversational contributions on par with utterances. Grunts are formalized as operations within EB-DEL that promote evidenced propositions to accepted facts in the common ground.

The third contribution is a prosodic analysis showing that token selection is the primary signal of grounding source, with prosody operating as a secondary channel. Population-level acoustic differences are largely absorbed by token identity, but where the token is held constant, in standalone *oh* in the Portal corpus, event-following productions are significantly higher and wider in pitch than speech-following ones. This is exploratory evidence that speakers can modulate acoustics to signal what is being grounded.

The fourth contribution extends the analysis to the structural dimension of turn position. Using a four-way positional typology (standalone, turn-initial, turn-medial, turn-final) annotated across all three corpora, position is shown to align with the same functional grouping found in the token and acoustic analyses acknowledgment tokens such as *mm-hm* are almost always standalone complete turns, while such as *oh* are predominantly turn openers. In the Weights Task, along with a furniture assembly collaborative video (Glider), position also tracks grounding source; speech-following grunts tend to stand alone as complete turns, while event-following grunts disproportionately open turns. In the faster-paced Portal, this positional contrast flattens somewhat. Across all three encoding channels, the speech–event distinction is found to be marked redundantly rather than carried by any single feature.

The Portal Dialogue Corpus, a structurally distinct virtual collaborative setting, serves as a test of generalization. Grunts are used differently to ground speech versus observable events across all three corpora, while their surface realization is likely shaped by the task. Taken together, this dissertation demonstrates that a complete account of grounding in multimodal dialogue must bring these small sounds into the formal and evidential picture. The grunt work, it turns out, is doing quite a lot.

TABLE OF CONTENTS

1 Introduction 1
1.0.1 What are grunts? 2
1.1 The Challenge with Verbal-only Dialogue Models 3
1.2 Grounding, Common Ground, and the Multimodal Challenge 5
1.3 Contributions 7
1.4 Organization 8
2 Representing Non-Lexical Acts: Gesture AMR 11
2.1 Gesture and Semantics 12
2.2 Gesture Act Taxonomy 13
2.2.1 Deictic Gesture *(deixis-GA)* 15
2.2.2 Emblematic Gesture *(emblem-GA)* 15
2.2.3 Iconic Gesture *(icon-GA)* 16
2.2.4 Metaphoric Gesture *(metaphor-GA)* 17
2.2.5 Gesture Units and Coordinated Gestures 18
2.3 Alignment in GAMR 20
2.4 Encoding Gesture in Multimodal Dialogue: The Multimodal AMR Corpus . 22
2.4.1 AMR and GAMR 22
2.4.2 Multimodal Multi-sentence AMR 23
2.5 Representable Content of Non-Lexical Acts 24
3 Theoretical Foundations: Grounding in Multimodal Dialogue 26
3.1 The Weights Task Dataset 26
3.1.1 Task Design and Motivation 27
3.1.2 Annotation Layers 28
3.2 Annotating Situated Actions in Dialogue 28
3.2.1 Actions as Discourse Contributions 28
3.2.2 Annotating Actions with AMR 28

3.2.3 Action Annotation Challenges . . . 29
3.3 Common Ground Tracking in Multimodal Dialogue . . . 30
3.3.1 Common Ground Structure . . . 30
3.3.2 Annotation and Experimental Pipeline . . . 33
3.4 A Model of Information State in Situated Multimodal Dialogue . . . 34
3.4.1 Integrating Speech, Gesture, and Action . . . 34
3.5 From Gesture to Grunt . . . 35
4 Grunts . . . 37
4.1 Corpora . . . 37
4.1.1 Weights Task Dataset . . . 37
4.1.2 Glider . . . 37
4.2 Annotation scheme . . . 38
4.2.1 What is a Grunt? . . . 38
4.2.2 Grunt Annotations . . . 39
4.2.3 Gruntable . . . 41
4.2.4 Annotation Procedure . . . 42
4.2.5 Inter-Annotator Agreement . . . 44
4.3 Formal Model . . . 45
4.3.1 Framework . . . 46
4.3.2 Detailed Walkthrough: Weights Task . . . 46
4.3.3 Illustrative Example: Cooking Scenario . . . 48
4.4 Findings . . . 49
4.4.1 Overall Token Distribution . . . 49
4.4.2 Token Specialization by Antecedent Type . . . 51
4.4.3 Task Effects: Glider vs. WTD . . . 51
4.4.4 Grunts as Complete Conversational Turns . . . 53
4.5 Summary . . . 54

5 Prosodic Encoding of Grounding Function 56
5.1 Motivation: Form and Function in Minimal Responses 56
5.1.1 Corpus and Preprocessing 57
5.2 Token Specialization by Antecedent Type 60
5.3 Prosodic Encoding of Grounding Source 61
5.3.1 ACCEPT Responses by Antecedent Type 61
5.3.2 Disentangling Token and Prosodic Modulation 62
5.3.3 The Speech-Describing-Event Category 63
5.3.4 Prosodic Variation by Dialogue Act 64
5.4 Discussion 65
5.5 Summary 66
6 The Portal Dialogue Corpus 67
6.1 Portal Distributional Analysis 68
6.2 Portal Token Specialization by Antecedent Type 71
6.3 Portal Prosody 73
6.4 Grunt Token Entrainment in Portal 75
6.5 Conclusion 76
7 Grunt Turn Position 78
7.1 Turn Position Typology and Annotation 78
7.2 Position by Antecedent 79
7.3 Position by Token 81
7.4 Conclusion 82
8 Conclusions and Future Directions 84
8.1 Summary of the Dissertation 84
8.2 How Grounding Source is Encoded 85
8.3 Caveats and Limitations 88
8.4 Future Directions 90

8.5 Conclusion 92
Bibliography 94

LIST OF TABLES

2.1 Distribution of gesture act types in EGGNOG annotations. 23

4.1 Grunt token inventory by sub-corpus. 43

4.2 Inter-annotator agreement for grunt detection and dialogue act classification. 44

4.3 Inter-annotator agreement on gruntable type for a one-hour segment ($n = 20$ gruntables, both annotators identifying the same spans). Krippendorff's α (nominal). 45

4.4 Grunt dialogue act distribution across Weights Task and Glider 50

4.5 Antecedent type distribution across Weights Task and Glider. 50

4.6 Acoustic features of standalone grunt responses by antecedent type (Weights Task). Mean (SD). 54

5.1 Acoustic features by antecedent type (ACCEPT responses, mean acoustic features). F0 reported in semitones relative to each speaker's median. 61

6.1 Grunt token inventory for the annotated portion of the Portal Dialogue Corpus 69

6.2 Grunt dialogue act distribution across Weights Task, Glider, and Portal . . . 70

6.3 Antecedent type distribution across Weights Task, Glider, and Portal 71

6.4 Expressive and reactive ACCEPT tokens in Portal outside the six-token inventory shared across corpora ($n = 90$, 28.8% of Portal ACCEPTs). Nearly all follow events. 73

6.5 Acoustic features of standalone ACCEPT grunts in Portal by antecedent type. Mean (SD). F0 in semitones relative to each speaker's median. 74

7.1 Turn position distribution by corpus (% of grunts). 79

7.2 Turn position by grounding source, per corpus. Each row gives the percentage of that corpus's speech- or event-following grunts in each position. Within the Weights Task, event-following grunts are turn-initial more than twice as often as speech-following grunts; in Portal the contrast is absent. 79

7.3 Turn position of ACCEPT grunts by grounding source, per corpus. Within the Weights Task, event-following ACCEPTs are turn-initial about three times as often as speech-following ones; in Portal position does not distinguish the two. 81

7.4 Turn position by token (percentage of each token's occurrences, pooled across corpora). 82

1 INTRODUCTION

Consider two people putting together a piece of flat-pack furniture. The first tightens a bolt, then the second responds *mm-hm*. The first turns the page of the instruction booklet and the task continues immediately, without further verbal confirmation and without explicit acknowledgment of what was accomplished. This minimal vocalization (non-lexical, non-propositional, and lasting a fraction of a second) has done real communicative work. It signals that the completed action has been perceived, integrated, and accepted as part of the shared record of the interaction between the two assemblers. The participants are able to proceed with the subsequent coordinated action (no matter what the Swedish cartoon instructions describe next), not because one spoke, but because one grunted.

This dissertation is about non-verbal communication, especially grunts, and how they contribute to the common ground as grounding acts between participants in a multimodal interaction [1].

Before proceeding, I want to clarify what this dissertation means by the term *grunt*. I focus the study on English spoken in the United States of America. These sounds have been studied under a variety of names: backchannels, filled pauses, disfluencies, interjections, paraverbals, hesitation markers, and planning signals to name a few. I largely follow Ward (2006) by treating conversational grunts as intentional non-lexical vocalizations that occur in the stream of speech and have systematic sound-meaning correspondence. Common examples include *oh*, *mm-hm*, *uh-huh*, and *hm*. Less common forms like *oof* and *ooh* are also in scope. I also follow Wilkins (1992) who treats them as propositional acts with goals. Like Ward (2006), I use *grunt* as a cover term. Sighs, laughs, clicks, as well as more lexicalized forms like *ok* and *yeah* are out of scope. Sighs and clicks are made during speech but typically use a different group of sounds and methods than speech (Ogden 2013). *Yeah* and *ok* were included in early drafts of the annotation scheme, but proved to be extremely common in the data, vastly outnumbering the other tokens. Also, these are more conventional lexemes.

[1] I looked for an etymological link between grunt and ground. There is not one as far as I could tell

The grunt tokens that I analyze sit between these two groups in a sense: more word-like than clicks and sighs, but less lexicalized than *yep* and *okay*. I use *token* throughout for the recurring vocalization form in a descriptive sense, without claiming that they have full word status.

1.0.1 What are grunts?

Mentioning this topic often brings to mind apes, babies, and early humans, but the grunts studied here are made by fluent adult human speakers in ordinary conversation. Non-lexical vocalizations have a long and varied descriptive history. They attract scholarly attention from a wide range of perspectives, and popular interest (Erard 2008). They go by various names, such as *interjections*, *paraverbals*, and *non-lexical vocalizations* and can be treated everywhere from *disfluencies* to words in their own right. Clark and Fox Tree (2002) argue that *uh* and *um* are purposeful and distinct studies. Yngve (1970) introduced the term *backchannel* for a listener's vocal feedback, and Ward (2006) argues that there are distinct meanings associated to different grunt tokens. Work on discourse structure considers a wider class of *discourse markers* that impact conversational segmentation and coherence (Schiffrin 1987).

The tradition of Conversation Analysis treats these tokens as critical to the interaction between speakers (Sacks, Schegloff, and Jefferson 1974). Schegloff (1982) treats *uh-huh* and *mm-hm* as *continuers* that allow the main speaker to maintain the floor. Heritage (1984) analyzes *oh* as marking a shift in the speaker's state of knowledge, and Gardner (2001) has a fuller account of *response tokens* and their roles. Keevallik (2014) extends this tradition to the body, showing that non-lexical vocalizations can coordinate with a speaker's own bodily conduct.

There is also a traditional of analysis on non-lexical objects that I draw outside the circle of grunts. Sighs (Hoey 2014), sniffs (Hoey 2020), and clicks (M. Wright 2011; Ogden 2013) help organize conversation but fall outside of speech.

A computational and formal strand models these phenomena more explicitly. Gravano and Hirschberg (2011) identify prosodic and lexical backchannel-inviting cues in task-oriented dialogue, and Hunter, Asher, and Lascarides (2018) give a formal semantics for situated conversation, in which non-linguistic events enter the discourse. Neither models the grounding relation as one whose antecedent may be either speech or an observable event, which this dissertation makes central.

1.1 The Challenge with Verbal-only Dialogue Models

Sophisticated language technologies have made progress toward modeling human dialogue over the course of this project. Large language models (LLMs) generate responses to complex prompts; dialogue state tracking (DST) can maintain user intents across several conversational turns; automatic speech recognition (ASR) can achieve near-human performance under the right conditions. Yet all of these advances share a common untenable assumption that the propositional content of a dialogue is carried only by its lexical content.

This assumption manifests in several ways. LLMs are predominantly used in text chat formats, despite advances in coding and multimodal LLM capabilities. Dialogue State Tracking (DST) frameworks (Budzianowski et al. 2018; Henderson, Thomson, and Young 2014) represent information state as a structure of slot-value pairs populated by text input from a chat interface. Budzianowski et al. (2018) uses a text interface, and the system described in Henderson, Thomson, and Young (2014) focus solely on the slot-filling values that come from the speech recognition layer. Meaning representation frameworks such as Abstract Meaning Representation (AMR) (Banarescu et al. 2013) were designed to capture the propositional content of written sentences. Extensions to AMR for dialogue (Bonial et al. 2020) handle illocutionary force and speech acts, but have no principled account of non-lexical modalities. ASR systems, such as OpenAI's Whisper, are commonly trained to filter out or drop "disfluency" tokens (Radford et al. 2023). In their *Text Standardization* section, the authors explain:

> We perform the following steps to normalize English texts in different styles into a standardized form, which is a best-effort attempt to penalize only when a word error is caused by actually mistranscribing a word, and not by formatting or punctuation differences. [...] 3. Remove any of the following words: `hmm, mm, mhm, mmm, uh, um`

As discussed in Erker and Vidal-Covas (2022), removing "filler words" is a selling point for commercial ASR and audio editing suites such as Otter.ai [2] and Descript [3]. BERT (Devlin et al. 2019) has been shown to use close underlying representations when semantically similar sentences are presented with or without a so-called disfluency (Tian et al. 2021). The same impulse appears in popular communication advice, where removing filled pauses and other non-lexical sounds might even improve your life (Hoeppner 2025). Recent theoretical work in super linguistics has argued for a unified treatment of linguistic and paralinguistic phenomena under a single analytical framework based in linguistic concepts (Patel-Grosz et al. 2023). Some of my and our prior work has begun to address the representational challenge of multimodality by extending Universal Meaning Representation (UMR) (Van Gysel et al. 2021) to situate multiple communicative channels, namely speech and gesture (Lai, Brutti, et al. 2021). Prior schemes have tended to annotate modalities separately rather than treating their interactions with each other as a means of emergent meaning, such as BML (Kopp, Krenn, et al. 2006).

The framework underlying these systems assumes that communication is fundamentally linguistic, and leads to a mischaracterization of how meaning is negotiated and constructed in human interaction. Human-computer interaction (HCI) and human-robot interaction (HRI) involve communicating intentions, goals, and attitudes through multiple modalities beyond language, including gesture, gaze, and prosodic manipulation. As systems become more advanced, there are increasing expectations that they will approximate interacting with other humans. In situated, co-present dialogue, participants gesture, interact with shared objects in their environment, and observe actions and their results, in addition to speaking

[2] https://otter.ai/
[3] https://www.descript.com/

(Pustejovsky and Krishnaswamy 2021). They acknowledge all of these modalities by minimal vocalizations whose systematic properties are commonly omitted from computational models of language.

This dissertation argues that non-lexical communicative acts are primary contributions to the common ground (Clark and Brennan 1991; Brennan and Hanna 2009). In addition, I argue that they can be formally represented and characterized, and that by analyzing their systematic properties, we can learn how grounding works and how information flows in multimodal interaction.

1.2 Grounding, Common Ground, and the Multimodal Challenge

The concept of *common ground* has roots in the philosophy of language. Successful communication between participants depends on more than the propositional and semantic content of the spoken utterances. Speakers and hearers rely on shared assumptions about their environment, cooperation, and intent (Grice 1975; Asher and Gillies 2003; Clark and Brennan 1991; Stalnaker 2002). In formalized versions of this notion for HCI and Human-Human Interaction (HHI), the common ground is treated as a set of propositions presupposed by participants (Traum 1995; Markowska et al. n.d.; Krishnaswamy and Pustejovsky 2019).

Other computational approaches to discourse have analyzed task-oriented dialogues in which participants share physical goals. Grosz and Sidner (1986) propose that discourse structure is organized around the interplay of three components: the linguistic structure, the intentional structure, and the attentional structure. The attentional state is a representation of the relevant entities, properties, and relations available to the participants, and is updated as the interaction progresses. It is managed through a stack data structure that tracks which entities are available for reference at a given point in the interaction. This account of shared attention is very relevant to situated dialogue, where what the participants can see and touch is as relevant to their shared context as to what they are talking about.

Formal models of dialogue have evolved these foundations into accounts of how the in-

formation state evolves over the course of an interaction (Traum and Larsson 2003; Farkas and Bruce 2010; Ginzburg 2012). The formal models take different approaches, but share an underlying approach of tracking what is mutually known, what is under discussion, and how communication acts move propositions and information from one epistemic belief state to another.

Most formal accounts do not adequately address the range of modalities through which grounding contributions are made. Many nod to other modalities but most formal methods are speech-centric. Common ground contributions are utterances and the Common Ground Structure (CGS) is filled with propositional utterances.

In situated interaction where participants are co-located, the environment itself contributes information (Hunter, Asher, and Lascarides 2015; Hunter, Asher, and Lascarides 2018; Pustejovsky and Krishnaswamy 2021). When a scale fails to balance, the observable outcome asserts a proposition about the relative weights of the blocks on it. When a furniture assembly step is completed, the event constitutes evidence that the previous instruction was correctly interpreted and executed. In previous work, we demonstrate that physical actions in a multimodal dialogue can serve as antecedents to speech, update object locations in a shared environment, and function as grounding contributions by extending AMR to account for certain actions (Tam et al. 2023a).

This is relayed in a favorite example from my sons, observed at our kitchen table:

(1) B: "Can we play with your fire truck?"
N: *thumbs-up gesture*
J: "N. said 'thumbs up'!"

More broadly, non-verbal events require grounding just as verbal utterances do. The furniture assemblers do not need to say *"the dowel is in"* for the information to enter their common ground. The event itself becomes the proposition, and the grunt acknowledges it.

This dissertation takes the position that a complete account of multimodal communication must treat gesture, observable events, and minimal vocalizations as first-class partic-

ipants in the grounding process. They are not peripheral phenomena, but communicative acts with representable content, functional properties, and a place in the formal setup of dialogue.

1.3 Contributions

This dissertation makes four interconnected contributions toward a theory of non-lexical acts in multimodal interaction.

The first is a framework and annotation scheme for representing the semantic content of gesture. Existing meaning representations, including AMR, were designed to capture the propositional content of written language, and have no extension to non-speech modalities. Along with colleagues, Gesture AMR (GAMR) was developed. It is an extension of AMR that captures the meaning of content-bearing gestures through a taxonomy of gesture act relations (Brutti, Donatelli, et al. 2022). Gesture act relations (deictic, iconic, metaphoric, and emblematic) are embedded in a predicate-argument structure compatible with other existing AMR extensions (Bonial et al. 2020; O'Gorman et al. 2018). This framework is realized in an annotated corpus of multimodal speech and GAMR (Lai, Brutti, et al. 2024), and demonstrates that non-lexical communicative acts have representable propositional content. The theoretical premise of this work is foundational to the rest of the dissertation; it establishes that non-lexical acts can be brought within the scope of formal meaning representation languages, which inspires and licenses the treatment of grunts in future chapters.

The second is a systematic annotation and formal model of grunts as communicative acts and grounding contributions. Drawing on collaborative task interaction from the Weights Task Dataset (Khebour, Brutti, et al. 2024) and the Glider video (Sidner 2025), I present an annotation scheme that captures grunt tokens, their antecedents (speech and observable events), and their dialogue act functions. The distributional analysis reveals that grunts respond to speech and observable events at nearly equal rates, demonstrating that non-verbal events function as conversational contributions on par with utterances (Brutti and

Pustejovsky 2026a). This includes a formalization of how grunts update the common ground using Evidence-Based Dynamic Epistemic Logic (EB-DEL) (Benthem, Fernández-Duque, and Pacuit 2014; Pacuit 2017), treating grunts as operations that promote evidence to accepted fact within the Common Ground Structure.

The third contribution is an acoustic characterization of the functional specialization encoded in grunt form. Through prosodic analysis of nearly 500 annotated grunts (Brutti and Pustejovsky 2026b), I show that token selection is the primary signal of grounding source: backchannels such as *mm-hm* predominantly acknowledge speech contributions, while reactive tokens such as *oh* and *ooh* preferentially follow observed events. Prosody operates as a secondary channel layered on top of this lexical choice. Population-level prosodic differences between speech- and event-following grunts are largely absorbed by token identity, but in additional analysis, where the token is held constant the modulation becomes visible: in standalone *oh* tokens, event-following productions carry substantially higher F0 and a wider pitch range than speech-following ones (Chapter 5). These findings show that the information encoded in a grunt is not only in its token; speakers can systematically modulate acoustics to signal what type of information is being grounded.

The fourth contribution is an analysis of grunt position within turns. Grunts that occur as standalone turns, at the initiation of a speaker turn, in the middle of a turn, or at a turn's conclusion occupy structurally distinct positions that reflect distinct grounding functions. Using a four-way typology (STANDALONE, TURN-INITIAL, TURN-MIDDLE, TURN-FINAL) annotated in ELAN across the Weights Task Dataset, Glider video, and the Portal Dialogue Corpus (Tomlin et al. 2025), I show that position is strongly associated with token identity. Portal's per-speaker audio, which the single-microphone recordings of the other corpora cannot support, further enables prosodic analysis of turn-initial grunts (Chapter 7).

1.4 Organization

The remainder of this dissertation is organized as follows.

Chapter 2 presents Gesture AMR. GAMR and its related work presents the challenge of extending meaning representations to non-speech-based modalities, develops the gesture act taxonomy and formal specification, and presents the Multimodal AMR corpus (Lai, Brutti, et al. 2024) which is built on top of the EGGNOG dataset (Wang et al. 2017). This chapter closes by drawing out the theoretical implication that grounds subsequent chapters, namely that non-lexical communicative acts can carry representable propositional content.

Chapter 3 develops the theory behind the grunt analysis. I present three supporting works to which I contributed that collectively motivate the central claim: the Common Ground Tracking framework and its operationalization of the QBank/EBank/FBank structure (Khebour, Lai, et al. 2024); the Weights Task Dataset (Khebour, Brutti, et al. 2024); and a formal model of information state in situated dialogue that integrates speech and gestural grounding (Lai, Donatelli, et al. 2025). I close with an account of what a complete theory of grunt grounding requires, motivating the remaining chapters.

Chapter 4 presents the core contribution with the annotation scheme for grunts and gruntables, the EB-DEL formal model of grunt-driven common ground updates, and the distributional findings establishing that observable events function as propositional contributions at rates comparable to speech.

Chapter 5 presents the prosodic analysis. I report acoustic findings on token specialization by antecedent type and prosodic variation by dialogue act function, and discuss their implications for theories of grounding and cognitive processing in co-situated interaction.

Chapter 6 recreates the distributional and prosodic analysis on a new data set, the Portal Dialogue Corpus (Tomlin et al. 2025). I repeat the analysis from Chapter 5, and add an additional analysis of lexical entrainment between participants over the pre-annotated Portal corpus.

Chapter 7 presents a supplement to the original grunt annotation scheme, looking at the position of the grunt token within the context of a longer utterance.

Chapter 8 synthesizes the dissertation's contributions, situates them within broader work

on multimodal dialogue and meaning representation, and outlines directions for future research.

2 REPRESENTING NON-LEXICAL ACTS: GESTURE AMR

Meaning representations are a central tool in natural language processing, providing structured accounts of sentence meaning that support human annotation, automatic parsing, and downstream inference. Abstract Meaning Representation (AMR) (Banarescu et al. 2013) is among the most widely adopted. It is a graph-based formalism that expresses the propositional content of a sentence in terms of its predicate-argument structure. An AMR for the English sentence *Put that piece there* is shown in Example 2:

(2) *Put that piece there.*

```
(p / put-01
   :mode imperative
   :ARG0 (y / you)
   :ARG1 (p / piece
             :mod (t / that))
   :ARG2 (t2 / there))
```

This representation captures the basic predicate-argument structure of the spoken command: an implicit addressee (`you`) is instructed to move a specific item (`that piece`) to a specified location (`there`). What it does not capture is the gesture that typically accompanies an utterance like this in situated interaction, likely a pointing gesture that specifies exactly which piece and where it should go. The propositional content of the utterance is incomplete without the gesture; the gesture is incomplete without the utterance. Yet AMR, as designed, provides no mechanism for representing either the semantic content of the gesture or the alignment between the two modalities.

This chapter presents Gesture AMR (GAMR), an extension of AMR that addresses this gap. I introduce the formal specification of GAMR, the gesture act taxonomy, and present the annotated corpus of multimodal speech and GAMR. I demonstrate the theoretical implication that motivates the rest of the work, that non-lexical communicative acts represent

propositional content and that the methods can be extended beyond gesture to other non-lexical modalities.

2.1 Gesture and Semantics

Gesture encompasses the visible bodily actions, particularly of the hands, that speakers produce as part of communicative behavior. Gesture research has established that co-speech gesture is not merely supplemental; it carries semantic content, co-refers to discourse objects, and interacts with speech in systematic ways, across languages and cultures (Kendon 1997; McNeill 2008; Goldin-Meadow 2003; Gawne 2025). Additionally, gesture can convey information that speech cannot, such as the manner of a motion, the shape of an object, and the location of a referent. This information is ampliative to the accompanying utterance (Lascarides and Stone 2009).

The theoretical situation of GAMR within the broader AMR extension landscape was developed in Donatelli et al. (2022a), our companion to (Brutti, Donatelli, et al. 2022) that places the work alongside Dialogue-AMR (Bonial et al. 2020), Spatial-AMR (Bonn et al. 2020), Multi-sentence AMR (O'Gorman et al. 2018), and UMR (Van Gysel et al. 2021). Each of these extensions preserves AMR's core structure while adding tools for their respective use cases. GAMR adopts the same strategy for gesture. The companion paper also develops the connection to the formal gesture semantics tradition (Lascarides and Stone 2009; Schlenker 2018; Lücking 2021) that treats co-speech gesture as contributing propositional content through systematic form-meaning correspondences, providing the theoretical grounding for the gesture act taxonomy below.

Formal representation of gesture poses challenges that are unique from language. The meaning of a gesture is determined holistically by its form, context and relationship to speech and is not compositional from its individual components (McNeill 2008). One gesture can accompany an entire utterance, and gesture components can be described in relation to the producer's body, as well as the nature of the movements. Like the person on a cell

phone, gesturers gesture while communicating, whether or not they are co-situated with their interlocutor.

Prior annotation schemes for gesture have addressed different aspects. Descriptive aspects like hand shape, movement trajectory, location relative to body, and palm orientation are addressed in schemes like (Rohrer et al. 2020; Kopp and Wachsmuth 2010; Kong et al. 2015). Other systems have proposed representations of gesture for conversational agents (Kopp, Krenn, et al. 2006), with an additional interpretation layer. None of these encode gesture meaning in a format that is compatible with speech representations. With schemas like this, gesture and speech can be represented in parallel, but are not unified for easier alignment and reasoning.

Gesture AMR was developed primarily using the EGGNOG corpus (Elicited Giant Gallery of Naturally Occurring Gestures) (Wang et al. 2017), a dataset of paired participants working on a shared task, collected in a university setting. One person, the `(actor)`, has a set of wooden blocks, and the other, the `(signaler)`, has a picture of a specific block arrangement that they are trying to get the `actor` to reproduce. Examples below refer to blocks and block-related predicates; I've continued playing with blocks long after my children moved on from their own.

2.2 Gesture Act Taxonomy

Gesture AMR addresses these challenges by extending AMR's predicate-argument structure to gesture. The central design is that each GAMR is anchored by a *gesture act* (GA) relation, analogous to the speech act relations of Dialogue-AMR (Bonial et al. 2020). The GA identifies the gesture's communicative function and provides the top-level node of the representation. The taxonomy reflects established classifications in gesture literature (Ekman and Friesen 1969; Mather 2005; McNeill 1992) and distinguishes four referential forms of content-bearing gesture: deictic, emblematic, iconic, and metaphoric. Gesture AMR does not physically describe the physical gesture shape, the movement or speed of the gesture.

Additionally, Gesture AMR does not consider non-referential beat or rhythmic gestures.

The general form of a Gesture AMR is as follows:

(3)

```
(g / [gesture]-GA
   :ARG0 [gesturer]
   :ARG1 [content]
   :ARG2 [addressee])
```

`ARG0` is the gesturer, `ARG2` is the addressee, and `ARG1` contains the semantic content of the gesture, which varies by gesture type as described below.

2.2.1 Deictic Gesture *(deixis-GA)*

A deictic gesture refers to an object or location by pointing. The semantic content of a deictic gesture is the pointed-to object or location, represented similarly to standard AMR. An indicating gesture with one finger is treated semantically the same to a gesture that might use the whole hand. A pointing gesture to a block on a scale is annotated as follows:

(4)

```
(d / deixis-GA
   :ARG0 (g / gesturer)
   :ARG1 (b / block)
   :ARG2 (a / addressee))
```

2.2.2 Emblematic Gesture *(emblem-GA)*

An emblematic gesture has a meaning that is set by convention rather than by any physical or metaphorical similarity between the gesture and its semantic content. In the US English context, a thumbs up and a head nod both express similar semantic meaning, so their gestures would be equivalent in GAMR in most cases. A head nod gesture, commonly used in English-speaking contexts to express positive acknowledgment, is annotated as:

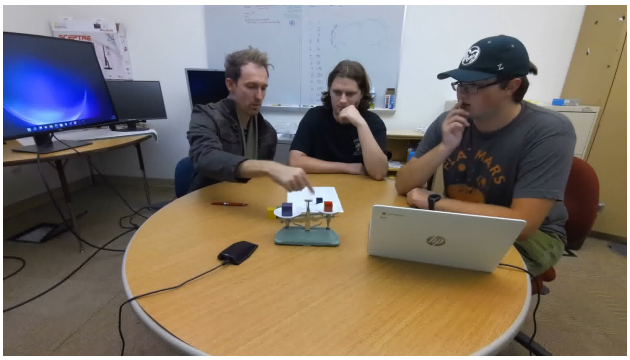


Figure 2.1: Deictic pointing gesture: leftmost participant points to a block on the scale.

(5)

```
(e / emblem-GA
   :ARG0 (g / gesturer)
   :ARG1 (y / yes)
   :ARG2 (a / addressee))
```



(a) Head up (b) Head down

Figure 2.2: Emblematic nod gesture: leftmost participant moves head down and up to signal affirmation to previous speech.

2.2.3 Iconic Gesture *(icon-GA)*

An iconic gesture depicts a concrete property of an object of action, like its shape, size, manner of motion, or spatial relationship. Action predicates are drawn from PropBank (Palmer, Gildea, and Kingsbury 2005), as in standard AMR. A two handed gesture mimicking the balance scale is annotated as:

(6)

```
(i / icon-GA
   :ARG0 (g / gesturer)
   :ARG1 (s / scale)
   :ARG2 (a / addressee))
```

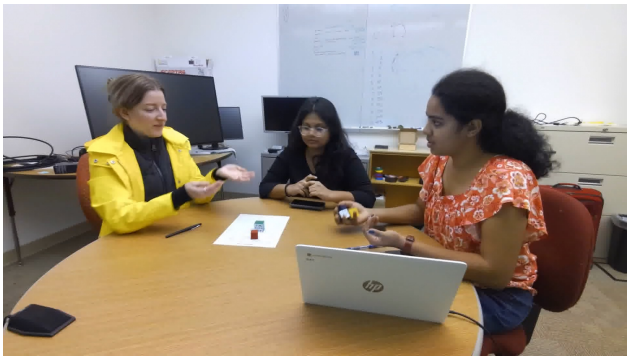


Figure 2.3: Iconic scale gesture: leftmost participant mimics a balance scale rightmost participant attempts to weigh blocks in their hands.

2.2.4 Metaphoric Gesture *(metaphor-GA)*

A metaphoric gesture depicts an abstract property of a concept or idea, contrasting with an iconic gesture. In our data, we observed a participant using both arms in an arc, encouraging the actor to arrange the blocks like a "sound wave". This is fittingly metaphoric, because as the prosody chapter will show, actual sound waves don't look like this.

(7)

```
(m / metaphor-GA
   :ARG0 (g / gesturer)
   :ARG1 (s / sound wave)
   :ARG2 (a / addressee))
```



(a) "Sound wave": hands closed



(b) Hands moving apart



(c) Hands apart

Figure 2.4: EGGNOG participant explaining to arrange blocks in a shape like a "sound wave".

2.2.5 Gesture Units and Coordinated Gestures

A single gesture can have multiple meaning components (Kendon 2004). When a gesturer simultaneously outlines a square with their hands (depicting a block) and moves them downward (depicting placement), both components are annotated and combined under a `g / gesture-unit` node.

(8)

```
(g / gesture-unit
   :op1 (i / icon-GA
            :ARG0 (g2 / gesturer)
            :ARG1 (b / block)
            :ARG2 (a / addressee))
   :op2 (d/ deixis-GA
             :ARG0 g2
             :ARG1 (l / location)
```

```
            :ARG2 a))
```

Coordinated gestures (independent gestures that occur simultaneously) should be connected using `(a / and)`. For example, if a signaler makes a pointing gesture with each hand to two different locations, meaning "put a block here and put another block there", we consider this as two gestures made at the same time.

(9)

```
(a / and
  :op1 (g / gesture-unit
         :op1 (d / deixis-GA
                :ARG0 (s / signaler)
                :ARG1 (l / location)
                :ARG2 (a2 / actor))
         :op2 (i / icon-GA
                :ARG0 s
                :ARG1 (p / put-01)
                :ARG2 a2))
  :op2 (g2 / gesture-unit
         :op1 (d2 / deixis-GA
                :ARG0 s
                :ARG1 (l2 / location)
                :ARG2 a2)
         :op2 (i2 / icon-GA
                :ARG0 s
                :ARG1 (p2 / put-01)
                :ARG2 a2)))
```

2.3 Alignment in GAMR

In multimodal communication, the same objects and events are frequently referenced in both speech and gesture. Capturing the full meaning of a communicative act requires not only representing the content of each modality separately but combining them into a single unified representation that supports cross-modal coreference and joint reasoning.

Consider a speaker who says *Put that block there* while pointing first to a specific block and then to a specific location. The full communicative act is represented as a single structure in which both gesture and speech are arguments of a shared communicative act node:

(10)

```
(s1c / command-00
  :ARG0 (g / gesturer)
  :ARG1 (c2 / communicative-act
    :gesture (d / deixis-GA
      :ARG0 g
      :ARG1 (b / block
                :ARG1-of (b2 / blue-01))
      :ARG2 (a / addressee))
    :gesture (d2 / deixis-GA
      :ARG0 g
      :ARG1 (b3 / block
                :ARG1-of (p / purple-02))
      :ARG2 a)
    :speech (p2 / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b4 / block
```

```
            :mod (t / that))
   :ARG2 (t2 / there))
:ARG2 a))
```

The `:gesture` and `:speech` roles are arguments of the same communicative act, placing the two modalities on equal footing. Coreference relations then mark that *that block* in the speech and the blue block pointed to in the first deictic gesture refer to the same entity, and that *there* in the speech and the purple block pointed to in the second gesture refer to the same location.

In the GAMR formalization, we encode the temporal relations between speech and gesture. This approach adopts Interval Temporal Logic (ITL) (Allen 1983) as used in TimeML (Pustejovsky, Lee, et al. 2010), but reduces the 13 relations to a coarser *overlap* (covering ITL's overlap and during, and their inverses). This builds on UMR (Van Gysel et al. 2021), whose temporal markup derives from the dependencies in Y. Zhang and Xue (2018). The formalism uses an AMR-native reentrancy to bind coreferent entities across modalities. For example, the block being pointed to and "that block" are the same entity, as are the gesture's addressee and "you". ", an implicit argument of the speech. We mark that gesture `d` temporally overlaps with "that", `d2` overlaps with "there", and that `d` precedes `d2`. This approach was formalized in the original GAMR paper, but was not operationalized for the associated annotations.

GAMR is distinguished from prior annotation schemes that treat modalities separately: by embedding speech and gesture as co-arguments of a single communicative act, GAMR enables reasoning over their joint content rather than over each modality in isolation. The co-argument alignment structure was described in more detail in Donatelli et al. (2022b).

2.4 Encoding Gesture in Multimodal Dialogue: The Multimodal AMR Corpus

2.4.1 AMR and GAMR

The GAMR annotation scheme was carried out on the EGGNOG video data, described above (Wang et al. 2017). The videos are approximately 1 minute each, and have natural continuous communication between the actor and signaler as they are in different rooms by connected by audio and video. The task-based nature of EGGNOG provides us with a simple concrete task and limits the possible range of gestures from the participants, making it very useful to develop the schema. However, it does not reflect a wide range of human communication, and participants are limited to an English-speaking U.S. university setting. The whole dataset contains over 300 videos, with annotations also including a physical description of the motion. We selected 21 videos for annotation with GAMR (Lai, Brutti, et al. 2024).

Annotation was first conducted in ELAN (Brugman and Russel 2004), with separate tiers for speech AMRs and gesture AMRs. A team of five student annotators was trained first on standard AMR, then on the GAMR schema. Each video was dually annotated and adjudicated by expert annotators to produce a gold standard. We also provided annotators with ASR transcripts as an ELAN tier. The annotators were responsible for segmenting the speech for AMR annotation.

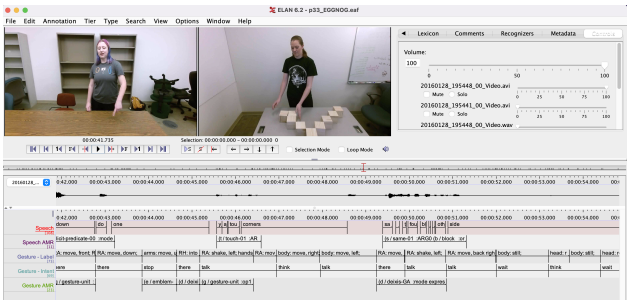


Figure 2.5: Annotating EGGNOG in ELAN.

The resulting corpus contains 662 AMRs: 343 speech AMRs and 319 gesture AMRs. Among gesture AMRs, iconic gestures are the most frequent (142 top-level instances), fol-

Gesture Act	Top-level	All
Icon	142	249
Deixis	63	129
Emblem	37	39
Gesture unit	50	-
Coordinated gestures	27	-

Table 2.1: Distribution of gesture act types in EGGNOG annotations.

lowed by deictic gestures (63) and emblematic gestures (37). The most common iconic gesture contents are *put-01* (40 instances), *block* (31), and *slide-01* (17); the vast majority of deictic gestures denote locations (99 instances).

Inter-annotator agreement was measured using SMATCH (Cai and Knight 2013) and the soft-matching variant S^2MATCH (Opitz 2023). SMATCH scores (48.9 for speech, 57.5 for gesture) are substantially lower than those for text-based AMR (typically 70–80), reflecting the genuine difficulty of the task: annotators segmented continuous speech into utterances, determined which body movements constitute content-bearing gestures, and resolved cross-modal reference without the surface form anchors available in text annotation. S^2MATCH scores (64.8 for speech, 71.5 for gesture) are considerably higher, reflecting that disagreements are often over near-synonymous lexical choices rather than structural differences. S^2MATCH accepts synonyms rather than relying on exact matches.

2.4.2 Multimodal Multi-sentence AMR

Our contribution of *Multimodal Multi-sentence AMR* (MM-MS-AMR) involved annotators marking coreference and bridging relations (e.g., `set-member` and `part-whole`) across the gold-standard AMRs and GAMRs using the Multi-sentence AMR MS-AMR framework (O'Gorman et al. 2018). MS-AMR links standalone AMRs using these relationships, resulting in a representation of meaning across a document or discourse.

Annotators used the Anafora tool (W.-T. Chen and Styler 2013), a web-based annotation tool designed for co-reference annotation. The multimodal multi-sentence portion of the corpus contains 436 cross-modal relations: 388 coreference chains, 28 set-member relations,

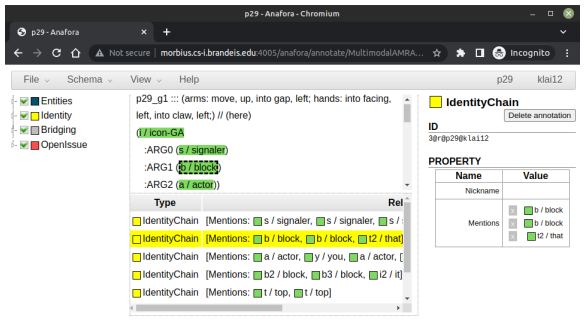


Figure 2.6: Anafora annotation tool.

and 20 part-whole relations, covering approximately half of all entities in the corpus including implicit roles.

Cross-modal coreference agreement (CoNLL-2012 F1 of 60.46) is below text-based MS-AMR results (69.86) but not substantially so, indicating that the framework is learnable and its annotation guidelines sufficiently well-specified. It is a difficult annotation task, requiring close reading of the AMRs.

2.5 Representable Content of Non-Lexical Acts

The GAMR framework and corpus establish a result that is foundational for the remainder of this dissertation: non-lexical communicative acts have representable propositional content. A pointing gesture, an iconic depiction of an action, a thumbs-up can all be assigned a formal semantic representation in a predicate-argument structure. These structures are compatible with representations of the speech they accompany, they support cross-modal coreference, and enable joint reasoning over the full content of a multimodal communicative act.

This result licenses the approach taken for grunts in subsequent chapters. The AMR framework is flexible enough to be extended to accommodate the semantic content of non-lexical communication. The grunt analysis will be more of a characterization of what grunts represent and how they function in dialogue, not just that they *can* be represented at all.

While AMR and its extensions are well-suited to representing the propositional content of communicative acts, grunts are a heterogeneous class. Some are turn starters and hesitation markers that do not contribute propositional content (Clark and Fox Tree 2002). Others

are about *epistemic states*: whether information has been received, whether evidence has been accepted, or whether a proposition has been promoted to accepted fact in the common ground. This is the domain of dynamic epistemic logic rather than predicate-argument representation, and it is why the grunt analysis in Chapter 4 employs Evidence-Based Dynamic Epistemic Logic (EB-DEL) (Benthem, Fernández-Duque, and Pacuit 2014; Pacuit 2017) rather than developing a *Grunt AMR* extension. The shift from AMR to EB-DEL is a theoretically motivated choice: AMR captures semantic content, EB-DEL captures state updates, and the grunts examined here are primarily the latter. We discuss the shift in Y. Zhu et al. (2024), where evidence-based dynamic epistemic logic was introduced as the appropriate formal apparatus for tracking belief updates across modalities in the Weights Task. Chapter 3 takes up that framework in detail.

The theoretical claim that non-lexical acts are first-class grounding contributions requires grounding itself. Existing corpora of multimodal dialogue were insufficient for this purpose; there are many strong datasets that do not quite check all of the boxes for this work. The Weights Task Dataset (WTD) was developed as a co-situated, shared collaborative task designed to study collaborative problem solving (Khebour, Brutti, et al. 2024). The WTD is the primary empirical resource for this dissertation. The grunt annotation (Chapter 4), prosodic analysis (Chapter 5), and turn-position annotation (Chapter 7) are all conducted on WTD data, with the flat pack furniture building Glider video as a secondary corpus before the addition of Portal in a later chapter.

The kind of interaction discussed here, where speech occurs with a background and environment of shared physical activity, has been recognized in other traditions. The notion of *empractical speech* (Bühler 1934) describes speech embedded in other non-linguistic activity, able to be interpreted against a shared situation. The concept has been revisited more recently by O'Connell and Kowal (2012).

This chapter builds on three works I co-authored. I describe the Weights Task Dataset in more detail, as it is a primary resource for the grunt schema development. I then present work on annotating physical actions as discourse contributions, establishing that observable events belong in the same formal account as speech and gesture. Next, I present the Common Ground Tracking framework that operationalizes multimodal common ground formally and computationally. Finally, I present a model of information state that integrates speech, gesture, and action into a unified theoretical account.

3.1 The Weights Task Dataset

To adequately study grounding in multimodal situated dialogue, we needed data that no existing resource provided. Existing corpora of multimodal interaction lacked at least one of the following: co-present participants with shared perceptual access to physical objects,



Figure 3.1: The Weights Task Group 3, with captions

physical object manipulation as a central task component, or multiparty interaction (Khebour, Brutti, et al. 2024). The Weights Task Dataset (WTD) was designed to address this gap, within the context of the NSF AI Institute for Student-AI Teaming (iSAT), specifically with collaborative problem solving (CPS) in mind (Stewart, Keirn, and D'Mello 2021).

3.1.1 Task Design and Motivation

The Weights Task is a collaborative problem-solving activity completed by triads at a round table. Participants are given a balance scale and five blocks of varying weight, size, and color. The facilitator tells the group the weight of one reference block (10 grams), and the group must determine the weights of the remaining blocks through collaborative experimentation. Once the individual weights are established, participants must identify the relation between them and predict the weight of the next hypothetical unseen block in the series, which turns out to be the Fibonacci sequence.

The task was designed to elicit genuine multimodal communication. Participants cannot complete it through any single modality alone: they must speak to discuss weights, gesture to indicate blocks, and physically manipulate the blocks and scale to test hypotheses. The structured nature of the task provides a consistent ground truth against which annotation and modeling can be evaluated: all groups successfully determine the correct weights, giving a stable endpoint for assessing common ground construction. The dataset consists of 10 videos totaling approximately 170 minutes, with 30 unique participants recruited from a university setting, plus one researcher-facilitator consistent in all videos.

3.1.2 Annotation Layers

The WTD is richly annotated across multiple modalities. Speech was transcribed by human annotators and multiple ASRs. Gestures were annotated with GAMR. Collaborative problem solving turns were annotated with Sun et al. (2020), labeling dimensions of group collaboration. Nonverbal behaviors including gaze, posture, and object usage were captured using the NICE coding scheme (Dey et al. 2023). Microsoft Azure Kinect cameras captured RGBD video from multiple angles, enabling extraction of 32-joint skeleton data for each participant.

3.2 Annotating Situated Actions in Dialogue

With the WTD in hand, the first theoretical question was how to extend formal meaning representation to physical actions. In Tam et al. (2023a), we address this question directly, arguing that actions are critical for interpreting dialogue and must be integrated with language and other communicative modalities in a shared annotation scheme.

3.2.1 Actions as Discourse Contributions

Many previous works have addressed the importance of non-linguistic modalities in communication, especially in the context of dialogue systems and agents (Cassell et al. 2000; Wahlster 2006; Foster 2007). Actions commonly serve as antecedents in speech (e.g., in my house one brother frequently asks another "why did you do that?"). Actions create relations between narrative events, and can themselves be referenced directly by participants. Most importantly for my purposes, actions update the common ground in ways that speech and gesture cannot, by adding, modifying, and deleting items within it.

3.2.2 Annotating Actions with AMR

Our work in Tam et al. (2023a) demonstrates the feasibility of annotating a limited set of physical actions, represented as Action AMR, on two datasets; the WTD and Epic Kitchens

(Damen et al. 2022). As we have seen, in the WTD the actions are primarily manipulating blocks and the scale, picking up blocks, and moving them to new positions. The movements are often prompted by speech or serve as antecedents to speech. The action of placing a block on the scale is represented as:

(11)

```
(p / put-01
   :ARG0 (p1 / participant)
   :ARG1 (b / block)
   :ARG2 (s / scale))
```

A parallel demonstration on Epic Kitchens shows the same approach generalizes to cooking actions such as transferring pasta from a pot to a bowl. This Action AMR is for a first-person action, where the home chef moves the cooked pasta from the pot of boiling water to a bowl for eating, using a utensil.

(12)

```
(t / transfer-01
   :ARG0 (p / participant)
   :ARG1 (p1 / pasta)
   :ARG2 (p2 / pot)
   :ARG3 (b / bowl)
   :instrument (s / spaghetti claw))
```

3.2.3 Action Annotation Challenges

Annotating actions poses difficulties that do not arise for speech and less so for gesture. Determining the appropriate level of granularity is a challenge for inter-annotator agreement, as well as automatic systems. Is picking up a block and putting it on the scale one action? Actions can happen simultaneously without the sequential structure of speech. Both

make consistent inter-annotator agreement, and reliable automatic segmentation, difficult to achieve.

Physical actions are discourse-relevant contributions: they serve as antecedents, update object locations, and modify common ground independent of speech. Hunter, Asher, and Lascarides (2018) formalize this, treating nonlinguistic events as discourse units that enter into coherence relations alongside linguistic ones. The annotation challenges of this work are the concrete version of the theoretical ones in the Tam et al. (2023a) paper.

3.3 Common Ground Tracking in Multimodal Dialogue

The WTD and subsequent Action AMR work helped set the conditions for a formal framework and computational pipeline for tracking multimodal common ground over the course of a collaborative interaction. In Khebour, Lai, et al. (2024), we present Common Ground Tracking (CGT) on the WTD dataset.

3.3.1 Common Ground Structure

The CGT framework is distinct from Dialogue State Tracking (DST), which represents the speaker's state at each point in a dialogue by considering the past conversational turns and dialogue moves. CGT aims to identify the shared beliefs of all participants in a task-oriented interaction. DST tracks what a user may want from an interactive computer system and CGT tracks the knowledge of all participants in a group. Other recent work has addressed challenges of tracking conversational grounding within text and chat data (Mohapatra et al. 2024), however text-based approaches do not have these grunts unless added performatively by users.

CGT adopts a modified version of the Dialogue Game Board (Ginzburg 2012), combined with an evidence-based model for belief from Benthem, Fernández-Duque, and Pacuit (2014) and Pacuit (2017). The grounding phenomena modeled here do not necessarily depend on this formalism. The ideas could be expressed in a related structure, such as the Table model

by Farkas and Bruce (2010). The Common Ground Structure (CGS) is a triple, consisting of three "banks", with propositional content moving from one bank to another:

- QBank: the set of questions under discussion. Following the QUD tradition (Ginzburg 2012), a question denotes a set of propositions, which are its possible answers. The questions in QBank are task-relevant and have not yet been resolved. *Is this the right dowel?*

- EBank: the set of evidenced propositions. These are claims for which some evidence of truth exists but which have not yet been accepted by all participants in the group. *The dowel slid in smoothly, suggesting that it fits.*

- FBank: the set of facts. These facts are propositions accepted as true by all participants. *The dowel fits (and the assembly can proceed).*

In Khebour, Lai, et al. (2024), the questions in QBank are yes/no questions, generated from the relations the task implicates for each object in the domain. In the Weights Task, where the goal is to determine the weight of each of five blocks, initialization enumerates a question for every candidate weight of every block. Because these are yes/no questions, each question $\varphi?$ denotes $\{\varphi, \neg\varphi\}$ and pairs with a unique proposition φ. This is what allows content to be tracked across banks of different types: QBank holds questions, while EBank and FBank hold propositions.

Dialogue moves update this structure. A statement move resolves a question $\varphi?$ in QBank and places its answer φ in EBank as evidence. An accept move promotes φ from EBank to FBank. A doubt move reverses the first step: φ leaves EBank and the question $\varphi?$ returns to QBank as unresolved.

These communicative acts of speech, gesture, and action are specified with Evidence-Based Dynamic Epistemic Logic (EB-DEL). The announcement operator $[!\varphi]$ represents a public announcement of φ, and the model distinguishes the evidence an announcement provides from the belief it establishes. An announcement made without prior evidence

updates the evidence sets of all participants to the worlds where φ holds: $[!\varphi][E]\varphi$. This is the STATEMENT move, and it is what places φ in EBANK.

Belief is not established by that first announcement. It requires a second one, made in a context where evidence for φ already exists: $[E]\varphi \to [!\varphi][B]\varphi$. The precondition $[E]\varphi$ is what the first announcement supplied; the re-announcement discharges it, and φ is promoted from EBANK to FBANK as a fact believed by all participants. This is the ACCEPT move. It is the apparatus that Chapter 4 extends to grunts, where the claim developed is that this second announcement, the step from evidence to fact, is frequently carried by a grunt rather than by speech.

We developed the theoretical motivation for this structure in Y. Zhu et al. (2024), which frames common ground tracking as an application of Simulation Theory of Mind (Goldman 2006). Consider a brief example to foreshadow the formal setup. A 10 gram weight sits on one side of a balance scale. A person puts an unknown block on the other side, creating a proposal that the new block weighs 10 grams too. This proposal is now available for acceptance. When the scale balances, there is evidence that the second block weighs 10 grams, supporting the initial proposal. That evidence becomes a fact when a participant confirms it, with a grunt like *oh*. The proposition moves to FBANK, and the group proceeds.

Each weighing works this way, resolving one question at a time and narrowing the set of live candidates until a single weight remains for each block. The task is a sequence of such eliminations, and the structure of the CGS records them.

Understanding another agent's mental state requires tracking what they perceive, what evidence is available to them, and what they believe. The EB-DEL statements make this concrete. Perception provides evidence, non-contradictory evidence supports belief, and public announcement restricts the evidence sets of all participants to worlds where the announced proposition holds. Y. Zhu et al. (2024) work through a Weights Task example in which participants balance two blocks on the scale, observe the outcome, and through announcement and agreement arrive at the group belief that the blocks weigh the same. Our model has a

common ground data structure for all participants; alternate formations could be possible where an individual's facts and beliefs are tracked separately.

3.3.2 Annotation and Experimental Pipeline

The CGT system uses a multimodal LSTM move classifier that draws on features from speech, prosody, gesture, action annotations, and collaborative problem-solving facets to predict the dialogue move expressed in each utterance, which is then translated into updates to the QBank, EBank, and FBank.

The CGT system operates through three components. A move classifier, implemented as a multimodal LSTM, predicts which dialogue move is expressed in each utterance: STATEMENT, ACCEPT, DOUBT, QUESTION, ANSWER, or OBSERVATION. The classifier takes input features derived from speech (BERT embeddings) (Devlin et al. 2019), prosody (openSMILE features) (Eyben, Wöllmer, and Schuller 2010), gesture (GAMR), action annotations, and collaborative problem-solving facets. A propositional extractor retrieves the task-relevant content of each utterance, either from common ground annotations or through language-model-based dense paraphrasing (Tu et al. 2023). A set of closure rules then translates the move classification and propositional content into updates to the QBank, EBank, and FBank.

The key empirical finding is that incorporating the multimodal features of gesture, action, and prosody improves common ground tracking over language alone, particularly for groups that rely heavily on non-verbal communication to complete the task.

The CGT framework in this paper provides the formal structure and methodology for tracking the multimodal common ground. However, the move classifier operates on utterances, gestures, and actions. It has no account for how minimal vocalizations contribute to grounding moves, especially if they are not registered by ASR. Grunts are the missing signal in many cases. Observationally, the majority of grunts in the WTD are not part of the provided ASR transcripts.

3.4 A Model of Information State in Situated Multimodal Dialogue

The CGT framework operationalizes common ground tracking empirically. In Lai, Donatelli, et al. (2025), we step back to ask what a complete formal model of information state in situated multimodal dialogue should look like, drawing on the WTD and EGGNOG corpora to ground the analysis in concrete examples.

3.4.1 Integrating Speech, Gesture, and Action

The model is illustrated through a worked example in which an EGGNOG signaler says:

(13) take another block put it next to it

while making a deictic gesture indicating a specific location. The AMR of the speech introduces two discourse referents - the new block b (`another block`) and a location l (`next to it`). It also adds *take* and *put* to the actor's To-Do List (Portner 2004). The deictic gesture specifies l as the location adjacent to a previously placed block, which is more specific than the *next to it* from the speech.

The signaler's communicative act consisting of speech plus gesture functions as a STATEMENT: it introduces the propositional content that a block should occupy location l. There is also a grounding function g that maps the discourse references to the actual elements in the environment. While our model in Lai, Donatelli, et al. (2025) does not formally adopt the CGS, we can put this act in those terms. The signaler's suggestion moves the corresponding question from QBank to EBank.



(a) Actor (b) Signaler

Figure 3.2: Actor giving instructions to Signaler in an EGGNOG video.

The actor picks up a block and places it at the specified location, proposing a grounding that the block she selected belongs in the location she placed it. The action is interpretable as a suggestion, or OBSERVATION, in EBank. They are resolved when the signaler accepts or rejects the actor's proposed interpretation. An ACCEPT move, either explicit such as *yes, right there* or as a minimal vocalization like *mm-hm*, promotes the grounded proposition from EBank to FBank (and clears the items from the actor's To-Do List). A DOUBT move, such as a head shake gesture or explicitly spoken phrase (*no, the other one*) would return the proposition to QBank for further negotiation. The bidirectionality is the key theoretical point of this paper. The physical action is a proposal, and requires acknowledgment to update the grounding function. Recent experimental corpus work demonstrates that listener feedback shapes the discourse using a storytelling premise (Boudin et al. 2024). Listener feedback is not incidental. The minimal grunt vocalizations that commonly perform that acknowledgment are a primary vehicle through which it occurs in task-based dialogue, and will be discussed in the remaining chapters.

The information state model provides the most complete account of how speech, gesture, and action work together to update the common ground in situated dialogue. A grunt following a successful action is not just a continuer; it is an acceptance move that updates the grounding function. The model describes the mechanism, and the grunt analysis aims to characterize the vocalizations that perform it.

3.5 From Gesture to Grunt

The three papers presented in this chapter collectively establish the following. Physical actions and observable events are discourse contributions that update the common ground independently of speech (Tam et al. 2023a). Multimodal common ground can be formally tracked through a QBank/EBank/FBank structure updated by dialogue moves, and incorporating gesture, action, and prosody into the tracking pipeline improves performance over language alone (Khebour, Lai, et al. 2024; Y. Zhu et al. 2024). Speech, gesture, and action

jointly construct and update an information state in situated dialogue, with a grounding function that maps discourse referents to objects in the shared physical environment, and whose updates require acknowledgment to become part of the common ground (Lai, Donatelli, et al. 2025).

This framework is still missing an account of the grounding grunts discussed in Chapter 4. The ASR systems used in the Weights Task Corpus, Google and Whisper, both filter or suppress minimal tokens (Whisper confirms this in (Radford et al. 2023); the technical situation is less clear with Google ASR, but observationally it appears that many are omitted). This means that the CGT tracker does not have access to grunts during the pipeline.

As I will show, grunts are not peripheral to the grounding process. They are actually among its primary signals in the right situations. In the WTD and Glider videos, grunts respond to speech and observed events at nearly equal rates, showing that they are first class grounding contributions. The grunt token selection and prosodic form encode grounding in systematic ways, and their position within turns often reflects their role. The following chapters take up this characterization in detail, beginning with the annotation scheme and formal model that establish what grunts do, and proceeding through the acoustic and structural evidence for how they do it.

The CGT framework established the formal model for tracking multimodal common ground, but has no account of the minimal vocalizations that perform the grounding moves it describes. The framework was designed to be extensible, so additional modalities can continue to be added. This chapter presents the core empirical and theoretical contribution of the dissertation: a systematic annotation scheme for grunts and *gruntables*, a formal model of how grunts update the common ground, and distributional findings from the corpus demonstrating that observable events function as grounding contributions at rates comparable to speech.

4.1 Corpora

4.1.1 Weights Task Dataset

For the grunt annotation, I used a mixed corpus of approximately 80% WTD (2 hours and 20 minutes), which was already introduced in Chapter 3, and the Glider video (40 minutes) for the remaining 20%. Nine of the ten Weights Task videos were suitable for grunt annotation. One was omitted due to quiet speech and poor audio quality. This was generally a challenge with the WTD, as all participant and researcher audio was captured with a single microphone and is not channel-separated.

4.1.2 Glider

The Glider video was recorded by Candace Sidner in 1989 and is previously unpublished (Sidner 2025). It was recorded on a camcorder in a home setting and was later digitized. The interaction is natural and collaborative, and the video is accompanied by a detailed hand transcription produced by the MIT AI Lab (Agre and Batali 1989). The Glider video was chosen for its similarity to the WTD videos: it is a dyad working on the shared task of assembling a piece of flat-pack furniture. The two participants use tools, read instructions, and collaborate on the assembly. In addition to the collaborative discussion, there is a thread



Figure 4.1: Glider furniture assembly

throughout about whether or not they are missing some hardware from the furniture kit. Their belief about this question changes over the course of the interaction. Combined with the nine WTD videos, the corpus totals approximately three hours of annotated video across 30 unique speakers.

4.2 Annotation scheme

4.2.1 What is a Grunt?

The term *grunt* and its relationship to adjacent terms of art in the literature is discussed in Chapter 1. Briefly, this work follows Ward (2006) in treating grunts as intentional non-lexical vocalizations that occur within the stream of speech and exhibit systematic sound-meaning correspondences. Semi-lexicalized forms such as *okay* and *yeah* are excluded, as are sighs, laughs, and clicks.

The goal of the annotation scheme is to analyze the functional role of non-lexical vocalizations in conversation and to understand their distribution across different dialogue contexts (Brutti and Pustejovsky 2026a). Full annotation conventions are available at this link[1]. To this end, annotators were asked to identify and label conversational grunts as well as the *gruntable*, which is the antecedent speech or event that the grunt responded to. The gruntable concept is inspired by the *laughable* in laughter research (Ginzburg, Mazzocconi,

[1]https://github.com/richardbrutti/grunts-corpus

and Tian 2020; Mazzocconi, Tian, and Ginzburg 2020), adapted here for the grunt context. In their framework, the laughable is selected by laughter as an argument. As with laughables, gruntables can be exophoric, pointing to something outside the relevant conversation. Similarly, the gruntable selects speech or an event as the argument of the conversational grunt. This is similar to the property that Wilkins (1992) attributes to interjections, that, notwithstanding the difference in terminology, interjections are anchored to the speech they refer to.

Annotators were given pre-labeled tiers in ELAN containing ASR transcripts for each participant as per the WTD corpus. These tiers were used as reference for identifying where non-lexical vocalizations occur in the interaction, but annotators were asked not to rely on the ASR transcription alone, as grunts are frequently omitted or misrecognized by ASR systems, as previously discussed.

4.2.2 Grunt Annotations

Each grunt annotation captures the following information, separated by commas:

Speaker Who produced the grunt (`researcher`, `p1`, `p2`, `p3`)

Transcription A phonetic transcription of the grunt (e.g., *oh*, *mm*, *uh*)

Dialogue Act The communicative function of the grunt, drawn from the taxonomy described below

Standalone Whether the grunt constitutes a complete conversational turn (`yes`/`no`); omitted for FILLED PAUSE

A complete annotation takes the form: `p1, um, STATEMENT, yes, +, no`.

Additional experimental labels were used for some of the videos during schema development, but proved too difficult to annotate consistently and effectively: *Expectedness* Whether the gruntable was expected in context (yes/no); *Polarity* Whether the grunt affirms or agrees

with the gruntable. Both could be useful for gathering more information about coordination around expectations and expectation violation, and how they are expressed (Stone and Stojnic 2015).

The dialogue act taxonomy is adopted from Khebour, Lai, et al. (2024), which in turn adapts the Dialogue Game Board (Ginzburg 2012):

STATEMENT A factual assertion about the world or the task procedure. Observable events are interpreted as STATEMENTs if they provide new information.

ACCEPT Agreement with a previous STATEMENT or OBSERVATION. A move by which a participant takes up a preceding statement or observation as available to the common ground.

DOUBT Disagreement or uncertainty regarding a previous STATEMENT.

QUESTION A yes/no or wh-question seeking information.

ANSWER A response to a QUESTION.

OBSERVATION An observation registers something noticed that does not advance the task, including off-task utterances.

INFERENCE A deduction bridging OBSERVATION and STATEMENT or DOUBT. It is rare for a grunt to fill this role.

Filled pause Filled pauses, turn starters, and disfluencies. Filled pauses do not require a preceding gruntable.

OTHER Any dialogue act that does not fit the categories above.

Some clarifications about ACCEPT are worth making explicit, since the category carries most of the analytic weight in future chapters. In practice, the majority of identified grunts were labeled ACCEPT or FILLED PAUSE.

Acknowledgment and agreement are ordinarily treated as distinct: registering that a contribution has been received is not the same as endorsing its content. This single ACCEPT category does not separate them. Acknowledging something without disputing it is a way of agreeing with it. The moment of acknowledgment is the moment where an objection or doubt could occur, so the acknowledged proposition stands and the group proceeds on it. This isn't just convenient: the taxonomy already has a category for the other case. A participant who doesn't accept a contribution can DOUBT it, and disagreement is coded there. ACCEPT covers everything else.

This is limited because the tacit agreement can be taken back. The annotation records the move made at the time; annotators do not know what participants believe but do not say.

4.2.3 Gruntable

The gruntable concept emerged from careful observation of the WTD videos. Participants commonly use minimal vocalizations, sometimes with marked prosodic modulations, to react to results on the scale. The more I watched, the more I observed that participants made these sounds in less surprising contexts too. The phenomenon extends well beyond experimental tasks. Think of an announcer responding to a slam dunk, children building a block tower together without explicit verbal coordination, or someone responding to a first taste of a new dish. None of these minimal responses are produced in isolation; they are responsive acts, and what they respond to is the gruntable.

The timestamps for gruntables are less sensitive than for grunts, as identifying the precise onset and offset of an event is inherently difficult, as discussed in Chapter 3. A complete annotation takes the form: `p2, speech, that one is 50, STATEMENT`. Each gruntable annotation captures the following:

Speaker Who produced the speech or initiated the event

Type Whether the gruntable is: `speech`, `event`, speech describing an event (`DE`), speech

describing a gesture (`DG`), or `other`

Content A transcription of the gruntable utterance or a brief description of the event (e.g., `scale does not balance`)

Dialogue Act The same taxonomy as for grunt annotations

The gruntable makes grunts responsive by definition, and not every grunt is. A speaker who says *Oh, I just realized that today is Tuesday* is not responding to anything in the shared environment. The *oh* projects forward, rather than grounding anything that has already happened.

This case falls outside the scheme. A gruntable has to be publicly available to both participants, either as speech or as an observable event, and that is what makes grounding analyzable. A projective *oh* selects a private realization that the addressee has no access to. There is nothing to ground. There is no evidence for the proposition in EBANK and nothing for a grunt to promote.

These kinds of grunts do not arise much in these corpora, given their settings. Co-situated task dialogue supplies observable antecedents continuously, and there are not many occasions to announce a private realization.

Where such cases do occur, they are absorbed by FILLED PAUSE, which covers turn starters as well as disfluencies. Most FILLED PAUSEs are planning hesitations, but not all of them are. These projective *oh*s or *oh* inside a phrase like *oh well* are here too. This category functions as a catchall. It collects grunts with no gruntable. Whether this scheme would extend to areas where projective grunts are common, such as casual conversation or storytelling, is an open question.

4.2.4 Annotation Procedure

All nine WTD videos were annotated by me. Four were independently annotated by graduate students in the Brandeis Natural Language Annotation for Machine Learning course

Table 4.1: Grunt token inventory by sub-corpus.

(a) Weights Task Dataset

Token	n	Token	n	Token	n
a-ee	1	gasp	2	ope	1
a-i-eh	1	gee	1	ouh	1
aa-um	1	hm	26	sharp inhale	1
ah	14	huh	7	sigh	1
aw-ow	1	mf	1	singing	2
aw-wa	1	mm	35	snort	3
ch	1	mm-hm	51	ugh	1
cheez	1	mm-mm	2	uh	38
clap	1	mm-mm-mm-mm	1	um	61
click	2	oh	56	wa-wa	1
ee	2	oo-hoo	1	whoa	2
eh	3	oof	2	wow	6
en	1	ooh	21	yawn	1
er	2	oops	3	yo	1
fake reading	1	oow	2		

(b) Glider

Token	n	Token	n	Token	n
ah	5	hah	2	singing	1
ah-huh	1	huh	6	uh	34
aha	1	mm	1	uh-huh	6
aw	1	mm-hm	15	um	7
dadada	1	oh	23	wa-wow	1
ech	1	oh-oh-oh-oh	1	whistling	7
eh	3	oops	2		
er	1	sigh	7		

(COSI 230B) in 2025. The Glider video came with a detailed transcript that clearly marked non-lexical sounds throughout; this transcript was converted to my annotation format and received a second annotation pass by me.

Annotators independently used different spelling for perceived syllable length (e.g., *ooh* vs. *oooooh*) and they also varied their conventions for dashes (e.g., *mm-hm* vs. *mmhm*). For inter-annotator agreement, spelling was normalized: dashes were standardized and repeated letters removed (e.g., *hmmmm* $\rightarrow$ *hm*), except where indicating distinct conventional tokens

Table 4.2: Inter-annotator agreement for grunt detection and dialogue act classification.

Task	**DSC (range)**	α **(range)**
Grunt detection	0.498 (0.400–0.638)	0.421 (0.379–0.479)
Dialogue act	— (0.379–0.625)	— (0.379–0.501)

(e.g., *ooh* vs. *oh*). Orthographic vowel differences were preserved (e.g., *em* vs. *um*). A temporal tolerance of 500 milliseconds was used when aligning grunt annotations across annotators to measure agreement.

My annotations are treated as the gold standard throughout. For the prosodic analysis, I identified all standalone grunts and refined their temporal boundaries using the waveform visualizer in ELAN, limiting the analysis to cases where a grunt serves as a complete conversational turn in order to minimize acoustic bleed from surrounding speech.

4.2.5 Inter-Annotator Agreement

Agreement is reported using both the Sørensen-Dice coefficient (DSC) and Krippendorff's alpha (Dice 1945; Sorensen 1948; Krippendorff 2011).

For grunt detection, DSC ranged from 0.400 to 0.638 across doubly annotated videos, with an overall DSC of 0.498. Krippendorff's alpha ranged from 0.379 to 0.479 per video, with an overall alpha of 0.421. While moderate, this is comparable to other subjective audio annotation tasks such as emotion annotation (Lotfian and Busso 2017). Grunt counts from the two annotators were typically within 10% of each other; observationally, disagreements primarily occurred with overlapping speech and quiet vocalizations.

Gruntable identification proved substantially more difficult than grunt detection, with agreement on the initial task too low to report meaningfully. This reflects the inherent difficulty of the task: identifying what triggered a grunt requires retrospective judgments about temporal and causal relationships, and multiple plausible antecedents often exist within a short window. Similar to challenges with the GAMR annotation work, having annotators complete multiple span-based sub-tasks simultaneously introduces the opportunity for cas-

cading disagreements. For a second iteration of the task, one annotator from the Brandeis Natural Language Annotation for Machine Learning course (COSI 230B) in 2026 was given the gold-standard grunt annotations and asked to identify the gruntable for each non-*filled pause* grunt. By limiting the task to antecedent identification, it removes the possibility of a single disagreement on a grunt span leading to additional disagreements on corresponding gruntable scope.

Table 4.3: Inter-annotator agreement on gruntable type for a one-hour segment ($n = 20$ gruntables, both annotators identifying the same spans). Krippendorff's α (nominal).

Distinction	α	Raw agreement
Three-way (speech / event / speech DE)	0.52	70%
Core (speech vs. event)	0.88	95%

The sample had 20 non-filled-pause grunts. Disagreements between the 2 annotators concentrate at the speech/event boundary: four involved the speech-describing-event subcategory, and the single core-level disagreement arose where an event coincided with a concurrent verbal description of it (the scale being removed from the table by the researcher). With this revised setup, annotators reliably identified the modality of the grounded contribution, with minor disagreements on cases that are intermediate.

4.3 Formal Model

While AMR and its extensions are well-suited to representing the propositional content of communicative acts, grunts are primarily about belief states: whether information has been received, whether evidence has been accepted, whether a proposition has been promoted to an accepted fact in the common ground. This is the domain of dynamic epistemic logic, not predicate-argument representation, and it is why the formal grunt analysis employs EB-DEL rather than extending GAMR to a *Grunt AMR*.[2].

[2]The acronym GR-AMR is hard to resist in a computational linguistics dissertation, but the formal apparatus of AMR is not the right fit here

Here, I provide a formal account of how grunts update the common ground in situated dialogue. The model treats observable events as contributing propositional content like speech acts, and grunts as operations that can promote evidence to accepted fact. While the formal model is not directly evaluated in this study, it provides a precise theoretical vocabulary for the claim that grunts can be fact-promoting moves rather than only continuers, and offers a computational target for future DST systems that incorporate minimal vocalizations.

4.3.1 Framework

The Common Ground Structure (CGS) and its Evidence-Based Dynamic Epistemic Logic (EB-DEL) semantics were introduced in Chapter 3, following our work Y. Zhu et al. (2024) and Khebour, Lai, et al. (2024). Briefly, the CGS comprises three banks: QBank (questions under discussion), EBank (available evidence), and FBank (accepted facts).

Agents are participants P_1, P_2, P_3. Propositions range over the task: colored blocks $C = \{r, b, g, p\}$ and weights $W = \{10, 20, 30, \ldots\}$. We use predicates $\mathrm{Eq}(c, w)$ (c *weighs* w *grams*), $\mathrm{Neq}(c, w)$ ($c \neq w$), and a scale relation $\mathrm{Bal}(X, Y)$. For sums, $r{+}b$ is the mass of red and blue on one side of the scale.

Following the dialogue act taxonomy, a STATEMENT publicly contributes evidence, adding to EBank; an ACCEPT promotes evidenced content to fact, from EBank to FBank.

4.3.2 Detailed Walkthrough: Weights Task

Applying this formalism to an interaction from the WTD demonstrates how it captures the systematic relationship between observable events and grunts. The walkthrough parallels the balance-scale example in Y. Zhu et al. (2024), which established the framework's applicability to this task. The extension here is that the acceptance move promoting evidence to fact is performed by a minimal vocalization (*oh*) rather than by the explicit verbal agreement (*"Yeah"*) in the original example. Before the sequence in Figure 4.2, the initial CGS is:

- $\mathrm{FBank}_0 = \{\mathrm{Eq}(r, 10),\ \mathrm{Eq}(b, 10),\ \mathrm{Eq}(g, 20)\}$

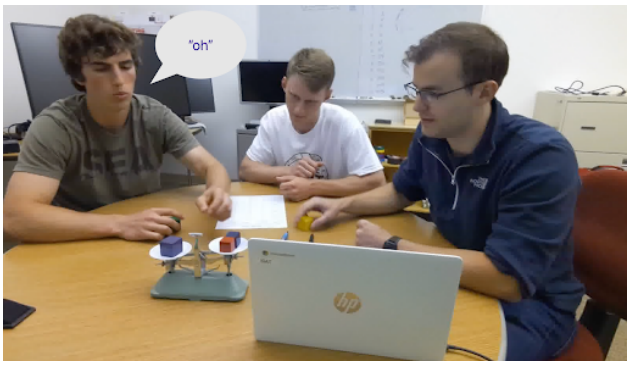


Figure 4.2: WTD Group 4 example with captions

- $\text{EBank}_0 = \varnothing$
- $\text{QBank}_0 = \{\text{Eq}(p, w)? : w \in W\} \cup \cdots$

Intuitively, red = 10, blue = 10, green = 20 are group facts; purple's weight is still questioned.

Step 1 (Event/Gruntable = statement). The second participant, P_2 places p against $r{+}b$ on the scale; it does not balance. The outcome, observed by all participants, is treated as a public announcement of an event proposition: $\varepsilon := \neg\text{Bal}(r{+}b, p)$. The event ε is defined as: it is not the case that the red and blue blocks together balance with purple. The banks after Step 1:

- $\text{FBank}_1 = \text{FBank}_0$ (no change to facts yet)
- $\text{EBank}_1 = \{\text{Neq}(p, 20)\}$ (new evidence)
- $\text{QBank}_1 = \text{QBank}_0 \setminus \{\text{Eq}(p, 20)?\} \cup \{\text{Eq}(p, 10)?,\ \text{Eq}(p, 30)?,\ \text{Eq}(p, 40)?,\ \text{Eq}(p, 50)?\}$

Step 2 (P_1 accepts with *oh*). The grunt *oh* is an ACCEPT of the evidenced proposition $\phi := \text{Neq}(p, 20)$. It is an acceptance announcement $[!\text{acc}_{P_1}(\phi)]$ whose precondition is $[E]\phi$ and whose postcondition promotes evidence to fact:

$$[E]\phi \ \wedge\ [!\text{acc}_{P_1}(\phi)] \quad \Rightarrow \quad [B]\phi.$$

Operationally in the CGS:

$$\text{EBank}_1 \ni \phi \xRightarrow{\text{ACCEPT}} \text{FBank}_2 \ni \phi$$

and ϕ is removed from EBANK. The banks after Step 2:

- $\text{FBank}_2 = \{\ldots,\ \text{Eq}(g, 20),\ \underline{\text{Neq}(p, 20)}\}$
- $\text{EBank}_2 = \varnothing$
- $\text{QBank}_2 = \text{QBank}_1$

The grunt functions as a *fact-promoting* move.

Step 3 (Speech referencing the new state). P_3 says *put the 20 on there.* This STATEMENT's felicity presupposes the updated FBank_2 — in particular, that $g = 20$ is known and that $p \neq 20$. Formally, it does not change the content above; it merely exploits the new CGS to propose the next action. This demonstrates how our formalism captures the systematic relationship between observable events, grunts, and common ground updates in collaborative problem-solving.

4.3.3 Illustrative Example: Cooking Scenario

The same mechanism applies across different situated contexts. Consider two participants P_1, P_2 cooking pasta together with a pot of water heating on the stove. Initially:

- $\text{FBank}_0 = \{\text{OnStove}(\textit{pot}),\ \text{Heating}(\textit{water})\}$
- $\text{EBank}_0 = \varnothing$
- $\text{QBank}_0 = \{\text{Ready}(\textit{pot})?\}$

The water begins to boil and the lid rattles. This observable event $\varepsilon := \text{Boiling}(\textit{water})$

functions as a STATEMENT, adding evidence that the pot is ready:

$$\text{EBank}_1 = \{\text{Ready}(pot)\}$$

P_1 responds with *ooh*, an ACCEPT that promotes the evidence to fact:

$$\text{EBank}_1 \ni \text{Ready}(pot) \quad \Downarrow \text{ ACCEPT} \quad \text{FBank}_2 = \text{FBank}_1 \cup \{\text{Ready}(pot)\}$$

P_2 adds pasta, presupposing the updated common ground. The grunt grounded the event and enabled coordinated action without a single word of explicit confirmation.

4.4 Findings

4.4.1 Overall Token Distribution

As shown in Table 4.4, grunts in these task-based settings most commonly function to ACCEPT a prior contribution (40.4%), whether verbal or non-verbal. This dominance of acceptance grunts supports the claim that these vocalizations serve critical grounding functions. The second most frequent category is FILLED PAUSE (39.8%), comprising primarily *um* and *uh* used as turn starters, consistent with findings in Clark and Fox Tree (2002).

There are some differences between the Glider and WTD distributions. The Glider participants tended to use filled pauses and turn-starters more often, possibly due to the longer periods of quiet work between the two men.

Interestingly, there are relatively few DOUBTs in either data set. As defined, DOUBT shows disagreement with a previous STATEMENT; in the formal model, DOUBT demotes propositions from the EBank back to the QBank. Commonly, a participant may be surprised at or not expecting a proposition, but they accept it anyway. The *oh* in the Detailed Walkthrough is a good example of this. Participant 1's expectation was violated (possibly indicated by the token choice and prosody, but he nevertheless accepted the proposition and moved to the next act.

Table 4.4: Grunt dialogue act distribution across Weights Task and Glider

	WTD		Glider		Total	
Dialogue Act	**N**	**%**	**N**	**%**	**N**	**%**
ACCEPT	160	43.6	41	31.5	201	40.4
Filled pause	126	34.3	72	55.4	198	39.8
OBSERVATION	32	8.7	1	0.7	33	6.6
ANSWER	21	5.7	3	2.3	24	4.8
DOUBT	20	5.4	4	3.1	24	4.8
Other	8	2.2	9	6.9	17	3.4
Total					497	100

Table 4.5 shows the distribution of antecedent types. (These are the same grunts as Table 4.4 cross-classified with different criteria.) While 41.6% of grunts lack clear antecedents (primarily filled pauses and turn starters), grunts responding to speech (25.4%) and observable events (25.1%) occur at nearly equal rates. This near-parity is the central finding of this chapter: non-verbal events function as conversational contributions on par with verbal utterances, a result with direct implications for multimodal approaches to information state tracking.

The distributions vary somewhat between the two sub-corpora, but the overall story of the gruntables is consistent. Even with the higher percentage of Glider grunts that have no antecedent, the remaining Glider grunts are split nearly evenly between having a speech antecedent and having an event one (16.9% to 13.8%).

Table 4.5: Antecedent type distribution across Weights Task and Glider.

	WTD		Glider		Total	
Antecedent Type	**N**	**%**	**N**	**%**	**N**	**%**
No antecedent	126	34.3	81	60.4	207	41.6
Speech	104	28.3	22	16.9	126	25.4
Event	107	29.1	18	13.8	125	25.1
Speech describing event	22	6.0	13	10.0	35	7.0

4.4.2 Token Specialization by Antecedent Type

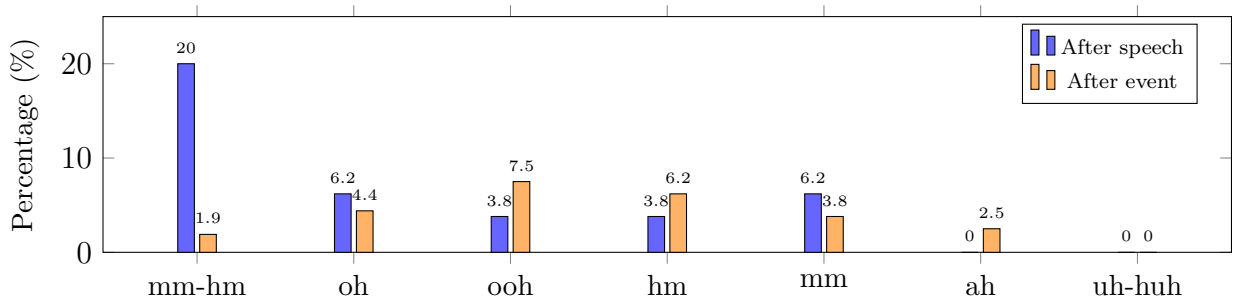
Figure 4.3c reveals functional specialization among grunt tokens based on antecedent type. Within ACCEPT responses, *mm-hm* shows a strong preference for speech antecedents (36.4%) over event antecedents (4.0%). Conversely, *oh*, *ooh*, and *hm* are used more frequently to acknowledge observed events. This specialization suggests that speakers maintain distinct repertoires for grounding verbal versus non-verbal information, with backchannels like *mm-hm* serving conversational continuity while reactive tokens like *oh* mark information state changes triggered by perceptual events.

In both sub-corpora, *mm-hm* strongly prefers speech, consistent with the description as continuers from Yngve (1970) and Schegloff (1982). *Uh-huh* does not appear in the WTD, but it is a common speech-following ACCEPT in the Glider corpus. Due to the small number of speakers in the data sets, and the large amount of time between their recordings, this may indicate preferences of the individual speakers, token entrainment among the groups in the interactions, or even changing vocabulary preferences due to eras. Tokens like *ooh* and *oh* may track the asymmetry in what is being grounded, and are not managing turn exchanges between speakers because they are registering new information to all participants.

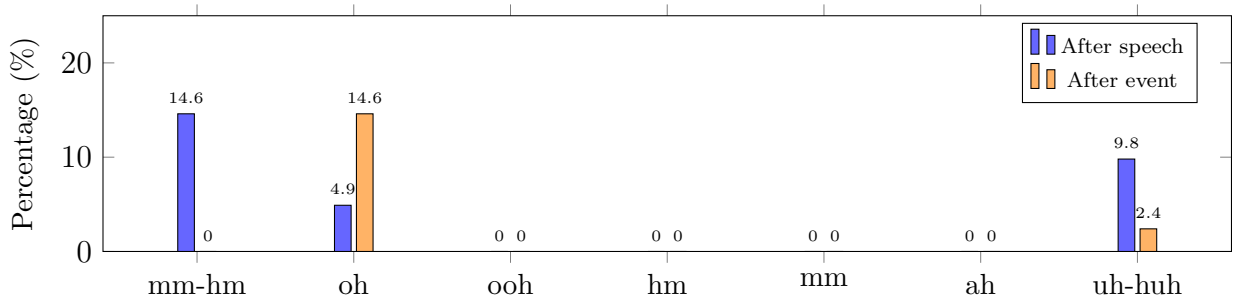
4.4.3 Task Effects: Glider vs. WTD

The two sections of corpora represent different interaction contexts: a problem-solving task among three university students in a laboratory setting versus two men assembling furniture at home. Despite comprising only 20% of the corpus, the Glider video shows distinctive patterns in antecedent distribution and token usage.

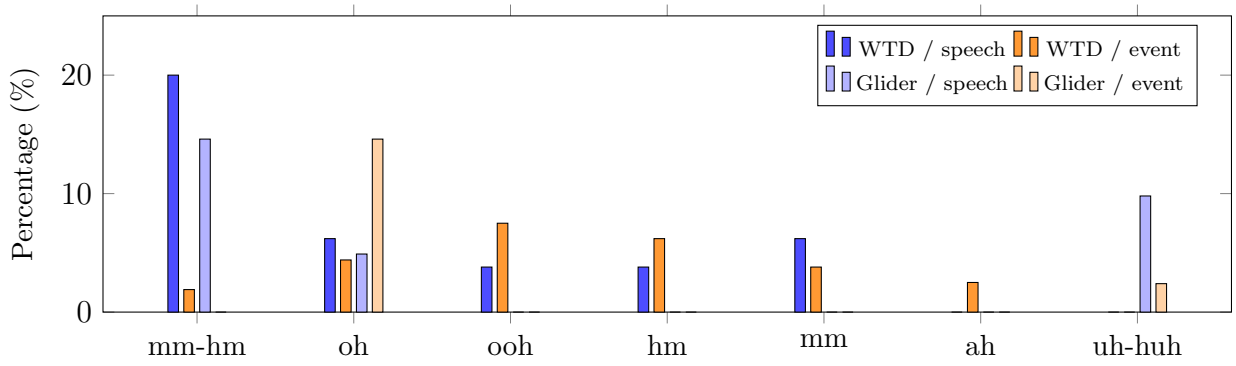
Glider contains more grunts without clear antecedents (56.9% vs. 36.2%) and fewer event responses (13.8% vs. 29.2%). This likely reflects the sequential nature of furniture assembly, where each step follows from the instructions rather than from collaborative inference about an uncertain outcome. In contrast, the WTD's problem-solving structure generates frequent observable events that require acknowledgment and interpretation.



(a) Weights Task ($n = 160$).



(b) Glider ($n = 41$).



(c) Both corpora.

Figure 4.3: Functional specialization among ACCEPT tokens. Each bar is a percentage of that corpus's ACCEPTs. In both corpora *mm-hm* prefers speech antecedents while *oh*, *ooh*, and *hm* lean toward events; divergent tokens reflect differing grunt inventories by the speakers.

Both corpora feature "meta-communicative" vocalizations, including whistling and humming, that seem to mark discourse segment transitions (Grosz and Sidner 1986). Celebratory vocalizations (*woo*, extended *oh*, singing) signal problem-solving success in WTD. The Glider participants whistle during long periods of otherwise quiet heads-down work. Speech resumes when a step is completed.

Token preferences also differ by task: Glider shows dominance of *uh* (25.4% vs. 10.1% in WTD), while *um* is more prevalent in WTD (16.6% vs. 5.4%). These differences may reflect individual speaker preferences (Cavalcanti and Skantze 2025), demographic factors (Bortfeld et al. 2001) task demands, or social dynamics, and underscore the importance of cross-corpus validation.

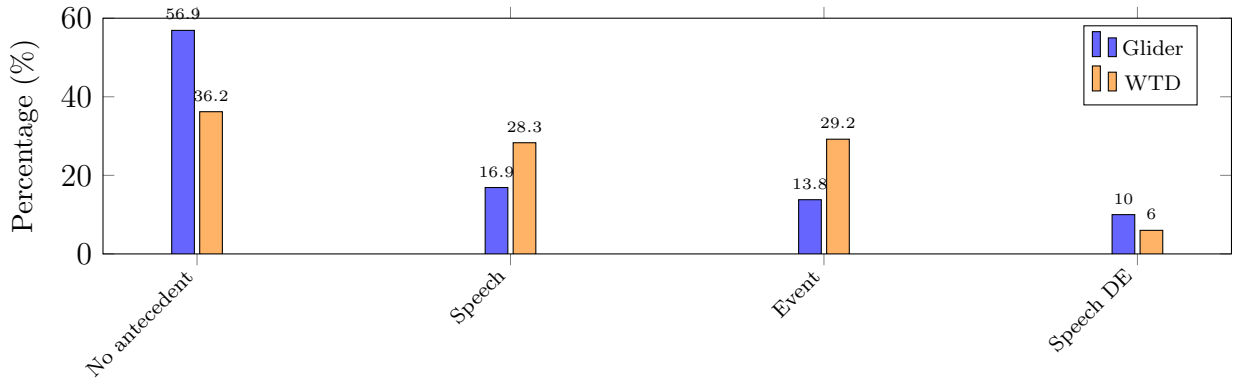


Figure 4.4: Antecedent type distribution by task context. Glider shows more grunts without clear antecedents and fewer event responses than WTD, reflecting differences in task structure.

4.4.4 Grunts as Complete Conversational Turns

Approximately 80% of non-filled-pause grunts in the corpus function as standalone conversational turns, that is complete responses that advance the dialogue state without verbal elaboration. A grunt following an observable event, such as *oh* after a scale reading, can constitute a sufficient turn that grounds the information and enables coordinated next actions without any further speech. This challenges dialogue models that treat minimal responses primarily as backchannels that hold the floor without contributing propositional content. It also motivates the prosodic analysis in Chapter 5: if grunts function as complete turns, their acoustic form should encode their grounding function in ways that go beyond token identity alone.

As a preview of the prosodic analysis, event-following grunts already show a markedly different acoustic profile than speech-following grunts in the Weights Task: they are 35%

Table 4.6: Acoustic features of standalone grunt responses by antecedent type (Weights Task). Mean (SD).

Antecedent	n	**Duration (s)**	**Pitch range (Hz)**	**F1 (Hz)**
Speech	90	0.472 (0.226)	93.7 (111.5)	515
Event	78	0.635 (0.530)	153.7 (160.8)	638
Speech DE	19	0.538 (0.197)	138.8 (172.7)	572

Note: Event vs. Speech differences were significant for duration, $t(101.1) = 2.53$, $p < .05$, $d = 0.41$; pitch range, $t(122.1) = 2.69$, $p < .01$, $d = 0.44$; and F1, $t(166.0) = 3.92$, $p < .001$, $d = 0.60$ (Welch's t-tests).

longer in duration, exhibit 64% greater pitch range, and show 24% higher F1 formant values. These differences suggest that speakers employ distinct strategies when reacting to observable events versus conversational speech, and that the grounding source is encoded not only in token choice but in acoustic realization. Chapter 5 develops this analysis in full.

The Glider corpus is excluded from the prosodic analysis. Its source recording and digitization process means that the audio is of lower fidelity than the Weights Task material. The Glider data yields unreliable formant and pitch estimates, with F1 values uniformly elevated and unstable pitch-range estimates. I believe this is likely a recording artifact rather than genuine acoustic differences. The Glider data is retained for the categorical dialogue act, token, and antecedent analyses, which do not depend on acoustic measurements.

4.5 Summary

This chapter established three results. First, the annotation scheme for grunts and gruntables provides a systematic framework for capturing the functional role of minimal vocalizations in situated dialogue, including the triggers to which they respond. Second, the EB-DEL formal model demonstrates that grunts are not merely continuers but fact-promoting moves: the grunt *oh* after a scale fails to balance is doing the same formal work as an explicit verbal acceptance of a statement, promoting a proposition from available evidence to accepted fact in the common ground. Third, the distributional findings establish the central empirical claim of the dissertation: grunts respond to speech and observable events at nearly equal

rates, and their token selection systematically encodes what type of information is being grounded. The following chapter asks how that encoding extends to the acoustic dimension.

5 PROSODIC ENCODING OF GROUNDING FUNCTION

The distributional findings in Chapter 4 established that grunts respond to speech and observable events at comparable rates, and that token selection encodes grounding source systematically. This chapter asks whether this encoding extends to the acoustic dimension. If grunts are doing different grounding work depending on what they respond to, does their form reflect that difference beyond the choice of token? I report findings from a prosodic analysis of standalone grunt responses (Brutti and Pustejovsky 2026b).

5.1 Motivation: Form and Function in Minimal Responses

It is well established that prosodic form encodes communicative function in minimal vocalizations (Ward 2000). Ward and Tsukahara (2000) identify low pitch regions as cues for backchannel opportunities in English and Japanese. Hockey (1993) shows that pitch contour distinguishes *okay* and *uh-huh* tokens with different dialogue functions. Most directly relevant, Clark and Fox Tree (2002) demonstrate that *uh* and *um* differ systematically in duration and planning function, with *um* signaling a longer upcoming pause. These works show precedence for form encoding function in non-lexical vocalizations that I extend to the grounding domain.

The literature on *oh* as a change-of-state token (Heritage 1984; Heritage 2002) is particularly relevant. Heritage characterizes *oh* as marking a transition in the speaker's informational state, as a signal that new information has been registered and integrated. This account is consistent with the token specialization findings in Chapter 4. The present analysis provides quantitative corpus-level support for what has largely been described through conversation analytic methods.

The mirativity literature offers a related theoretical frame. *Mirativity* in linguistics is the grammatical marking of surprise or novelty with information (DeLancey 1997; Aikhenvald 2012). Kraus (2019) describes English discourse particles like *oh* and *huh* as mirative markers whose prosodic contours are systematically heightened when the speaker's expectations are

violated, or when there is new information. Expectation-violating observable events in the task-based settings include a scale failing to balance, or a furniture piece not fitting. Moments like this predict that event-following grunts should show more marked prosodic profiles than speech-following grunts.

Event Segmentation Theory (Radvansky and Zacks 2014) provides a cognitive account of why this might be so. Observable events are parsed into discrete units at points of high prediction error, triggering updating of mental event models. If event-following grunts mark these segmentation points, they may reflect increased processing demands relative to grunts that simply acknowledge a verbal contribution. This is a nice theoretical connection, though the step from acoustic features to cognitive mechanisms is speculative without experimental evidence.

A related observation, suggested by an anonymous Cognitive Science (CogSci) reviewer, is that LLMs employing chain-of-thought reasoning produce tokens such as *wait*, *hmm*, and *oh* at points of internal state transition. These are the contexts where human speakers produce change-of-state tokens. Whether this reflects a general mechanism for marking information state updates or is an artifact of training data remains an open question, but the parallel is worth noting.

5.1.1 Corpus and Preprocessing

The prosodic analysis draws on the same corpus as Chapter 4: nine WTD videos and the Glider video. The prosodic analysis is restricted to *standalone* grunts, i.e., those that constitute a complete conversational turn. This approach minimizes acoustic bleed from surrounding speech. Temporal boundaries for standalone grunts were refined by hand using the waveform visualizer in ELAN prior to feature extraction. Speaker-normalized prosodic features have previously been used to distinguish interactional and affective states in collaborative dialogue (Goodman et al. 2006).

The WTD recordings were captured with a single room microphone for all 4 speakers (3

unique participants per video, plus the same researcher-facilitator for all videos), and are not channel-separated. The Glider video was recorded on a camcorder and subsequently digitized. As a result, all videos have variability in recording quality and acoustic quality across speakers. No additional filtering was applied beyond the boundary refinement step described above. Formant estimation is sensitive to recording quality and background noise, so the results from the two corpora may not be directly comparable.

Acoustic features were extracted using Parselmouth (Jadoul, Thompson, and De Boer 2018), a Python interface to Praat (Boersma and Weenink 2021). Four features were extracted for each standalone grunt: duration (total vocalization length in seconds), mean F0 (fundamental frequency, reported in Hz), pitch range (difference between maximum and minimum F0 in Hz), and mean F1 (first formant frequency in Hz). Additionally, as pointed out by a helpful Cognitive Science review, Parselmouth does not currently support the most recent pitch detection algorithm introduced in newer releases of Praat; pitch results presented here should therefore be interpreted with some caution. Ward Ward (2006) suggests loudness/intensity plays a role, but I omitted that measurement due to the single microphone setup.

The choice of features is motivated by the cognitive and affective processing literature. Duration indexes processing effort and arousal. Pitch range reflects prosodic expressiveness and has been associated with information status marking. F1 correlates with vowel openness and has been used as an index of vocal effort, following Scherer (2003).

Pitch perception is non-linear, and F0 values are conventionally expressed in semitones normalized to a speaker-specific baseline. This chapter's analysis converts F0 measurements to semitones relative to each speaker's mean F0.

A note on F1 interpretation: F1 reflects vowel quality and is therefore sensitive to lexical choice as well as to paralinguistic modulation. Because the grunt tokens in this analysis are lexically distinct (*mm-hm* is nasal with low F1, while *oh* and *ooh* involve open vowels with high F1), direct cross-token comparison of F1 values conflates token identity with prosodic

modulation. The cross-token F1 reflect the combined effect of token selection and prosodic realization rather than purely paralinguistic encoding. A within-token analysis comparing F1 values for *oh* tokens across antecedent types provides a cleaner test.

Statistical analysis was conducted using linear mixed-effects models implemented in statsmodels (Seabold and Perktold 2010), with speaker included as a random intercept to control for between-speaker variation in baseline characteristics. The primary analysis is restricted to ACCEPT dialogue acts to control for the effect of dialogue act function on prosodic realization.

Each speaker is allowed their own baseline level of the outcome (duration, F0, pitch range, or F1), so the antecedent coefficients β_1 and β_2 estimate how much event-following and Sp→Ev-following grunts differ from speech-following grunts (the reference category). This keeps speakers who happen to contribute many grunts from disproportionately driving the estimated effects. Participant labels (p1, p2, p3) are not unique across groups in the WTD corpus, so speaker identity was constructed as a composite of group and within-group ID. The researcher (r) is the same for all WTD groups. Significance was assessed by t-tests on fixed effects with effective degrees of freedom approximated based on the within- and between-speaker variance components. Cohen's d is reported with 95% confidence intervals as effect-size estimates alongside model coefficients.

A token-controlled model added grunt token as an additional fixed effect (*mm-hm* as reference, restricted to the six tokens with $n \geq 5$ in the ACCEPT subset) to test whether antecedent effects on prosody persist after accounting for which token speakers selected. We also fit a within-token model for *hm*, the only token with sufficient cells in both speech and event conditions, as an exploratory analysis.

For token distribution analysis, chi-square tests evaluate whether token choice (e.g., *mm-hm* vs. *oh*) varied systematically by antecedent type.

5.2 Token Specialization by Antecedent Type

A chi-square test restricted to the six most frequent tokens (n = 190 standalone grunts) revealed significant specialization by antecedent type: $\chi^2(10) = 48.95$, $p < .001$. Backchannels predominantly follow speech: *mm=hm* follows speech contributions in 75.8% of cases, rising to 95.2% when speech-describing-event antecedents are folded in with speech. Reactive tokens pattern in the opposite direction, with *ooh* and *hm* preferentially following observable events (66.7% and 57.7% respectively).

This functional specialization suggests that speakers select tokens based on the type of information being grounded. Tokens traditionally described as backchannels function as acknowledgments of verbal contributions, while reactive tokens may encode surprise or novelty at unexpected events. The predominant use of mm-hm for speech and preference for oh and ooh following events indicates that lexical selection may provide coarse-grained categorical encoding of grounding source, operating in coordination with prosodic modulation to signal multimodal grounding.

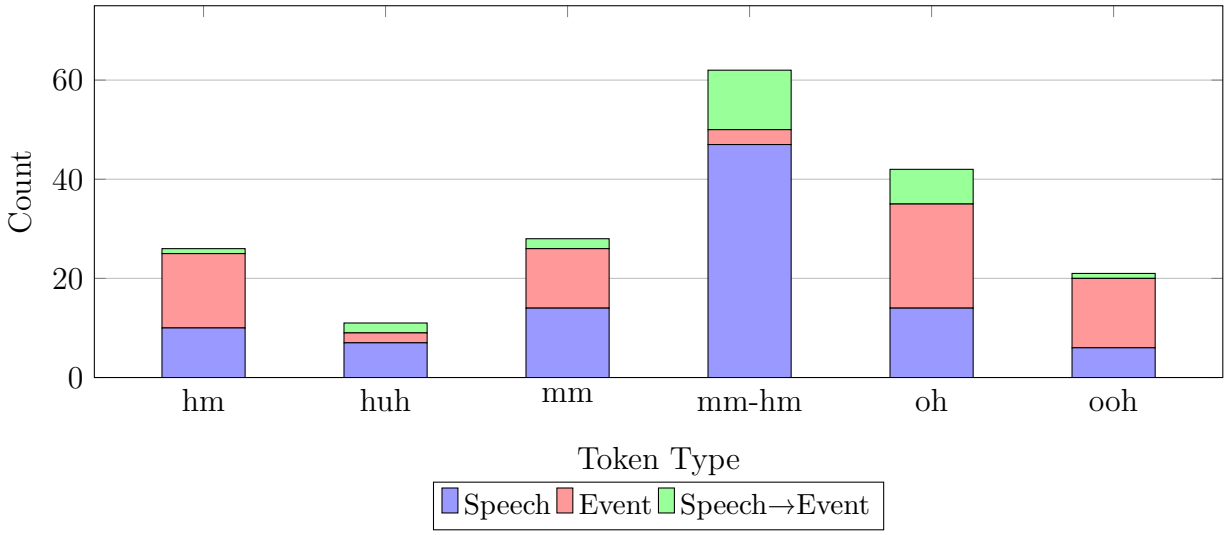


Figure 5.1: Distribution of grunt tokens by antecedent type ($n = 190$ standalone grunts across the six most frequent tokens). Backchannels (*mm-hm*, *mm*) predominantly follow speech contributions, while reactive tokens (*oh*, *ooh*, *hm*) preferentially follow observable events. Stacked bars show counts for each token–antecedent combination.

Token specialization in backchannels is a well-established *entrainment* phenomenon.

Speakers align lexical and prosodic choices to those of their interlocutors over the course of an interaction (Levitan and Hirschberg 2011). Without controlling for entrainment, it is difficult to fully disentangle contextual effects from alignment processes. A group-level analysis of entrainment in grunt token choice is a natural extension of this work, and is discussed with the Portal Dialogue Corpus work in Chapter 6, as the same pair of participants completes multiple puzzles together.

5.3 Prosodic Encoding of Grounding Source

5.3.1 ACCEPT Responses by Antecedent Type

The primary prosodic analysis compares standalone ACCEPT grunts following speech antecedents ($n = 77$) and observable event antecedents ($n = 50$). Table 5.1 shows the acoustic features by antecedent type.

Table 5.1: Acoustic features by antecedent type (ACCEPT responses, mean acoustic features). F0 reported in semitones relative to each speaker's median.

Antecedent	n	Dur (s)	F0 (ST)	Range (Hz)	F1 (Hz)
Speech	77	0.43	−0.24	101	609
Event	50	0.62	+1.21	140	677
Sp→Ev	23	0.49	+1.87	134	683

A speculative account is that longer duration iconically reflects increased processing demands at event segmentation boundaries: the grunt is literally extended in time to accommodate the additional cognitive work of grounding a perceptual event. A cleaner test would involve manipulating event complexity or predictability experimentally.

Mixed-effects models (random intercept on speaker) showed that event-following grunts were significantly longer than speech-following grunts ($\beta = 0.147\,\text{s}$, $SE = 0.064$, $t = 2.28$, $p = .024$; $d = 0.49$, 95% CI $[0.13, 0.85]$) and exhibited higher F1 values ($\beta = 98.4\,\text{Hz}$, $SE = 36.7$, $t = 2.68$, $p = .009$; $d = 0.27$, 95% CI $[-0.09, 0.62]$). Mean F0 in semitones ($\beta = 1.40\,\text{ST}$, $p = .31$) and pitch range ($\beta = 22.1\,\text{Hz}$, $p = .47$) did not differ significantly.

Speech-describing-event responses (Sp→Ev) (grunts produced when the immediately pre-

ceding utterance was a verbal description of an event) showed an acoustic profile that mixes features of both reference categories. On vowel quality (F1 = 683 Hz), pitch (F0 = +1.87 ST relative to speaker median), and pitch range (134 Hz), Sp→Ev grunts pattern with event-following responses; on duration (0.49 s), they pattern more closely with speech-following acknowledgments. Mixed-effects models showed no statistically significant differences between Sp→Ev and either reference category at the current sample size ($n = 23$; all $p > .19$), so this pattern is descriptive only.

5.3.2 Disentangling Token and Prosodic Modulation

Because grunt tokens differ inherently in their acoustic properties (*mm-hm* is a brief nasal murmur with low F1, while *oh* and *ooh* require an open vowel of longer duration), the population-level prosodic differences reported above could reflect either (a) systematic prosodic modulation that operates independently of token choice, or (b) the inherent acoustics of the tokens that speakers select for each antecedent type. This is a question of how much prosody can tell us about grounding source beyond what the token choice conveys. Recent work estimates channel quantities in information-theoretic terms for features such as sarcasm and affect (Yadavalli et al. 2025).

To distinguish these possibilities, I re-fit the prosodic models with grunt token included as a fixed effect (restricted to the six tokens, *hm, huh, mm, mm-hm, oh, ooh*, with $n \geq 5$ in the ACCEPT subset; $n = 107$).

After controlling for token, the duration effect was no longer detectable ($\beta = 0.012$ s, $p = .79$) and the F1 effect was reduced to a marginal trend ($\beta = 81.0$ Hz, $p = .087$). F0 in semitones and pitch range remained non-significant. This pattern indicates that the population-level prosodic differences in these corpora may be largely attributable to token selection rather than independent prosodic modulation by antecedent.

An exploratory within-token analysis was possible only for *hm*, the single token with adequate cells in both conditions (speech $n = 5$, event $n = 7$). Effect sizes were consistent

with prosodic modulation (duration $d = 0.89$; F1 $d = 1.28$), but the within-token mixed-effects coefficients did not reach significance given the very small sample (all $p > .23$). For *oh* ($n_{\text{event}} = 4$, $n_{\text{speech}} = 7$) and *mm-hm* ($n_{\text{event}} = 3$, $n_{\text{speech}} = 34$), the cell distributions were too imbalanced for meaningful within-token testing. Whether speakers prosodically modulate a given token by antecedent therefore remains an open empirical question. The larger sample size of the Portal corpus may provide additional insight.

Splitting the corpus by source confirmed that the population-level effects (duration, F1) were carried by the WTD corpus ($n = 125$, 26 speakers), where they replicated (duration $p = .032$, F1 $p = .001$). The Glider corpus ($n = 25$, 2 speakers, 4 event tokens in ACCEPT) was too small to detect effects in either direction. This pattern is consistent with the recording quality discussion in Chapter 4. The effects replicate in the WTD corpus, and the Glider sample is too small and acoustically variable. This also motivates the Portal section, where each speaker has their own microphone.

5.3.3 The Speech-Describing-Event Category

The speech-describing-event (Sp→Ev) antecedent type is for cases where a participant verbally describes or narrates an observable event, rather than the event itself serving as the gruntable. This condition offers a useful probe of where the lexical-prosodic boundary falls. These show an intermediate acoustic profile. Duration (0.485s) sits between the speech (0.453s) and event (0.626s) values, as does F1 (686Hz between speech 582Hz and event 652Hz).

When a partner verbally describes a co-perceived event, the listener's grunt selects from a token inventory whose acoustic properties (open vowel, wider pitch range, F0 elevated relative to the speaker's baseline) resemble those of grunts following directly perceived events, even though the proximate antecedent is speech. Duration, however, patterns more like a speech-following acknowledgment. One possibility is that token selection tracks the kind of contribution being grounded (a state change in the world) regardless of whether it was

acquired perceptually or verbally, while duration tracks the proximate processing demands of the antecedent.

This is a tentative interpretation given the small Sp→Ev sample (n = 23), but it may suggest that the Sp→Ev case is worth treating as theoretically distinct. Verbal framing of a physical event may partially shift the grunt toward speech-type patterns, but does not fully override the event grounding signature. This suggests that grounding source is gradient rather than categorical. Cognitive demands of grounding an event described in speech are between grounding direct perception and grounding a purely verbal assertion. The small n makes this finding suggestive rather than conclusive, and the Sp→Ev category warrants dedicated analysis in future work with larger samples.

5.3.4 Prosodic Variation by Dialogue Act

Prosodic variation within individual tokens across dialogue acts further demonstrates that speakers draw on acoustic resources to disambiguate grunt function when token alone is insufficient. *mm-hm* produced as an ACCEPT is 18% shorter in duration and shows 57% less pitch range than *mm-hm* produced as an ANSWER, for example, suggesting that the same token is prosodically modulated to signal different grounding functions. *hm* shows the most dramatic variation: duration ranges from 0.245s in question-like uses to 0.934s in DOUBT expressions. While the within-token analysis above could not confirm prosodic modulation across antecedent types, a complementary pattern is visible across dialogue acts: the same token is realized differently depending on the grounding function it performs. This dialogue-act variation indicates that prosodic modulation operates within tokens for at least some functional contrasts, even though the antecedent-level modulation tested earlier remained inconclusive.

5.4 Discussion

These findings point to grunt token selection as the primary mechanism of differentiating grounding sources. Token choice provides a coarse-grained reliable signal about STATEMENT source. At the corpus level, event-following grunts are also longer in duration and higher in F1 than speech-following grunts. Once grunt token is entered into the model, however, the duration effect is less detectable and the F1 effect is reduced to a small trend. The narrowest interpretation is that speakers do most of the differentiating through grunt token choice, and that the prosody of the resulting grunts follow largely from the inherent acoustics of the selected vocalizations rather than from independent prosodic modulation by antecedent.

In this data, the interpretation is demonstrated by the link between the tokens and their vowels. The grunt tokens that prefer event gruntables typically have more open vowels than those that prefer speech ones. So, the F1 difference is likely expected based on token alone. Cross-token F1 comparisons reflect a combination of token choice and vowel quality; it is unclear if it is a purposeful modulation of the tokens. The larger sample size of the Portal corpus allows for this analysis.

This does not rule out independent within-token modulation, but indicates that the present corpus cannot establish it. The exploratory within-token analysis of the only token with adequate counts in both antecedent conditions, namely *hm*, showed large effect sizes in the predicted direction, though these did not reach significance at just 12 examples. The prosodic variation observed across dialogue acts points the same way: the same token is realized differently depending on the function it performs. Taken together, these patterns suggest that within-token modulation is real, but small, relative to token selection, and that detecting it requires denser sampling and annotation than either sub-corpus provides. Whether speakers prosodically modulate a given token by antecedent therefore remains an open empirical question from WTD and Glider.

The prosodic analysis reported in this chapter was published at Cognitive Science 2026 (Brutti and Pustejovsky 2026b). The peer-review process directly shaped the analysis pre-

sented here: F0 is reported in semitones normalized to each speaker's median rather than in raw Hz; F1 is interpreted as a property of vowel quality; and the token-controlled model was added to separate token selection from prosodic modulation. The larger size of the Portal corpus will allow for an analysis of lexical entrainment between participants, because the same dyads complete multiple tasks together. The larger size of the corpus will also allow for at least one version of the within-token question.

5.5 Summary

This chapter reported a prosodic analysis of standalone ACCEPT grunts. On the whole, event-following grunts are longer in duration and higher in F1 than speech-following grunts; once grunt token is included in the model, however, the duration effect shrinks. The primary mechanism by which speakers differentiate grounding sources is token selection, which encodes information source at a categorical level, while the corpus-level prosodic differences follow largely from the inherent acoustics of the selected tokens. Whether speakers additionally modulate prosody within a token by antecedent remains an open question: the exploratory *hm* analysis and the variation observed across dialogue acts are consistent with that idea, but the WTD + Glider corpus is too sparse to establish it. The intermediate acoustic profile of speech-describing-event grunts is likewise suggestive of gradient rather than categorical grounding, though too small at the current sample size.

The analysis in this chapter was published at Cognitive Science 2026 (Brutti and Pustejovsky 2026b), and the methodological refinements raised in peer review—speaker-normalized F0, the token-controlled model, and the corpus-split robustness check—are incorporated above. Two questions remain open: prosodic and lexical entrainment between interlocutors, and the within-token modulation question, both of which the Portal corpus is positioned to address. Chapter 8 turns to the structural dimension of grunt function—where in a turn grunts occur, and what that position reveals about their grounding role.

6 THE PORTAL DIALOGUE CORPUS

To this point, the dissertation establishes three results. Non-lexical communicative acts have representable propositional content (Chapter 2). Grunts perform fact-promoting moves in the common ground at rates comparable to speech (Chapter 4). Their acoustic form encodes grounding source systematically (Chapter 5).

The preceding two chapters established their central findings on a limited corpus base: the distributional analysis of Chapter 4 and the prosodic analysis of Chapter 5 both drew primarily on the Weights Task, supplemented by the shorter Glider recording. A natural question is whether these findings reflect properties of grunts in general or artifacts of these particular tasks and speakers. This chapter addresses that question by reproducing both the distributional and prosodic analyses on the Portal corpus (Tomlin et al. 2025), a structurally distinct collaborative setting. Portal comprises 18 dyads each completing multiple video game puzzle tasks together in a virtual co-situated environment, recorded with individual microphones, as the participants are not physically co-located. These differences make Portal a test of generalization rather than a confirmation: the dyadic structure, distinct task demands, and repeated pairing each represent a dimension along which the earlier findings might or might not hold. They also enable analyses the earlier corpora could not support, including more reliable prosodic measurement and denser within-speaker sampling.

Pairs work together in the first-person gameplay in Portal 2's cooperative mode. Portal 2 is a puzzle video game [1] where players can place portals and teleport between them in order to move objects around in order to solve the puzzles. Players' avatars and actions are mutually visible in-world from each other's point-of-view.

Portal was selected from a broader search of multimodal dialogue corpora as a complement to WTD and Glider. Several multimodal corpora were considered before Portal; many are surveyed in Anikina, Leippert, and Ostermann (2025). Each was evaluated against the requirements that the analyses in this dissertation impose. Any corpus setting must involve

[1] https://www.thinkwithportals.com/

two or more participants engaged in a shared task that generates observable, groundable events; this excludes single-participant recordings (Damen et al. 2022), speech-primary interactions (Kraaij et al. 2005), and corpora whose interactions are captured primarily in the chat modality (Afantenos et al. 2012; Lopes, Hemmingsson, and Åstrand 2018; R. Zhang and Eickhoff 2021). The study of prosody requires clean, per-speaker audio, which rules out corpora distributed without audio (Zarrieß et al. 2016; A. Zhu et al. 2023) or recorded on a single shared microphone (Mughal et al. 2024). Also, verifying that events are mutually observable requires video, excluding audio-only resources (Anderson et al. 1991), or resources that cannot distribute video (Kontogiorgos, Sibirtseva, and Gustafson 2020). And lastly, to keep the grunt inventory comparable with the Weights Task and Glider and to avoid confounding cross-linguistic differences in non-lexical vocalization, the corpus had to be in US English; this set aside otherwise suitable corpora in other languages (Kok and Heylen 2011; Ueda et al. 2024). Many otherwise excellent corpora fail one of these requirements, or carry access restrictions on their video portions that placed them beyond reach.

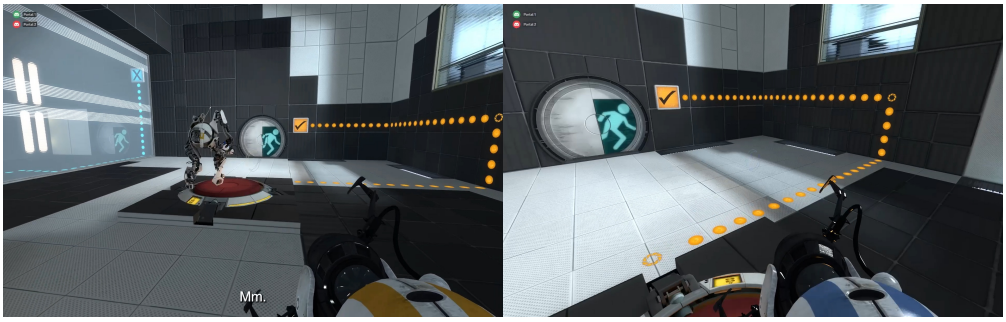


Figure 6.1: The Portal Dialogue Corpus: Stitched view. Blue and Orange players have their own perspectives in the same virtual space.

6.1 Portal Distributional Analysis

As shown in Table 6.2, the Portal corpus tracks closely with the Weights task in terms of the distribution of the grunt tokens as Dialog Acts. Portal Sessions are organized in dyads, where the same people work together completing 6 multiple puzzle levels each. Session 01 Course 1 and Session 12 Course 1 were annotated with the grunt scheme described in Chapter 4,

Table 6.1: Grunt token inventory for the annotated portion of the Portal Dialogue Corpus

Token	n	Token	n	Token	n
agh	1	nn	1	thinking noises	1
ah	28	oh	287	ugh	4
aw	2	oh-oh-oh	1	uh	171
ay	2	oof	11	uh-huh	3
blech	1	ooh	15	uh-oh	5
doo-doo-doo-doo	1	ooh-ooh	1	um	30
ee	2	oop	2	whee	17
eh	9	oops	4	whoa	7
hm	25	ope	1	woo-hoo-hoo	1
hu-um	1	ow	1	wow	3
huh	10	oy vey	2	yay	4
mm	9	phew	2	yippee	1
mm-hm	8	singing	3	yo	1

totaling 12 videos lasting approximately 65 minutes. There were two unique speakers in each Session, for a total of four. Full token counts are shown in Table 6.1.

An observation about OBSERVATION and the *oh* token: annotating OBSERVATION vs ACCEPT was difficult because players walk through a virtual world. They are constantly seeing different rooms, portals, and other things in the environment. I made the decision to annotate standalone *oh* as ACCEPT with the rationale of the visualization of new environments or items in the world served as proposition which the player accepts and uses for navigating the situation and the puzzle. Since different items may or may not be useful to the players, at the time of their appearance, I annotated as ACCEPT since the state of the items may end up as part of the Common Ground for the players solving the puzzle. Grunts occur far more frequently in Portal (10.5 per minute) than in the Weights Task (2.6 per minute) or Glider (3.2 per minute).

One additional note on the token *oh* in Portal. In the dyads I chose, Session 12 in particular, used the phrase *oh my god* several times. During annotation, I did not have a graceful way to handle these, so I parked them as FILLED PAUSEs because they are part of a larger bound phrase. In an analysis of the transcripts, I found that approximately 10% of

Table 6.2: Grunt dialogue act distribution across Weights Task, Glider, and Portal

	WTD		**Glider**		**Portal**	
Dialogue Act	**N**	**%**	**N**	**%**	**N**	**%**
ACCEPT	160	43.6	41	31.5	312	45.9
FP + Bound	126	34.3	72	55.4	296	43.5
OBSERVATION	32	8.7	1	0.7	6	0.9
ANSWER	21	5.7	3	2.3	30	4.4
DOUBT	20	5.4	4	3.1	29	4.3
Other	8	2.2	9	6.9	7	1.0
Total	367	100	130	100	680	100

the *oh*s in Portal FILLED PAUSEs were *oh my god*s, so in the tables in this chapter, I will label this category as *Filled pause + Bound*. This does not impact either the standalone or ACCEPT analysis.

Table 6.3 shows the distributions of grunts to all antecedent types. This counts all grunts, regardless of dialogue act and utterance position. As before, these are the same grunts as Table 6.2 classified with different criteria. In the Portal corpus, grunts respond to events more often than they do to speech by a significant margin. By far, this is the highest proportion of grunts responding to events of the 3 data sets. The Portal task is a far more event-dense environment: participants contend with moving mechanical elements (doors, elevators, lasers) and with information on in-game signs and screens, while collaborative verbal reasoning is comparatively sparse. The event-grounding rate thus scales with the event density of the task. That events remain a major groundable category across all three corpora supports the broader claim that observable events function as contributions requiring acknowledgment rather than a minor or task-specific phenomenon. They are dominant in Portal, and near parity in the Weights Task.

Table 6.3: Antecedent type distribution across Weights Task, Glider, and Portal

	WTD		Glider		Portal	
Antecedent Type	**N**	**%**	**N**	**%**	**N**	**%**
No antecedent	126	34.3	81	60.4	296	43.5
Speech	104	28.3	22	16.9	88	12.9
Event	107	29.1	18	13.8	274	40.3
Speech describing event	22	6.0	13	10.0	22	3.2

6.2 Portal Token Specialization by Antecedent Type

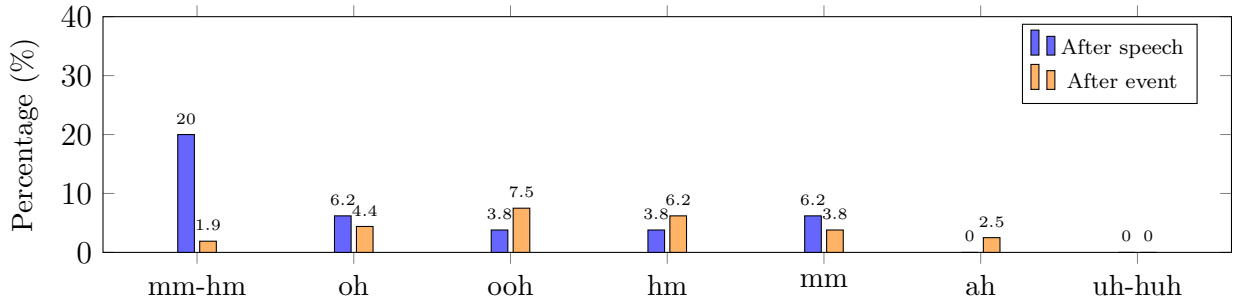
The token specialization in Portal replicates my earlier findings on the event side, but the effect is not visible on the speech as antecedent side. In all three corpora, *oh*, *ooh*, and *hm* all lean toward event gruntable antecedents. However, *mm-hm*'s strong preference for speech, which is robust in the Weights Task and Glider, is not apparent in Portal.

This follows from Portal's observed interactional profile. There is little sustained verbal exchange in the annotated dyads, so there is relatively less conversational speech available to acknowledge. Only 12.9% of Portal grunts follow speech, so the speech-continuer function that *mm-hm* fills has fewer occasions to appear. The event-reactive tokens, by contrast, have abundant triggers and dominate: *oh* alone accounts for 49% of Portal ACCEPTs, mostly following events.

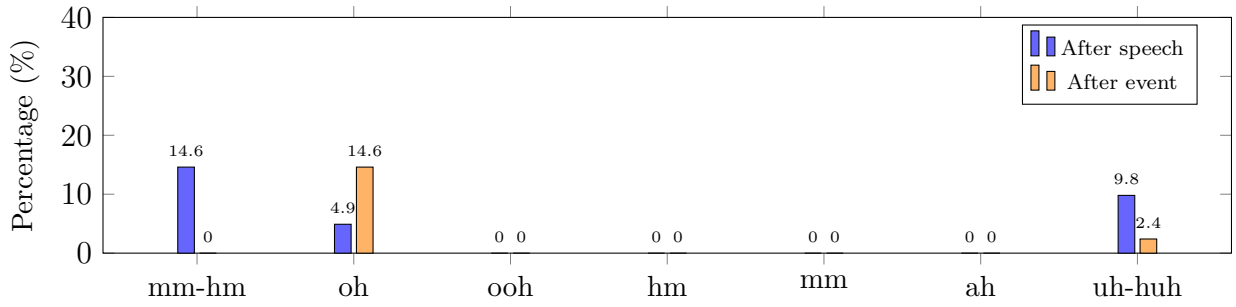
Portal also shows a wider inventory of reactive tokens than the other corpora—*whee*, *oof*, *whoa*, *uh-oh*, *yay*—reflecting the affective stakes of game success and failure. These fall outside the seven-token set common to all corpora and are event-driven as shown in Table 6.4.

The event-reactive specialization thus generalizes across all three settings, while the speech-continuer specialization is contingent on the presence of verbal contributions to be grounded. This is consistent with the account on which token choice tracks the source of the grounded contribution: where events predominate, so do event-grounding tokens.

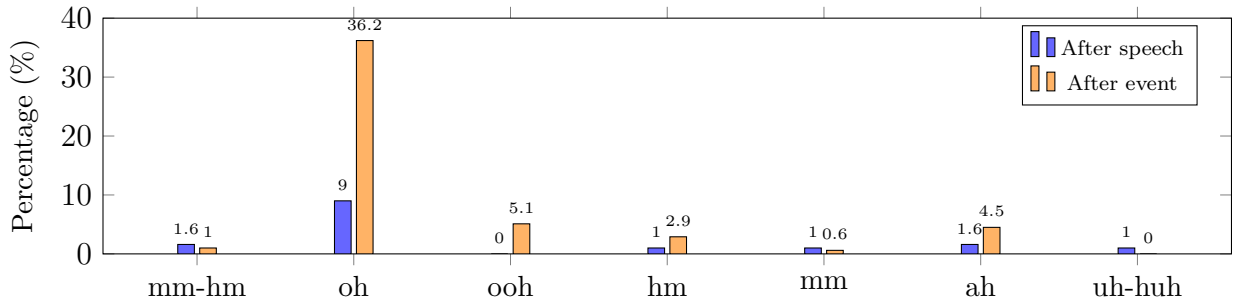
Roughly half of Portal grunts constitute complete turns, compared to about 80% in the Weights Task and Glider. The remaining Portal grunts are embedded in larger turns—often,



(a) Weights Task ($n = 160$).



(b) Glider ($n = 41$).



(c) Portal ($n = 312$).

Figure 6.2: Functional specialization among ACCEPT tokens. Each bar is a percentage of that corpus's ACCEPTs.

it appears, opening them rather than standing alone. Why event-dense interaction should shift grunts from complete turns toward turn-initial position is a question about the structural role of grunts, which the next chapter takes up directly: Chapter 7 examines where in a turn grunts occur and what that position reveals about their grounding function.

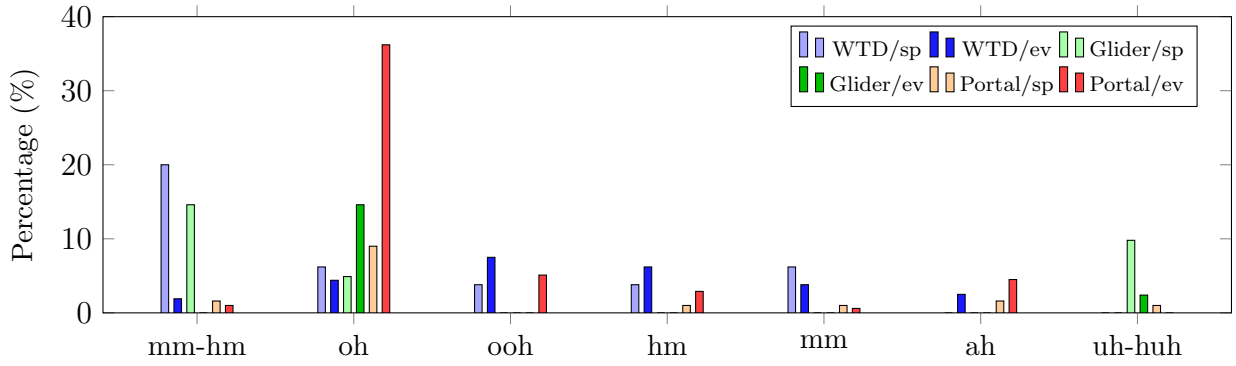


Figure 6.3: Combined view for all 3 corpora. Divergent tokens reflect differing grunt inventories across the speaker groups.

Table 6.4: Expressive and reactive ACCEPT tokens in Portal outside the six-token inventory shared across corpora ($n = 90$, 28.8% of Portal ACCEPTs). Nearly all follow events.

Token	n	Token	n
whee	17	oops	4
oof	9	um	4
whoa	7	eh	4
huh	5	yay	4
uh	5	ugh	3
uh-oh	5	*13 singletons*	13

6.3 Portal Prosody

Chapter 5 found that event-following ACCEPT grunts in the Weights Task differ prosodically from speech-following ones. This section asks whether that pattern extends to Portal. Because the Portal subset was sampled for depth rather than breadth, there are two dyads recorded over 6 videos each. This means that the prosodic analysis here relies on just 4 speakers, so I report it as exploratory corroboration.

As with the WTD and Glider analysis in Chapter 5, I focus the prosodic analysis on standalone ACCEPTs. This clearly delineates the tokens from surrounding ones. Portal's time-stamped transcription allowed for easier extraction based on a list of grunt tokens, so I did not refine word boundaries myself. This leaves 157 standalone ACCEPTs for the analysis. Fundamental frequency is reported in semitones relative to each speaker's median F0, as before. Because only four speakers contributed standalone ACCEPTs in both an-

tecedent conditions, the primary test aggregates to per-speaker means and compares speech- and event-following grunts within speaker. I also omitted loudness/intensity for the Portal analysis due to the likely presence of Automatic Gain Control (AGC) on the voice-chat microphones.

Event-following grunts are higher and wider in pitch than speech-following grunts (Table 6.5): roughly five semitones higher in mean F0 (+3.69 vs. −1.21 ST) and more than double the pitch range (166 vs. 70 Hz). The pitch effect survives speaker-level aggregation: on per-speaker means, event-following grunts are significantly higher in F0 (paired $t(3) = 3.77$, $p = .033$) and wider in range ($t(3) = 2.71$, $p = .073$). Duration, by contrast, does not differ (0.52 vs. 0.49 s, n.s.). F1 is higher for event-following grunts (649 vs. 590 Hz) but does not differ reliably given the small speaker sample.

Table 6.5: Acoustic features of standalone ACCEPT grunts in Portal by antecedent type. Mean (SD). F0 in semitones relative to each speaker's median.

Antecedent	n	Duration (s)	F0 (ST)	Pitch range (Hz)	F1 (Hz)
Speech	29	0.490 (0.388)	−1.21	70	590
Event	118	0.519 (0.298)	+3.69	166	649
Speech DE	10	0.651 (0.253)	−1.34	71	644

Note: On per-speaker means (four speakers contributing both conditions), event-following grunts are significantly higher in F0 ($t(3) = 3.77$, $p = .033$) and wider in pitch range ($t(3) = 2.71$, $p = .073$). Duration, F1, and the speech-describing-event category do not differ reliably.

This pattern is clearest within a single token. *Oh* is the only token with sufficient productions in both conditions (speech $n = 15$, event $n = 41$). For *oh*, event-following productions are substantially higher in pitch (F0: $t = 2.75$, $p = .009$, $d = 0.65$) and wider in range ($t = 4.68$, $p < .001$, $d = 0.93$), with no difference in duration. Because token identity is constant, this provides some within-token evidence for prosodic modulation by antecedent, which the smaller Weights Task and Glider corpora were underpowered to detect.

6.4 Grunt Token Entrainment in Portal

A natural question for a repeated-interaction corpus is whether the two members of a dyad *entrain* on grunt-token choice. The Portal corpus is well suited to this question. Each dyad plays together for an extended session, and the released transcripts provide grunt token productions for all eighteen dyads. This analysis uses the full corpus rather than the hand-annotated subset, since entrainment requires only token, speaker, and timestamp of the speech.

Grunt tokens were extracted from the player transcripts using a lexicon observed from the corpus. To avoid measuring similarity that might be driven by common filled pauses, i.e., *uh* and *um*, these tokens were excluded. The analysis is restricted to the contentful grunts where convention formation would be expected to occur. Entrainment was assessed three ways.

Distributional similarity. For each dyad, a grunt-token frequency vector was built for each player, and the cosine similarity between the two players' vectors was computed. To distinguish partner-specific similarity from similarity driven by the shared task, real-partner similarity was compared against a permutation baseline of cross-dyad (artificial) pairings. Real partners were no more similar than randomly paired non-partners (real mean cosine = 0.97, artificial pairs = 0.96). 41% of the random pairings were as similar as, or more similar than, the real partners.

Convergence over time. Each dyad's session was divided into three ordered windows across the levels, and the change in player-to-player similarity across them was estimated as a slope. There was no convergence: the mean slope was slightly negative (-0.012), and only four of the eighteen dyads showed positive trend.

Adoption of idiosyncratic tokens. Convention formation is also possibly visible in rare or novel token use, as in the lexical "catch" example documented in the Portal paper Tomlin et al. 2025. Restricting to grunts used by at most three dyads, only one such token was shared by both players of any dyad. Shared expressive tokens such as *whee* observed

multiple times during annotation of Session 12 proved to be distributed across many dyads rather than partnership-specific, indicating common reactions to similar game events rather than adoption.

Across all three measures, grunt-token use showed no evidence of interpersonal entrainment. This contrasts with the lexical convention formation that the same players exhibit for content words. This can be interpreted as a lack of alignment on grunts, even as participants align on content words. This fits the account in Chapters 4 and 5: grunts track the source of the contribution being grounded (speech or observable event).

6.5 Conclusion

This chapter tested whether the distributional and prosodic findings of the preceding chapters hold beyond the WTD and Glider corpora, using the Portal Dialogue Corpus as a structurally distinct third setting. It is dyadic like Glider, virtual rather than physically co-situated, and far more event-dense than either the Weights Task or Glider. The central phenomenon generalizes, but its surface realization is shaped by the task.

The distributional findings replicate and, in one respect, strengthen. Grunts ground observable events at a higher rate in Portal than in either earlier corpus. Grunts occur roughly four times more often per minute, and events are the dominant antecedent (40.3%) rather than at parity with speech. The rate of event-grounding scales with the event density of the task, while events remain a major groundable category throughout. This is positive evidence that observable events function as contributions requiring acknowledgment rather than as an artifact of any one corpus scenario.

Token specialization replicates only in part. The event-reactive preference of *oh*, *ooh*, and *hm* holds across all three corpora, but the speech-continuer role of *mm-hm* does not replicate in Portal, where there is little sustained verbal exchange to be acknowledged. The specialization is thus better understood as tracking the source of the grounded contribution. Where speech is sparse, the speech-grounding tokens have few occasions to appear.

The prosodic finding replicates on a different acoustic dimension. Where the Weights Task marked event-following grunts through duration, Portal marks them through pitch: event-following ACCEPTs are higher and wider in pitch, and, for *oh*, holds within a single token. This evidence for prosodic modulation is a new finding that the smaller corpora were underpowered to detect, though it rests on four speakers and is reported as exploratory. The specific means differs across corpora, even as event-following grunts are prosodically marked in each, reinforces the claim that grounding source can be encoded in acoustic form rather than in any single fixed feature.

Finally, grunt-token use showed no evidence of interpersonal entrainment on any of three measures, even though the same players form lexical conventions for content words. Grunt selection appears to be governed by what is being grounded rather than by alignment to a partner. This is consistent with the information source-tracking account, for which entrainment could have been a competing explanation.

Taken together, the Portal results support a unified conclusion. The grounding functions of grunts established in earlier chapters are present across different collaborative settings. Their surface form (tokens, acoustic feature, frequency) is sensitive to the structure of the task.

Only about half of Portal grunts constitute complete conversational turns, against roughly 80% in the other corpora, with the remainder embedded in larger turns. Where grunts fall within a turn, and what that position reveals about their grounding role, is the subject of Chapter 7.

7 GRUNT TURN POSITION

The distributional and acoustic analyses in the preceding chapters characterize grunts at the token, functional, and acoustic levels. This extension examines whether a grunt stands alone as a complete turn, prefaces a longer verbal contribution, sits embedded within a turn, or ends an utterance. Each of these positions suggests a different grounding role.

Turn-position annotation was conducted manually in ELAN across all nine WTD videos and the Glider video, using the grunt annotations from Chapter 4 as the input. Manual annotation was chosen over an automated approach because grunt tokens in naturalistic speech are not reliably detectable by ASR systems, as discussed in Chapter 1.

7.1 Turn Position Typology and Annotation

The annotation uses a four-way positional typology:

standalone The grunt constitutes the entire conversational turn. A: *put green on the scale* / B: *mm-hm*

turn-initial The grunt precedes verbal content in the same turn. *oh, that one's heavier*

turn-medial The grunt is embedded within a turn, with other speech both before and after. *the red one is, uh, ten grams*

turn-final The grunt follows verbal content in the same turn. *so it balances, huh*

A standalone grunt is a complete move (e.g., acceptance, doubt, observation) with no surrounding speech. This is the minimal case in the prosodic analysis in the previous chapters. A turn-initial grunt precedes additional verbal content in the same turn, typically registering a contribution before the speaker fills in additional detail. A turn-medial grunt is embedded in a speaker's turn, often serving as a filled pause or internal acknowledgment. A turn-final grunt follows verbal content in the same turn, often signaling that the topic can advance. These positions may predict the different distributions of dialogue act, antecedent type, and token selection.

Approximately 80% of non-filled-pause grunts in the WTD and Glider corpus are STANDALONE, as reported in Chapter 4; compared to 50% of Portal. The other three categories are drawn primarily from the remaining grunts, which means the counts for non-standalone positions are smaller but analyzable.

Table 7.1 gives the overall position distribution in each corpus. The corpora differ markedly: Weights Task grunts are most often standalone (52%), Portal grunts are most often turn-initial (48%), with Glider between (34%). This aligns with the observation that Portal grunts far less frequently constitute complete turns than those of the other corpora.

Table 7.1: Turn position distribution by corpus (% of grunts).

Corpus	n	Standalone (%)	Initial (%)	Middle (%)	Final (%)
Weights Task	367	52	31	13	4
Glider	130	30	34	28	8
Portal	680	32	48	15	5

7.2 Position by Antecedent

If turn position encodes grounding source like token and prosody do, event- and speech-following grunts would differ in where they occur in a turn. Table 7.2 shows the position distribution of grunts in each corpus.

Table 7.2: Turn position by grounding source, per corpus. Each row gives the percentage of that corpus's speech- or event-following grunts in each position. Within the Weights Task, event-following grunts are turn-initial more than twice as often as speech-following grunts; in Portal the contrast is absent.

Corpus	Antecedent	n	Standalone (%)	Initial (%)	Middle (%)	Final (%)
Weights Task	Speech	106	82	13	1	4
	Event	110	61	32	6	1
Glider	Speech	22	50	32	5	14
	Event	18	28	44	28	0
Portal	Speech	88	48	51	1	0
	Event	274	47	43	7	3

In the Weights Task the two diverge clearly. Speech-following grunts are overwhelmingly

standalone (82%) and rarely turn-initial (13%): an acknowledgment of something a partner said tends to stand on its own as a complete contribution. Event-following grunts are standalone less often (61%) and turn-initial more than twice as often (32%): a grunt that registers an observed event tends to open a turn that continues into further talk.

Glider shows the same direction. Event-following grunts are turn-initial more often than speech-following ones (44% vs. 32%), and speech-following grunts are standalone more often (50% vs. 28%). With only forty classified grunts, however, the association does not reach significance, though it is consistent with the Weights Task.

Portal departs from this pattern. Grunts there are turn-initial at similar rates whether they follow speech (51%) or an event (43%), and standalone at similar rates as well (48% vs. 47%). Position does not distinguish the two grounding sources in the same way. Part of this is a matter of token composition. Portal ACCEPTs are dominated by *oh*, which makes up roughly half of them, and *oh* is turn-initial in about two-thirds of its uses regardless of what it follows (Table 7.4). When a single turn-initial token dominates the inventory, the contrast is absorbed. The rapid, event-dense character of the Portal interaction could favor grunts that open turns, since players are reacting continuously to new changes in the environment while acting in parallel.

Similar to the restricted analyses in previous chapters, the trend holds when limited to ACCEPTs as shown in Table 7.3.

In the Weights Task the two grounding sources diverge clearly. Speech-following ACCEPTs are overwhelmingly standalone (84%) and rarely turn-initial (12%): an acknowledgment of something a partner said tends to stand on its own as a complete contribution. Event-following ACCEPTs are standalone less often (58%) and turn-initial three times as often (37%): a grunt that registers an observed event tends to open a turn that continues into further talk. Glider shows the same direction. Portal departs from the pattern: ACCEPTs are turn-initial at similar rates whether they follow speech (44%) or an event (43%)

The Weights Task and Glider show that turn position can encode grounding source.

Table 7.3: Turn position of ACCEPT grunts by grounding source, per corpus. Within the Weights Task, event-following ACCEPTs are turn-initial about three times as often as speech-following ones; in Portal position does not distinguish the two.

Corpus	Antecedent	n	Position (%)			
			Standalone (%)	Initial (%)	Middle (%)	Final (%)
Weights Task	Speech	74	84	12	1	3
	Event	65	58	37	3	2
Glider	Speech	16	62	31	6	0
	Event	10	30	60	10	0
Portal	Speech	52	56	44	0	0
	Event	242	48	43	7	2

For these situations, event-following grunts open turns, speech-following grunts complete them. Similar to the prosodic findings of Chapter 6, this positional contrast may be task-sensitive rather than fixed. The effect is flattened in the faster-paced, event-saturated Portal interaction. A fixed antecedent-to-position mapping does not generalize, but is a resource available for grounding.

7.3 Position by Token

Position is strongly associated with token identity, and the association recovers the same functional grouping found in the distributional and prosodic analyses (Table 7.4). Some patterns are clear. The acknowledgment tokens *mm-hm*, *mm*, and *hm* are overwhelmingly *standalone*. For example, *mm-hm* constitutes a complete turn in 97% of its occurrences across the three corpora. These tokens ground a contribution and yield the floor. The reactive tokens *oh* and *ah*, by contrast, are predominantly *turn-initial* (65% and 67%). They open a turn, registering and grounding something and leading to further talk. Finally, the filled pauses *uh* and *um* are the only tokens that frequently occur in the middle of turns, which is consistent with their function as planning moves rather than grounding acts. Turn-final grunts were rare across all three. Observationally, they typically involve speakers trailing off incomplete thoughts, or appending polar questions.

(Chapter 4).

Table 7.4: Turn position by token (percentage of each token's occurrences, pooled across corpora).

Token	n	Standalone (%)	Initial (%)	Middle (%)	Final (%)
mm-hm	74	97	3	0	0
mm	43	81	16	0	2
hm	50	70	30	0	0
oh	357	22	65	12	1
ah	43	23	67	7	2
ooh	36	50	39	6	6
uh	240	14	38	37	12
um	98	24	34	35	7

7.4 Conclusion

This chapter examined turn position as a third channel through which grunts encode grounding role, alongside token selection and acoustic form. Three findings emerge.

Position is tied to token identity, and the mapping aligns with the groupings in earlier chapters. Acknowledgment tokens (*mm-hm*, *mm*, *hm*) are almost always standalone complete turns. Reactive tokens (*oh*, *ah*) overwhelmingly open turns; and the filled pauses (*uh*, *um*) are alone in occurring turn-medially, consistent with their treatment as planning markers rather than grounding acts.

In the Weights Task, position encodes the grounding source. This also holds in Glider with the smaller sample size. Event-following grunts open turns much more often than speech-following ones. The same result holds when restricted to ACCEPTs. Position is additional evidence that grounding source can be marked in grunt form.

The position contrast seems to be task- or environment- sensitive, not fixed. In dyads I annotated in Portal, the grunt inventory is dominated by the turn-initial *oh* token. Grunts open turns regardless of speech or event gruntable antecedent. This is similar to the prosodic findings of Chapter 6. There is not a generalizable mapping from antecedent to position, but grunt position can be a resource speakers recruit for grounding work, varying to the setup

of the interaction.

Combined with the distributional and acoustic results, the turn position analysis shows a consistent picture. The gruntable source is inferrable from the grunt token choice, how the grunt is expressed, and where it might fall in the turn. None of these dimensions is singular, but across them and across the corpora, the distinction between grounding speech and grounding events is encoded in the form of the minimal vocalizations.

8 CONCLUSIONS AND FUTURE DIRECTIONS

8.1 Summary of the Dissertation

This dissertation has argued that non-lexical communicative acts are first-class contributions to the common ground in multimodal dialogue. The minimal vocalizations that follow spoken assertions or completed actions are not noise to be filtered by automatic systems. They are not only continuers holding the conversation turn. These grunts are grounding acts, performing communicative work on par with explicit verbal acknowledgments, and doing so very frequently and efficiency in co-situated interactions.

The argument proceeded sequentially, starting with extensible formal frameworks for non-verbal modalities. Gesture AMR established the premise that non-verbal communicative modalities have representable propositional content. The formal tools designed to represent speech can be extended to other modalities. The scaffolding chapter put together an account of grounding that the subsequent grunt analysis depended on. The Weights Task Dataset corpus was designed to elicit genuine multimodal communication, and the action annotation approach put physical actions into the same representational framework as speech and gesture. The Common Ground Tracking work, along with the supporting theory papers, developed the QBank/EBank/FBank structure. The work on MM-AMR and MM-MS-AMR, as well as the information state work, showed how speech, action, and gesture can be integrated in a common account. With this work as a backdrop, the grunt analysis made its central claim, i.e. grunts respond to speech and observable events at nearly equal rates. The EB-DEL model formalizes how a grunt can promote evidence to accepted fact in the common ground. Observable events function as propositional contributions on par with utterances. My annotation scheme and distributional analysis show that grounding source is encoded by token choice. The prosodic analysis showed that grounding source is also encoded acoustically, and the turn-position analysis showed that the gruntable antecedent is encoded structurally, by where a grunt falls within a turn. The Portal Dialogue Corpus demonstrated that these findings hold beyond the Weights Task and the Glider video, in a

very different environment. The central phenomenon, that people use grunts differently to respond to speech and events, generalizes across corpora while its surface realization can be shaped by the task.

It establishes the foundation through four interconnected contributions. Gesture AMR demonstrates that non-lexical acts have representable propositional content and can be brought within the scope of formal meaning representation. The Weights Task Dataset, Common Ground Tracking framework, and action annotation model provide the empirical and theoretical scaffolding for tracking multimodal grounding over the course of an interaction. The grunt annotation scheme and EB-DEL formal model characterize grunts as fact-promoting moves in the common ground, with distributional findings establishing that they respond to speech and observable events at comparable rates. The prosodic analysis shows that token selection and acoustic realization jointly encode grounding source.

8.2 How Grounding Source is Encoded

The empirical chapters in the dissertation approached grunts from three perspectives, and analyzed them separately: token identity, acoustic realization, and position within a turn. Taken together, these perspectives tell a single story. Across all three dimensions, the form of a grunt consistently encodes its grounding source, whether the grunt responds to something a co-interlocutor said, or to something observed in the shared environment. This unified the distributional, prosodic, and structural results, and is my central claim. The speech-event contrast is not carried by any single feature of a grunt, but is distributed redundantly across token, acoustics, and position.

Consider how each channel carries the contrast. At the level of grunt token selection, the corpora show a clear functional split. Acknowledging tokens, *mm-hm*, *mm*, and less often *hm*, lean toward speech antecedents, consistent with their long-standing description as continuers and backchannels that pass the floor back to a speaker. Reactive tokens, *oh*, *ooh*, and *hm*, lean toward observable events, consistent with the change-of-state account

of *oh* in the conversation-analytic literature. *Uh* and *um* showed up most consistently as filled pauses and planning signals, consistent with previous work. A test over the most frequent tokens showed that token choice varies systematically by the antecedent gruntable type. Token selection provides a coarse, categorical signal about what is being grounded. Hearing *mm-hm* is good evidence that someone else has spoken, and hearing *oh* or *ooh* is good evidence that something has happened external to the speakers. At the level of acoustic form, the same contrast appears. Event-following grunts are, at the population level, longer in duration, higher and wider in pitch, and higher in F1 than speech-following grunts. However, once grunt token is entered into the model, the duration and F1 effects in the Weights Task is reduced, indicating that much of the population-level acoustic difference follows from the inherent acoustics of the tokens speakers select rather than from independent prosodic modulation. The best evidence that prosody carries antecedent information within a fixed token comes from Portal, where *oh* (which is the only token with enough post-event and post-speech examples) is reliably higher and wider in pitch when it follows an event than when it follows speech. This is a genuine finding for in-token modulation, but it rests on a small sample of 4 speakers. At the level of turn position, the contrast is encoded a third way. In the Weights Task, speech-following grunts are overwhelmingly standalone complete turns, while event-following grunts are disproportionately turn-initial; a grunt that registers an observed event tends to open a turn that continues into further talk, whereas a grunt that acknowledges a partner's utterance tends to stand alone as a complete contribution. The effect holds in the same direction in Glider and persists when the analysis is restricted to ACCEPT moves.

These three analyses, conducted on different data sets, come to the same grouping. The tokens that prefer events are the tokens that open turns and the tokens with the more open, marked acoustic profiles; the tokens that prefer speech are the tokens that stand alone and the tokens with the reduced, nasal profiles. Grounding source is encoded redundantly, across token, acoustic form, and turn position, with no single channel as decisive. Token selection

is the most reliable and does the most work, acoustic form and position add further marking, and the three overlap rather than fully partition the signal.

From the speaker's perspective, this means that token, acoustics, and turn position constitute a repertoire of resources available for marking grounding source. The predominating resource is conditioned by what the task and situation scenario affords its participants. Utilizing the idea of affordances from Gibson (Gibson 1977), a collaborative task affords its participants particular communicative resources to solve the challenge. Participants may use sustained verbal exchange, gestures and physical movement, the stream of observable events, as well as opportunities to take or yield the floor. The grounding work of grunts is distributed across the available channels accordingly. I am not claiming that speakers deliberately select among channels. In the corpora examined for the dissertation, the cross-corpus shifts are distributional, and most can be explained by what each task affords rather than by any strategic choice by the participants. Where an interaction supplies sustained verbal exchange, the speech-grounding tokens and standalone turn position have ample occasion to do the work; in the Weights Task, where participants reason about block weights, *mm-hm* predominantly follows speech, speech-following grunts are standalone over eighty percent of the time, and speech and observable events serve as antecedents at near-equal rates. The same speech-side pattern holds in Glider, a two-person furniture-assembly task in which long stretches of quiet work are punctuated by talk: *mm-hm* and *uh-huh* appear as speech-following acknowledgments, and speech-following grunts again tend to stand alone, even though the more meticulous, sequential task yields fewer groundable events overall. In the Portal corpus, the interaction instead supplies a near-continuous stream of observable events and comparatively little talk. Here, *mm-hm*'s speech preference disappears, because there is less conversational speech for it to acknowledge. The reactive *oh* carries the distinction in Portal. It accounts for nearly half of ACCEPTs and has higher and wider pitch in my sample. Across the three settings, speakers draw on the same resources; which ones do the grounding work depends on the task.

The perspectives of the previous two paragraphs are two views of the same analysis. From the outside, grounding source is redundantly marked; from the speaker's side, the repertoire of channels is utilized according to the interactional setup. Because the grounding source is marked in more than one place, the speech-event distinction stays recoverable when one channel carries less information. In Portal, where there's little speech to acknowledge, mm-hm's speech preference flattens and position stops separating the sources, but the pitch of *oh* can still differentiate its source. No corpus in the dissertation rests the distinction on a single feature, so none leaves it unmarked. The corpus can't quite make this redundancy claim on its own. It would take a perceptual or interactional study, maybe by artificially manipulating pitch or duration of tokens, to demonstrate that speakers and listeners actually use this redundancy. I'll leave that as a future possibility.

8.3 Caveats and Limitations

The findings in this dissertation are subject to several caveats and limitations, largely due to the nature of the data and the difficulties of the annotation task. Each of the listed limitations is also an opportunity for more esting of the ideas and account of the dissertation.

The most consequential limitation is the sample size of the annotated data, from which I base several of the acoustic takeaway claims. The data set is substantial in grunt tokens, with approximately 1000 annotated across WTD, Glider, and Portal. But there are relatively few speakers. The Weights Task portion has fewer than thirty unique speakers, and Glider has only two. Portal, where I demonstrate the in-token modulation for *oh*, has just four, across two pairs. The *oh* finding is the cleanest demonstration that grounding source differs prosodically within a fixed token, so it is vulnerable to the small speaker sample.

The audio recording quality constrains what could be measured and where. The Weights Task was captured on a single room microphone shared by all speakers and is not diarized. There is overlapping speech, and varying distance to the microphone. The Glider video was recorded on a camcorder and later digitized. Glider's audio yielded unreliable formant and

pitch estimates and was therefore excluded from the prosodic analysis entirely, retained only for the categorical token, dialogue-act, and antecedent results. The shared-microphone setup also placed the theoretically motivated feature of intensity out of reach. Ward's work ties it to confidence and emphasis, but intensity cannot be reliably be recovered when mic-to-speaker distance varies so much, so loudness plays no role in the acoustic account despite its possible relevance. Portal's per-speaker audio is captured with individual microphones per speaker, which is part of why it could supply the within-token test, but it introduces its own uncertainty in the form of possible automatic gain control in the voice-chat pipeline, which would compress exactly the intensity differences that might otherwise be measured.

The gruntable annotation reliability poses another limitation. Identifying a grunt is moderately reliable, with agreement shown to be on par with comparable tasks. But identifying what a grunt responded to proved to be substantially more difficult. The retrospective, causal judgment of which prior utterance or event triggered a grunt is genuinely difficult, since multiple plausible antecedents often fall within a short window. Combining gruntable and grunt annotation in a single annotation pass increased the difficulty of the task. Annotations were subject to cascading disagreements; meaning if annotators disagreed on a grunt annotation, the dependent gruntable annotations would necessarily be misaligned. Reframing the task by giving annotators the gold-standard grunts and asking only for the antecedent gruntable showed reliable agreement. Even with the reframing, the annotations are more secure at the coarse speech/event level than in the finer subcategories. The speech-describing-event category in particular remained the area of most disagreement, likely due to a lack of refinement in the guidelines for ambiguous cases.

Also in the realm of annotation, the dissertation largely relies on my single-annotator gold standard throughout. The subset annotated for agreement is small, and the annotators did not have formal training in the subject matter. The initial grunt agreement is promising, but over-reliance on a single annotator's judgment is a risk.

One additional limitation is narrower. The different setup of the Portal data required an

annotation judgment call. Because the players move continuously through a virtual world in which new rooms and objects are constantly appearing, the boundary between OBSERVE and accept was blurred. I annotated standalone *oh* primarily as ACCEPT on the rationale that the newly visible environment element constitutes an event proposition the player takes up, similar to the event-to-acceptance structure used for the Weights Task, even where no prior question was under discussion.

8.4 Future Directions

Several additional threads emerge from the work presented here that are beyond the scope of the dissertation but represent natural extensions of the work's contributions. There are some broad themes: integrating grunts into computational models of grounding, recovering and recognizing grunts from the speech signal, experimental testing of the encoding claims, and extending the theoretical account beyond grunts.

Grunts in computational models The Common Ground Tracking pipeline of Chapter 3 did not consider minimal vocalizations. Its move classifier operates over utterance-derived segments, and because the ASR systems used to produce those segments filter exactly the tokens studied here, grunts that constitute complete turns are largely absent from the input. A first step is to measure that gap directly: how many grounding-relevant grunts, particularly the standalone ACCEPT and DOUBT moves that promote and demote propositions in the common ground. fall outside the segments the tracker operates on, and are therefore invisible to it by construction.

Recovering grunts from the speech signal The filtering problem motivates a dedicated grunt-detection system. I have carried out some preliminary experimentation on a Wav2Vec2-based (Baevski et al. 2020) classifier trained to identify the difference between output from a voice activity detection algorithm (VAD) (Team 2024) and ASR using faster-whisper [1]. These experiments use podcast audio, providing groundwork for a more robust

[1] `https://github.com/SYSTRAN/faster-whisper`

grunt detection pipeline that could be integrated with systems to recover the tokens they currently filter. These regions that have vocal activity but no recognized speech are candidate grunts, but in early results, it's common to identify music and laughter alongside grunts.

A more robust detector, integrated with downstream dialogue systems, would let those systems recover the tokens they currently discard. A natural companion is a systematic benchmark of how leading ASR systems such as Parakeet, Qwen, and DeepSpeech handle non-lexical vocalizations (Kuchaiev et al. 2019; Chu et al. 2023; Hannun et al. 2014). Beyond the manual annotation approach used here, simulation-based validation (Tam et al. 2023b) offers another path forward for verifying multimodal annotation schemes. Quantifying what fraction of grunts each system captures would inform both system selection for multimodal dialogue work and the design of grunt-aware recognition, and with luck turn up another disfluency-removal policy as quotable as the one in the Whisper paper.

Testing the encoding claim experimentally The central claim of this dissertation, that the form of a grunt encodes its grounding source, has been established observationally from the distributions and measurements from the three corpora. An experimental test is a natural next step. A perceptual study could ask whether a listener, given a single grunt in isolation, can recover what it was responding to, and whether artificially manipulating pitch or duration shifts that judgment. This would establish whether the redundancy described earlier in this chapter is something speakers and listeners actually exploit. The three encoding channels could also be modeled jointly: a single classifier predicting grounding source from token, acoustic form, and turn position together would quantify how much each channel contributes independently of the others, which the separate analyses here cannot. The denser sampling that such a model would need is the kind of data an expanded Portal annotation could supply. On a related note, the token-entrainment analysis of Chapter 6 could be extended to the acoustics. It would be interesting to see whether dyads converge on grunt prosody over an interaction, as they do not on grunt tokens. The paired structure

of Portal is suited to answer with per-speaker audio across more dyads than were annotated for this dissertation.

Extending the account The grunts studied here sit between non-lexical sounds and conventional words, and the question of how they move along that continuum is open. *Wow* seems to have made the journey to full word: it can be used as a noun ("that was a wow"), verb ("wowed the committee") and is drifting towards adjectival use ("wow moment"). Perhaps one day we will be *uh-huh*-ing and *uh-uh*-ing each other (once we work out the spellings). Meta-communicative vocalizations observed in all corpora (e.g., whistling, humming, singing during problem-solving) appear to mark discourse segment transitions and task-completion boundaries. These fall outside the grunt category as defined in this work but are plausibly part of the same broader phenomenon: non-lexical vocalizations that communicate structural information about the interaction (Keevallik and Ogden 2020). Finally, the relationship between grunts and co-occurring gesture is unexamined here: whether speakers grunt differently when they also gesture, and whether gesture and grunt trade off in carrying grounding work, is a question the multimodal corpora are positioned to address, in tandem with GAMR annotation.

8.5 Conclusion

This work had argued that non-lexical communicative acts are first-class contributions to the common ground in multimodal dialogue. It establishes the foundation through four interconnected contributions. Gesture AMR demonstrated that non-lexical acts have representable propositional content and can be brought within the scope of formal meaning representation. The Weights Task Dataset, Common Ground Tracking framework, and action annotation model provided the empirical and theoretical scaffolding for tracking multimodal grounding over the course of an interaction. The grunt annotation scheme and EB-DEL formal model characterize grunts as fact-promoting moves in the common ground, with distributional findings establishing that they respond to speech and observable events at comparable rates. The

prosodic and turn-position analyses showed that token selection, acoustic realization, and structural position jointly encode grounding source, and the Portal corpus established that these findings hold across a structurally distinct collaborative setting.

Taken together, this work makes the case that grunts are not peripheral to the grounding process but among its primary signals in co-present, task-based interaction. In situated dialogue, the minimal vocalization that follows a completed action or a spoken assertion is doing the same communicative work as an explicit verbal acknowledgment. That acknowledgment is efficient, frequent, and more systematic in how it encodes its function than existing models of dialogue have recognized. A complete account of grounding in multimodal dialogue must bring these small sounds into the formal and evidential picture.

The grunt work, it turns out, is doing quite a lot.

BIBLIOGRAPHY

Afantenos, Stergos, Nicholas Asher, Farah Benamara, Anaıs Cadilhac, Cédric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, et al. (2012). "Developing a corpus of strategic conversation in the settlers of catan". In: *Proceedings of the 1st Workshop on Games and NLP*.

Agre, Philip E. and John Batali (1989). "Transcript of the Glider Video". Unpublished transcript. Cambridge, MA.

Aikhenvald, Alexandra Y (2012). "The essence of mirativity". In: *Linguistic typology* 16.3, pp. 435–485.

Allen, James F. (1983). "Maintaining knowledge about temporal intervals". In: *Communications of the ACM* 26.11, pp. 832–843.

Anderson, Anne H, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. (1991). "The HCRC map task corpus". In: *Language and speech* 34.4, pp. 351–366.

Anikina, Tatiana, Alina Leippert, and Simon Ostermann (2025). "Building Common Ground in Dialogue: A Survey". In: *Proceedings of the 2nd LUHME Workshop*, pp. 3–28.

Asher, Nicholas and Anthony Gillies (2003). "Common ground, corrections, and coordination". In: *Argumentation* 17, pp. 481–512.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33, pp. 12449–12460.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (Aug. 2013). "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: https://aclanthology.org/W13-2322.

Benthem, Johan van, David Fernández-Duque, and Eric Pacuit (2014). "Evidence and plausibility in neighborhood structures". In: *Annals of Pure and Applied Logic* 165.1, pp. 106–133.

Boersma, Paul and David Weenink (2021). *Praat: doing phonetics by computer [Computer program]*. Version 6.1.38, retrieved from https://www.praat.org/. https://www.praat.org/.

Bonial, Claire, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss (May 2020). "Dialogue-AMR: Abstract Meaning Representation for Dialogue". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 684–695. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.86/.

Bonn, Julia, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner (May 2020). "Spatial AMR: Expanded Spatial Annotation in the Context of a Grounded Minecraft Corpus". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Marseille, France: European Language Resources Association, pp. 4883–4892. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.601.

Bortfeld, Heather, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan (2001). "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender". In: *Language and speech* 44.2, pp. 123–147.

Boudin, Auriane, Stéphane Rauzy, Roxane Bertrand, Magalie Ochs, and Philippe Blache (2024). "The distracted ear: how listeners shape conversational dynamics". In: *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pp. 15872–15887.

Brennan, Susan E and Joy E Hanna (2009). "Partner-specific adaptation in dialog". In: *Topics in Cognitive Science* 1.2, pp. 274–291.

Brugman, Hennie and Albert Russel (May 2004). "Annotating Multi-media/Multi-modal Resources with ELAN". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

Brutti, Richard, Lucia Donatelli, Kenneth Lai, and James Pustejovsky (June 2022). "Abstract Meaning Representation for Gesture". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1576–1583. URL: https://aclanthology.org/2022.lrec-1.169.

Brutti, Richard and James Pustejovsky (2026a). "More Than "Oh": Grounding Observable Events with Grunts in Multimodal Dialogue". In: *Proceedings of the Fifteenth Language Resources and Evaluation Conference*.

— (2026b). "The Grunt Work of Conversational Grounding: Minimal Responses in Multimodal Dialogue". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 48.

Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic (2018). "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling". In: *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 5016–5026.

Bühler, Karl (1934). *Sprachtheorie*. Vol. 2. 2. Jena Fischer.

Cai, Shu and Kevin Knight (Aug. 2013). "Smatch: an Evaluation Metric for Semantic Feature Structures". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 748–752. URL: https://aclanthology.org/P13-2131.

Cassell, Justine, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost (2000). *Embodied conversational agents*. MIT Press.

Cavalcanti, Julio Cesar and Gabriel Skantze (2025). "" Dyadosyncrasy", Idiosyncrasy and Demographic Factors in Turn-Taking". In: *Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA.

Chen, Wei-Te and Will Styler (June 2013). "Anafora: A Web-based General Purpose Annotation Tool". In: *Proceedings of the 2013 NAACL HLT Demonstration Session*. Atlanta, Georgia: Association for Computational Linguistics, pp. 14–19. URL: https://aclanthology.org/N13-3004.

Chu, Yunfei, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou (2023). "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models". In: *arXiv preprint arXiv:2311.07919*.

Clark, Herbert H and Susan E Brennan (1991). "Grounding in communication". In: *Perspectives on socially shared cognition* 13.1991, pp. 127–149.

Clark, Herbert H and Jean Fox Tree (2002). "Using uh and um in spontaneous speaking". In: *Cognition* 84.1, pp. 73–111.

Damen, Dima, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray (2022). "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100". In: *International Journal of Computer Vision (IJCV)* 130, pp. 33–55. URL: https://doi.org/10.1007/s11263-021-01531-2.

DeLancey, Scott (1997). "Mirativity: The grammatical marking of unexpected information". In.

Devlin, Jacob, 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)*, pp. 4171–4186.

Dey, Indrani, Sadhana Puntambekar, Ruofan Li, Diane Gengler, Rachel Dickler, Leanne M Hirshfield, Charis Clevenger, Sierra Rose, Mariah Bradford, and Nikhil Krishnaswamy (2023). "The NICE framework: analyzing students' nonverbal interactions during collaborative learning". In: *Pre-conference workshop on Collaboration Analytics at 13th International Learning Analytics and Knowledge Conference (LAK 2023)*. Society for Learning Analytics Research.

Dice, Lee R (1945). "Measures of the amount of ecologic association between species". In: *Ecology* 26.3, pp. 297–302.

Donatelli, Lucia, Kenneth Lai, Richard Brutti, and James Pustejovsky (2022a). "Towards Situated AMR: Creating a Corpus of Gesture AMR". In: *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design*. Ed. by Vincent G. Duffy. Cham: Springer International Publishing, pp. 293–312. ISBN: 978-3-031-06018-2.

— (2022b). "Towards situated AMR: Creating a corpus of gesture AMR". In: *International Conference on Human-Computer Interaction*. Springer, pp. 293–312.

Ekman, Paul and Wallace V Friesen (1969). "The repertoire of nonverbal behavior: Categories, origins, usage, and coding". In: *Semiotica* 1.1, pp. 49–98.

Erard, Michael (2008). *Um...: Slips, stumbles, and verbal blunders, and what they mean*. Anchor.

Erker, Daniel and Lee-Ann Marie Vidal-Covas (2022). "What we say when we say nothing at all: Clues to contact-induced language change in Spanish conversational pause-fillers". In: *Estudios del Observatorio Cervantes en Harvard* 80, pp. 1–31.

Eyben, Florian, Martin Wöllmer, and Björn Schuller (Oct. 2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. MM '10. New York, NY, USA: Association for

Computing Machinery, pp. 1459–1462. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246. URL: https://doi.org/10.1145/1873951.1874246 (visited on 10/27/2022).
Farkas, Donka F and Kim B Bruce (2010). "On reacting to assertions and polar questions". In: *Journal of semantics* 27.1, pp. 81–118.
Foster, Mary Ellen (2007). "Enhancing human-computer interaction with embodied conversational agents". In: *International Conference on Universal Access in Human-Computer Interaction*. Springer, pp. 828–837.
Gardner, Rod (2001). "When listeners talk". In.
Gawne, Lauren (2025). *Gesture: A slim guide*. Oxford University Press.
Gibson, James J (1977). "The theory of affordances". In: *Hilldale, USA* 1.2, pp. 67–82.
Ginzburg, Jonathan (2012). *The interactive stance*. Oxford University Press.
Ginzburg, Jonathan, Chiara Mazzocconi, and Ye Tian (2020). "Laughter as language". In: *Glossa: a journal of general linguistics (2021-...)* 5.1.
Goldin-Meadow, Susan (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.
Goldman, Alvin I (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.
Goodman, Bradley A., Jill Drury, Robert D. Gaimari, Laura Kurland, and Jeffrey Zarrella (Oct. 2006). *Applying User Models to Improve Team Decision Making*. MITRE Technical Report MTR 060150. Bedford, MA: The MITRE Corporation.
Gravano, Agustín and Julia Hirschberg (2011). "Turn-taking cues in task-oriented dialogue". In: *Computer Speech & Language* 25.3, pp. 601–634.
Grice, H. Paul (1975). "Logic and Conversation". In: *The logic of grammar*. Ed. by Donald Davidson. Dickenson Pub. Co., pp. 64–75.
Grosz, Barbara J and Candace L Sidner (1986). "Attention, intentions, and the structure of discourse". In: *Computational linguistics* 12.3, pp. 175–204.
Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. (2014). "Deep speech: Scaling up end-to-end speech recognition". In: *arXiv preprint arXiv:1412.5567*.
Henderson, Matthew, Blaise Thomson, and Steve Young (2014). "Word-based dialog state tracking with recurrent neural networks". In: *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pp. 292–299.
Heritage, John (1984). "13. A change-of-state token and aspects of its sequential placement". In: *Structures of social action*. Cambridge University Press, pp. 299–345.
— (2002). "Oh-prefaced responses". In: *The language of turn and sequence*, p. 196.
Hockey, Beth Ann (1993). "Prosody and the role of okay and uh-huh in discourse". In: *Proceedings of the eastern states conference on linguistics*. Citeseer, pp. 128–136.
Hoeppner, Michael Chad (Jan. 2025). *Don't Say Um: How to Communicate Effectively to Live a Better Life*. Balance. ISBN: 978-0306834509.
Hoey, Elliott M (2014). "Sighing in interaction: Somatic, semiotic, and social". In: *Research on Language and Social Interaction* 47.2, pp. 175–200.
— (2020). "Waiting to inhale: On sniffing in conversation". In: *Research on Language and Social Interaction* 53.1, pp. 118–139.

Hunter, Julie, Nicholas Asher, and Alex Lascarides (2015). "Integrating non-linguistic events into discourse structure". In: *Proceedings of the 11th international conference on computational semantics*. Association for Computational Linguistics, pp. 184–194.

— (2018). "A formal semantics for situated conversation". In: *Semantics and Pragmatics* 11, pp. 1–52.

Jadoul, Yannick, Bill Thompson, and Bart De Boer (2018). "Introducing Parselmouth: A Python interface to Praat". In: *Journal of Phonetics* 71, pp. 1–15. DOI: 10.1016/j.wocn.2018.07.001.

Keevallik, Leelo (2014). "Turn organization and bodily-vocal demonstrations". In: *Journal of Pragmatics* 65, pp. 103–120.

Keevallik, Leelo and Richard Ogden (2020). "Sounds on the margins of language at the heart of interaction". In: *Research on Language and Social Interaction* 53.1, pp. 1–18.

Kendon, Adam (1997). "Gesture". In: *Annual review of anthropology* 26.1, pp. 109–128.

— (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Khebour, Ibrahim, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. (2024). "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, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky (May 2024). "Common Ground Tracking in Multimodal Dialogue". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 3587–3602. URL: https://aclanthology.org/2024.lrec-main.318/.

Kok, Iwan de and Dirk Heylen (2011). "The MultiLis corpus–Dealing with individual differences in nonverbal listening behavior". In: *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues: Third COST 2102 International Training School, Caserta, Italy, March 15-19, 2010, Revised Selected Papers*. Springer, pp. 362–375.

Kong, Anthony Pak-Hin, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam (2015). "A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a Database of Speech and GEsture (DoSaGE)". In: *Journal of nonverbal behavior* 39.1, pp. 93–111.

Kontogiorgos, Dimosthenis, Elena Sibirtseva, and Joakim Gustafson (2020). "Chinese whispers: A multimodal dataset for embodied language grounding". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 743–749.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson (2006). "Towards a Common Framework for Multimodal Generation: The Behavior Markup Language". In: *Intelligent Virtual Agents*. Ed. by Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 205–217. ISBN: 978-3-540-37594-4.

Kopp, Stefan and Ipke Wachsmuth (2010). *Gesture in embodied communication and human-computer interaction*. Vol. 5934. Springer.

Kraaij, Wessel, Thomas Hain, Mike Lincoln, and Wilfried Post (2005). "The AMI meeting corpus". In: *Proc. International Conference on Methods and Techniques in Behavioral Research*, pp. 1–4.

Kraus, Kelsey (2019). "Intonation and expectation: English mirative contours and particles". In: *Proceedings of Sinn und Bedeutung*. Vol. 23. 2, pp. 19–36.

Krippendorff, Klaus (2011). "Computing Krippendorff's alpha-reliability". In: *Computing* 1, pp. 25–2011.

Krishnaswamy, Nikhil and James Pustejovsky (2019). "Generating a Novel Dataset of Multimodal Referring Expressions". In: *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pp. 44–51.

Kuchaiev, Oleksii, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. (2019). "Nemo: a toolkit for building ai applications using neural modules". In: *arXiv preprint arXiv:1909.09577*.

Lai, Kenneth, Richard Brutti, Lucia Donatelli, and James Pustejovsky (Sept. 2021). "Situated UMR for multimodal interactions". In: *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*. Potsdam, Germany: SEMDIAL. URL: http://semdial.org/anthology/Z21-Lai_semdial_0028.pdf.

— (May 2024). "Encoding Gesture in Multimodal Dialogue: Creating a Corpus of Multimodal AMR". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 5806–5818. URL: https://aclanthology.org/2024.lrec-main.515/.

Lai, Kenneth, Lucia Donatelli, Richard Brutti, and James Pustejovsky (2025). "A Model of Information State in Situated Multimodal Dialogue". In: *Proceedings of the 16th International Conference on Computational Semantics*, pp. 282–288.

Lascarides, Alex and Matthew Stone (2009). "A formal semantic analysis of gesture". In: *Journal of Semantics* 26.4, pp. 393–449.

Levitan, Rivka and Julia Hirschberg (2011). "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions". In: *Proc. Interspeech 2011*, pp. 3081–3084.

Lopes, José, Nils Hemmingsson, and Oliver Åstrand (2018). "The Spot the Difference corpus: a multi-modal corpus of spontaneous task oriented spoken interactions". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lotfian, Reza and Carlos Busso (2017). "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings". In: *IEEE Transactions on Affective Computing* 10.4, pp. 471–483.

Lücking, Andy (2021). "Gesture". In: *Head-Driven Phrase Structure Grammar: The handbook*. Ed. by Robert D. Borsley & Jean-Pierre Koenig Stefan Müller Anne Abeillé. Berlin: Language Science Press. Chap. 27.

Markowska, Magdalena, Adil Soubki, Gary Mar, Seyed Abolghasem Mirroshandel, Owen Rambow, and Anita Wasilewska (n.d.). "Formal Representation of Common Ground in Dialogue". In: ().

Mather, Susan M (2005). "Ethnographic research on the use of visually based regulators for teachers and interpreters". In: *Attitudes, Innuendo, and Regulators*, pp. 136–161.

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.

McNeill, David (1992). "Hand and mind". In: *Advances in Visual Semiotics* 351.

— (2008). *Gesture and thought*. University of Chicago press.

Mohapatra, Biswesh, Seemab Hassan, Laurent Romary, and Justine Cassell (May 2024). "Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 3967–3977. URL: https://aclanthology.org/2024.lrec-main.352/.

Mughal, Muhammad Hamza, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, Varsha Suresh, Vera Demberg, and Christian Theobalt (2024). *DnD Group Gesture Dataset*. Version V4. Edmond. DOI: 10.17617/3.IPFYCC. URL: https://doi.org/10.17617/3.IPFYCC.

O'Connell, Daniel C and Sabine Kowal (2012). *Dialogical genres: Empractical and conversational listening and speaking*. Springer Science & Business Media.

O'Gorman, Tim, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer (Aug. 2018). "AMR Beyond the Sentence: the Multi-sentence AMR corpus". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3693–3702.

Ogden, Richard (2013). "Clicks and percussives in English conversation". In: *Journal of the International Phonetic Association* 43.3, pp. 299–320.

Opitz, Juri (May 2023). "SMATCH++: Standardized and Extended Evaluation of Semantic Graphs". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1595–1607. DOI: 10.18653/v1/2023.findings-eacl.118. URL: https://aclanthology.org/2023.findings-eacl.118.

Pacuit, Eric (2017). *Neighborhood semantics for modal logic*. Springer.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The proposition bank: An annotated corpus of semantic roles". In: *Computational linguistics* 31.1, pp. 71–106.

Patel-Grosz, Pritty, Salvador Mascarenhas, Emmanuel Chemla, and Philippe Schlenker (2023). "Super linguistics: An introduction". In: *Linguistics and Philosophy* 46.4, pp. 627–692.

Portner, Paul (2004). "The semantics of imperatives within a theory of clause types". In: *Semantics and Linguistic Theory*, pp. 235–252.

Pustejovsky, James and Nikhil Krishnaswamy (2021). "Embodied human computer interaction". In: *KI-Künstliche Intelligenz* 35.3, pp. 307–327.

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary (May 2010). "ISO-TimeML: An International Standard for Semantic Annotation". In: *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/55_Paper.pdf.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2023). "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR, pp. 28492–28518.

Radvansky, Gabriel A and Jeffrey M Zacks (2014). *Event cognition*. Oxford University Press.

Rohrer, Patrick Louis, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve-Gibert, Ada Ren, Stefanie Shattuck-Hufnagel, and Pilar Prieto (2020). "The MultiModal MultiDimensional (M3D) labelling scheme for the annotation of audiovisual corpora". In: *submitted at this conference*.

Sacks, Harvey, Emanuel A Schegloff, and Gail Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation". In: *Language* 50.4 Part 1, pp. 696–735.

Schegloff, Emanuel A (1982). "Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences". In: *Analyzing discourse: Text and talk* 71.93.

Scherer, Klaus R (2003). "Vocal communication of emotion: A review of research paradigms". In: *Speech communication* 40.1-2, pp. 227–256.

Schiffrin, Deborah (1987). *Discourse markers*. 5. Cambridge University Press.

Schlenker, Philippe (2018). "Gesture projection and cosuppositions". In: *Linguistics and Philosophy* 41.3, pp. 295–365.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.

Sidner, Candace L (2025). "Glider Video".

Sorensen, Thorvald (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: *Biologiske skrifter* 5, pp. 1–34.

Stalnaker, Robert (2002). "Common ground". In: *Linguistics and Philosophy* 25.5-6, pp. 701–721.

Stewart, Angela E. B., Zachary Keirn, and Sidney K. D'Mello (Sept. 2021). "Multimodal modeling of collaborative problem-solving facets in triads". en. In: *User Modeling and User-Adapted Interaction* 31.4, pp. 713–751. ISSN: 1573-1391. DOI: 10.1007/s11257-021-09290-y. URL: https://doi.org/10.1007/s11257-021-09290-y (visited on 06/02/2022).

Stone, Matthew and Una Stojnic (2015). "Meaning and demonstration". In: *Review of Philosophy and Psychology* 6.1, pp. 69–97.

Sun, Chen, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello (2020). "Towards a generalized competency model of collaborative problem solving". In: *Computers & Education* 143, p. 103672.

Tam, Christopher, Richard Brutti, Kenneth Lai, and James Pustejovsky (June 2023a). "Annotating Situated Actions in Dialogue". In: *Proceedings of the Fourth International Workshop on Designing Meaning Representations*. Ed. by Julia Bonn and Nianwen Xue. Nancy, France: Association for Computational Linguistics, pp. 45–51. URL: https://aclanthology.org/2023.dmr-1.5/.

Tam, Christopher, Richard Brutti, Kenneth Lai, and James Pustejovsky (2023b). "Evaluating Multimodal Behavior Schemas with VoxWorld". In: *International Conference on Human-Computer Interaction*. Springer, pp. 578–591.

Team, Silero (2024). *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*.

Tian, Ye, Tim Nieradzik, Sepehr Jalali, and Da-shan Shiu (July 2021). "How does BERT process disfluency?" In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li. Singapore and Online: Association for Computational Linguistics, pp. 208–217. DOI: 10.18653/v1/2021.sigdial-1.22. URL: https://aclanthology.org/2021.sigdial-1.22/.

Tomlin, Nicholas, Naitian Zhou, Eve Fleisig, Liangyuan Chen, Téa Wright, Lauren Vinh, Laura X Ma, Seun Eisape, Ellie French, Tingting Du, et al. (2025). "Characterizing Language Use in a Collaborative Situated Game". In: *arXiv preprint arXiv:2512.03381*.

Traum, David (1995). "A computational theory of grounding in natural language conversation". UMI Order No. GAX95-23171. PhD thesis. USA: University of Rochester.

Traum, David and Staffan Larsson (2003). "The information state approach to dialogue management". In: *Current and new directions in discourse and dialogue*. Springer, pp. 325–353.

Tu, Jingxuan, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky (2023). "Dense paraphrasing for textual enrichment". In: *Proceedings of the 15th International Conference on Computational Semantics*, pp. 39–49.

Ueda, Nobuhiro, Hideko Habe, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino (2024). "J-CRe3: A Japanese conversation dataset for real-world reference resolution". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9489–9502.

Van Gysel, Jens E. L., Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue (2021). "Designing a Uniform Meaning Representation for Natural Language Processing". In: *KI-Künstliche Intelligenz* 35.3, pp. 343–360.

Wahlster, Wolfgang (2006). "Dialogue systems go multimodal: The SmartKom experience". In: *SmartKom: foundations of multimodal dialogue systems*. Springer, pp. 3–27.

Wang, Isaac, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J. Ross Beveridge, Bruce A. Draper, and Jaime Ruiz (2017). "EGGNOG: A Continuous, Multi-modal Data Set of Naturally Occurring Gestures with Ground Truth Labels". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 414–421. DOI: 10.1109/FG.2017.145.

Ward, Nigel (2000). "The challenge of non-lexical speech sounds". In: *Proc. ICSLP 2000*, vol–2.

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

Ward, Nigel and Wataru Tsukahara (2000). "Prosodic features which cue back-channel responses in English and Japanese". In: *Journal of pragmatics* 32.8, pp. 1177–1207.

Wilkins, David P (1992). "Interjections as deictics". In: *Journal of Pragmatics* 18.2-3, pp. 119–158.

Wright, Melissa (2011). "On clicks in English talk-in-interaction". In: *Journal of the International Phonetic Association* 41.2, pp. 207–229.

Yadavalli, Aditya, Tiago Pimentel, Tamar I Regev, Ethan Wilcox, and Alex Warstadt (2025). "What Do Prosody and Text Convey? Characterizing How Meaningful Information is Distributed Across Multiple Channels". In: *arXiv preprint arXiv:2512.16832*.

Yngve, Victor H (1970). "On getting a word in edgewise". In: *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pp. 567–578.

Zarrieß, Sina, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen (2016). "Pentoref: A corpus of spoken references in task-oriented dialogues". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 125–131.

Zhang, Ruochen and Carsten Eickhoff (2021). "SOCCER: An information-sparse discourse state tracking collection in the sports commentary domain". In: *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pp. 4325–4333.

Zhang, Yuchen and Nianwen Xue (May 2018). "Structured Interpretation of Temporal Relations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Zhu, Andrew, Karmanya Aggarwal, Alexander Feng, Lara J. Martin, and Chris Callison-Burch (July 2023). "FIREBALL: A Dataset of Dungeons and Dragons Actual-Play with Structured Game State Information". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 4171–4193. DOI: 10.18653/v1/2023.acl-long.229. URL: https://aclanthology.org/2023.acl-long.229/.

Zhu, Yifan, Hannah VanderHoeven, Kenneth Lai, Mariah Bradford, Christopher Tam, Ibrahim Khebour, Richard Brutti, Nikhil Krishnaswamy, and James Pustejovsky (2024). "Modeling theory of mind in multimodal HCI". In: *International Conference on Human-Computer Interaction*. Springer, pp. 205–225.