More Than "Oh": Grounding Observable Events with Grunts in Multimodal Dialogue

Richard Brutti and James Pustejovsky
Brandeis University
Waltham, MA, USA
{brutti, jamesp}@brandeis.edu

Abstract

Conversational grunts (minimal vocalizations like *oh*, *mm-hm*, and *uh-huh*) ground information and coordinate understanding in human dialogue, yet computational systems typically treat them as noise rather than meaningful communicative acts. We present a systematic annotation and analysis of 497 grunts across 3 hours of multimodal collaborative tasks, introducing an annotation scheme that captures grunts, their antecedents, and dialogue act functions. Our analysis reveals that grunts respond to speech and observable events at nearly equal rates, demonstrating that non-verbal events function as conversational contributions requiring acknowledgment. Tokens exhibit functional specialization: *mm-hm* predominantly acknowledges speech, while *oh* preferentially acknowledges events. Prosodic analysis shows speakers systematically modulate duration and pitch based on antecedent type, with event responses typically longer and having greater range. These findings have implications for dialogue state tracking, multimodal grounding, and turn-taking in conversational AI systems.

Keywords: multimodal dialogue, backchannels, dialogue acts

1. Introduction

Modern dialogue systems track user intents and generate coherent responses, yet they fail to leverage minimal vocalizations like *oh*, *mm-hm*, and *uh-huh* that ground information and coordinate understanding. These conversational *grunts*, also called non-lexical utterances (Ward, 2006), backchannels (Yngve, 1970), and interjections (Wilkins, 1992), are often treated as noise to be filtered (Erker and Vidal-Covas, 2022), rather than meaningful communicative acts.

Consider two people cooking together: water begins to boil and the lid rattles. One person says *ooh*. The other immediately adds pasta to the boiling water. This minimal vocalization grounds the observable event and enables coordinated action without lexical confirmation, demonstrating how grunts respond to and ground events in everyday interaction.

Despite their ubiquity, grunts remain poorly represented in traditional semantic annotation schemes such as AMR (Banarescu et al., 2013) and inadequately modeled in dialogue systems. They are inherently co-referential (Cassell et al., 2000; Foster, 2007), yet retain characteristics that make them partially interpretable on their own. Our analysis of 497 grunts reveals that speakers respond to speech and observable events at nearly equal rates (25.4% vs. 25.1%), with systematic token specialization and prosodic encoding of discourse context.

Contributions This work provides:

- An annotation scheme capturing grunts, antecedents (speech and events), and dialogue act functions in situated multimodal interaction
- A formal model of events and grunts contributing to common ground in collaborative tasks
- Annotated extensions to the Weights Task Dataset (Khebour et al., 2024a,b) and analysis of the Glider video corpus
- Empirical analysis demonstrating functional specialization, prosodic encoding, and the role of grunts as complete conversational turns

2. Background / Related work

2.1. Conversational Grunts

Despite the ubiquity of conversational grunts in natural interaction, they have received limited systematic attention in computational models of dialogue.

We follow Ward (2006) in treating grunts as having sound-meaning correspondences. Beyond emblematic examples like *oh*, *um*, and *uh-huh*, our data includes diverse forms like *wa-how* and *oof*. Our focus centers on grunts that emerge as vocal productions rather than those derived from existing lexical items (e.g., *damn*, *hurray*), emphasizing the more noise-like qualities of these phenomena.

Individual grunt tokens have been studied in detail, including *oh* as a change-of-state marker (Heritage, 2002), *um* and *uh* as planning signals (Clark and Tree, 2002), and *mm* in response contexts (Gardner, 1997). However, systematic analysis of grunts across multimodal situated tasks remains limited, particularly regarding how they respond to and ground non-verbal events.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1677–1687
11–16 May 2026.

Critically, grunts often function as complete conversational turns. In task-oriented dialogue, a grunt can constitute a sufficient response that advances the dialogue state; acknowledging an observation, accepting a proposal, or signaling understanding, without requiring elaboration. This challenges dialogue models that do not consider grunts as contributing propositional content.

2.2. Information State and Common Ground in Dialogue

Understanding and responding appropriately in dialogue requires tracking what information is shared between participants, or their *common ground* (Clark and Brennan, 1991). Information state theories conceptualize this as a dynamic process where each communicative act updates the available context (Fernández, 2022). Modern computational approaches operationalize these concepts through dialogue state tracking (DST), which maintains estimates of participants' evolving beliefs throughout an interaction (Budzianowski et al., 2018; Henderson et al., 2014; Chen et al., 2020).

However, existing DST frameworks focus almost exclusively on propositional content conveyed through complete utterances. Minimal responses like grunts, which signal understanding or acceptance without introducing new factual information, are typically ignored or treated as continuers that maintain turn-taking without updating dialogue state. Moreover, these approaches assume verbal grounding: information enters common ground through speech acts, overlooking how observable events in situated contexts can be acknowledged and integrated. This gap becomes critical in multimodal settings where participants coordinate around shared perceptual access to their environment. Khebour et al. (2024b) introduced multimodal common ground tracking to address some of these gaps, though systematic analysis of how grunts function in such contexts is limited.

2.3. Multimodal Context in Embodied Interaction

Full understanding of information state in situated dialogue requires multimodal context beyond speech alone. Non-verbal behaviors including gesture, gaze, and physical actions provide essential layers of meaning that complement and sometimes substitute for verbal communication (Kendon, 2004). Gesture can carry propositional content equivalent to speech acts (Brutti et al., 2022; Goldin-Meadow, 2005), while physical actions in embodied settings create observable changes that participants must coordinate around (Hunter et al., 2015; Tam et al., 2023). Such actions directly impact common ground, requiring participants to acknowledge and integrate what they observe (Pustejovsky and Krishnaswamy, 2021).

Figure 1: Participants in the Weights Task learning and acknowledging new information

2.4. Situated Grounding, Data Sets, and Annotation Frameworks

Several corpora capture situated grounding processes. The Weights Task Dataset (WTD) provides multimodal situated interaction with speech and gesture coordination (Khebour et al., 2024a,b). The Chinese Whisper corpus documents furniture assembly similar to our Glider video (Kontogiorgos et al., 2020). The SCOUT corpus provides human-robot interaction data annotated with Dialogue-AMR (Lukin et al., 2024), while the DUEL corpus was specifically designed to study disfluencies, laughter, and exclamations in task-based settings across multiple languages (Hough et al., 2016).

Our work was inspired by observations of the WTD: participants observe an action (e.g., scale balancing), acknowledge it with a minimal vocalization, and proceed to a subsequent step that incorporates knowledge of the result (Figure 1). In these videos, grunts regularly serve as complete conversational turns that ground observable events and advance dialogue state.

3. Multimodal Tasks Corpus

Our corpus comprises 10 videos totaling approximately 3 hours of collaborative task interaction with 32 unique speakers. The data includes two distinct task types that enable systematic study of grunts across different collaborative contexts: a structured problem-solving task (Weights Task Dataset) and a naturalistic furniture assembly task (Glider video).

Weights Task Dataset The Weights Task Dataset (WTD) (Khebour et al., 2024a) consists of 10 videos where three university students collaborate to determine the weights of five blocks using a balance scale. Given the weight of a reference block, participants discover through experimentation that the remaining blocks follow the Fibonacci sequence. Each session runs approximately 15 minutes, featuring continuous

multimodal interaction as participants physically manipulate blocks, observe scale behavior, and verbally coordinate their reasoning. The WTD was designed for studying multimodal situated interaction and was previously augmented with Common Ground annotations (Khebour et al., 2024b). Nine of the videos were annotated for grunts; one was excluded due to poor audio quality.

Glider Video The Glider video is an unpublished recording collected by Candace Sidner in 1989 (Sidner, 2025), documenting two men assembling flat-pack furniture in a home. The 40-minute video captures naturalistic task-oriented interaction with extended silent work punctuated by coordination dialogue. Unlike the WTD, participants were not part of a formal research protocol. An accompanying transcript is from an unpublished MIT AI Lab technical report (personal communication, 2025).

Figure 2: Assembling furniture, where grunts frequently acknowledge observable progress

Corpus Design These tasks share key properties for studying grunts in situated contexts: co-located participants with shared perceptual access, hands-on manipulation creating observable events, and collaborative goals requiring coordination. However, they differ in task complexity, social context, and interaction style. The WTD features structured problem-solving with clear sub-goals and frequent discoveries, while the Glider video captures procedural assembly with sequential instructions. This variation enables analysis of how task characteristics may influence grunt usage patterns (see Section 7.1). Together, the dataset is approximately 20% Glider and 80% WTD.

4. Annotation Scheme

Our annotation scheme captures grunts, their transcriptions, and their dialogue act functions. The scheme is designed for two broad categories: grunts and gruntables. This work takes inspiration from research on laughter (Ginzburg et al., 2020; Mazzocconi et al., 2020), borrowing and adapting the concept of the *laughable* to create the *gruntable*; that is, the conversational trigger that elicits a grunt. The fine-grained token annotation enables downstream prosodic analysis.

4.1. Grunt Annotations

Each grunt annotation captures: its producer, orthographic transcription, dialogue act (Section 4.3), and whether the grunt is standalone or embedded in a longer turn. Annotators also labeled Expectedness and Polarity, though these dimensions are not analyzed here due to space constraints.

4.2. Gruntable Annotations

Gruntables are the interactional triggers that precede and potentially elicit grunts. Each annotation includes: its producer, gruntable type (Speech, Event, Speech describing Event/Gesture), content, and dialogue act (Section 4.3).

4.3. Dialogue Act Taxonomy

Our dialogue act classification system adapts categories from (Khebour et al., 2024b), which adopts a simplified version of the Dialogue Game Board (DGB) (Ginzburg, 2012). The taxonomy includes:

- **OBSERVATION**: Reports on a perceived physical action or state change
- **INFERENCE**: Deductive reasoning from available evidence
- **STATEMENT**: Introducing new propositional content, via speech or event
- **QUESTION**: Information-seeking queries
- **ANSWER**: Direct responses to questions
- **ACCEPT**: Agreement with previous evidence or conclusions
- **DOUBT**: Disagreement or uncertainty regarding prior assertions
- **FILLED PAUSE**: Hesitation phenomena and turn-starters without propositional content

Critically, events can function as STATEMENTS: when a scale balances or water begins to boil, these state changes assert propositional content about the world. Grunts responding to such events often carry the ACCEPT act, grounding the information conveyed by the event without elaboration.

4.4. Annotation Procedure

Annotations were created in ELAN, a video and audio annotation tool (Brugman et al., 2004). All of the WTD videos were annotated by the first author. Four were independently annotated by a second

graduate student. The Glider video came with a detailed transcription, which was converted to our annotation format. It received a second annotation following our conventions. The temporal boundaries of all standalone grunt annotations were refined to enable subsequent prosodic analysis.[1].

Spelling Normalization Annotators independently used varied spelling for perceived length (e.g., *ooh* vs. *oooooh*) and were creative with dashes (e.g., *mm-hm* vs. *mmhm*). For inter-annotator agreement, spelling was normalized: dashes were standardized and repeated letters removed (e.g., *hmmmm* → *hm*), except where indicating distinct tokens (e.g., *ooh* vs. *oh*). Vowel differences were preserved (e.g., *em* vs. *um*). A temporal tolerance of 500 ms was used when aligning grunts across annotators.

4.4.1. Example

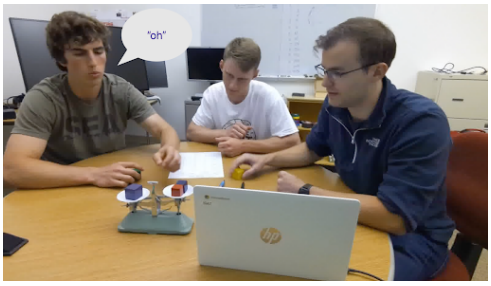

Figure 3: Grunting at new information

At the moment of Figure 3, the group knows *red = 10 g, blue = 10, green = 20*. P_2 places the purple block on the scale opposite red and blue. The scale does not balance [*GRUNTABLE: event, STATEMENT*]. P_1 responds "oh" [*GRUNT: ACCEPT*] They now know *purple* $\neq$ *20*, and P_3 suggests trying another block.

This example illustrates how grunts facilitate collaborative reasoning. P_1's *oh* signals understanding that the purple block does not weigh 20 g. This information was conveyed by the scale failing to balance. P_3 recommends testing another block (green) without explicitly restating the experimental result, demonstrating that the information has been grounded.

5. Formal Model

Here, we provide a formal account of how grunts update the common ground in situated dialogue. Our model treats observable events as contributing propositional content (like speech acts), and grunts as operations that can promote evidence to accepted fact. We use the scenario in Figure 3 as a basis for the example. While the formal model is not directly evaluated in this pilot study, it provides a precise theoretical vocabulary for the claim that grunts can be fact-promoting moves (rather than only continuers). Additionally, it offers a computational target for future DST systems that use minimal vocalizations. The subsequent walkthrough is a proof of concept for such a representation.

5.1. Framework

We formalize how grunts update common ground using evidence-based dynamic epistemic logic (EB-DEL) (van Benthem et al., 2014; Pacuit, 2017) combined with Common Ground Tracking (CGT) (Ginzburg, 2012).

The Common Ground Structure (CGS) comprises: *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, in this case colored blocks $C = \{r, b, g, p\}$ and weights $W = \{10, 20, 30...\}$. 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 taxonomy, a STATEMENT publicly contributes evidence, adding to EBank; an ACCEPT promotes evidenced content to fact, from EBank to FBank.

5.2. Detailed Walkthrough: WTD

We apply this formalism to the interaction in (Figure 3), demonstrating how it captures the systematic relationship between observable events and grunts. Before the sequence in Figure 3, the initial CGS is:

- $\mathrm{FBank}_0 = \{\mathrm{Eq}(r, 10),\ \mathrm{Eq}(b, 10),\ \mathrm{Eq}(g, 20)\}$
- $\mathrm{EBank}_0 = \varnothing$
- $\mathrm{QBank}_0 = \{\mathrm{Eq}(p, w)? : w \in W\} \cup \cdots$

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

Step 1 (Event/Gruntable = STATEMENT). P_2 places p against $r{+}b$ on the scale; it *does not* balance. The jointly observed outcome is treated as a public announcement of an *event proposition*: $\varepsilon := \neg\mathrm{Bal}(r{+}b, p)$. The Banks after Step 1:

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

[1]Annotation conventions available at: https://github.com/richardbrutti/grunts-corpus

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

Operationally in the CGS:

$$\mathrm{EBank}_1 \ni \phi \xrightarrow{\text{ACCEPT}} \mathrm{FBank}_2 \ni \phi,$$

and ϕ is removed from EBank. The Banks effect after Step 2 (the grunt update):

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

Thus 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 detailed walkthrough demonstrates how our formalism captures the systematic relationship between observable events, grunts, and common ground updates in collaborative problem-solving. The same mechanism applies across different situated contexts.

6. Corpus Statistics and Quality

6.1. Inter-annotator agreement

We report agreement using both the Sørensen-Dice coefficient (DSC) and Krippendorff's alpha (Dice, 1945; Sorensen, 1948; Krippendorff, 2011).

Grunt Detection For doubly annotated videos, DSC ranged from 0.400 to 0.638, with an overall DSC of 0.498 across all grunt annotations. 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 interpretation tasks such as emotion annotation (Lotfian and Busso, 2017). Grunt counts from the two annotators were typically within 10%. Disagreements primarily occurred with overlapping speech and quiet vocalizations.

Dialogue Act Classification Agreement on dialogue act labels showed similar patterns. DSC ranged from 0.379 to 0.625 across videos, while Krippendorff's alpha ranged from 0.379 to 0.501.

Gruntable Identification Gruntable annotation proved substantially more difficult than grunt detection, with agreement too low to be meaningfully reported. This reflects both 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. All analyses in this paper treat the first author's annotations as the gold standard.

Table 1: Grunt Dialog Act Distribution

Dialog Act	Count	%
ACCEPT	201	40.4
Filled Pause	138	27.8
Turn Starter	48	9.7
OBSERVATION	33	6.6
ANSWER	27	5.4
DOUBT	24	4.8
OTHER	26	5.2

7. Distributional Analysis

Grunt usage varies across multiple factors in our corpus: task type, participant, dialogue act, and antecedent type. These patterns reveal systematic functional specialization, with specific tokens preferentially used for different grounding contexts.

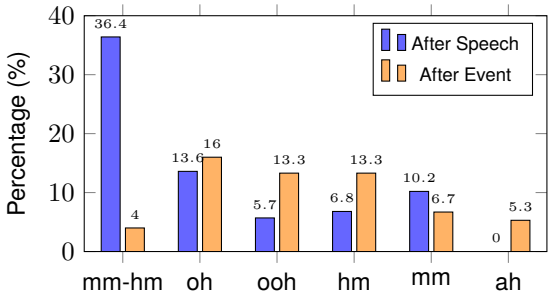

Figure 4: Functional specialization among ACCEPT tokens (n=201 ACCEPTs). Bars show percentage of ACCEPTs using each token after speech (blue) vs. after observable events (orange). *mm-hm* strongly prefers speech antecedents (36.4% vs. 4.0%), while *oh*, *ooh*, and *hm* prefer event antecedents.

7.1. Overall Distribution Patterns

As shown in Table 1, grunts in these task-based settings most commonly function to ACCEPT the previous contribution (40.4%), whether verbal or non-verbal. This dominance of acceptance grunts supports our claim that these vocalizations serve critical grounding functions. The second most frequent category is FILLED PAUSE (27.8%), comprising primarily *um* and *uh* used as turn-starters, consistent with findings in (Clark and Tree, 2002).

Table 2 shows the distribution of antecedent types. While 41.6% of grunts lack clear an-

tecedents (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 demonstrates that non-verbal events function as conversational contributions on par with verbal utterances—a key finding supporting multimodal approaches to information state tracking.

Table 2: Antecedent Type Distribution

Antecedent Type	Count	%
No antecedent	207	41.6
Speech	126	25.4
Event	125	25.1
Speech Descr. Event	35	7.0
Other	4	0.8

7.2. Token Specialization by Antecedent

Figure 4 reveals functional specialization among grunt tokens based on antecedent type. Within ACCEPT responses, *mm-hm* shows dramatic 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.

7.3. Task Effects: Glider vs. WTD

The videos represent different contexts: friends assembling furniture at home vs. strangers in an experimental task. Despite comprising only 20% of the corpus, Glider 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, with less need for collaborative inference. In contrast, WTD's problem-solving structure generates frequent observable events requiring acknowledgment and interpretation.

Both corpora feature meta-communicative vocalizations, including whistling and humming that mark discourse segment transitions (Grosz and Sidner, 1986). Celebratory vocalizations (*woo*, extended *oh*, singing) signal problem-solving success in WTD.

Grunt preferences varied by task. Glider shows dominance of *uh* (25.4% vs. 10.1% in WTD), while *um* was more prevalent in WTD (16.6% vs. 5.4%). Additionally, *uh-huh* appeared uniquely in Glider, while *hm* and *ooh* were absent. *Mm* was much more common in WTD. These differences may reflect individual preferences, task demands, or social dynamics.

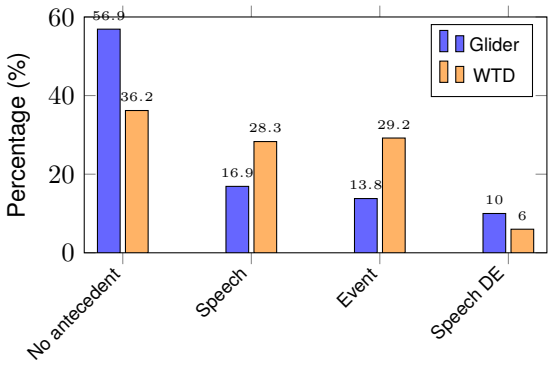

Figure 5: Antecedent type distribution by task context. Glider shows more grunts without antecedents and fewer event responses than WTD, reflecting differences in task structure.

8. Prosody of Standalone Grunts

We extracted acoustic features (duration, mean pitch, pitch range, F1) from all standalone grunts using Parselmouth (Jadoul et al., 2018). This analysis reveals that speakers encode both antecedent type and dialogue act function through prosodic modulation of identical grunt tokens.

8.1. Prosodic Differences by Antecedent

Across both WTD and Glider, responses to non-verbal events showed different prosodic profiles compared to responses to verbal speech turns. Responses to events were 38% longer in duration, exhibited 42% greater pitch range, and showed higher F1 formant values (Table 3), possibly indicating more open vowels. These differences suggest that speakers employ distinct strategies when reacting to observable events versus conversational exchanges. The increased duration and pitch variation in event responses may reflect additional processing demands or a need for more emphatic acknowledgment when grounding non-verbal information.

To control for the possibility that prosodic differences simply reflected different distributions across dialogue act contexts, we examined ACCEPT responses exclusively. Within this category, antecedent type exerted substantial effects: ACCEPT responses following events were 45% longer and showed 39% greater pitch range (Table 5). This demonstrates that speakers in both tasks systematically adjust the realization of functionally equivalent responses based on what they are responding to, suggesting that prosody encodes information about

Table 3: Acoustic features of responses by antecedent type. Mean (SD).

Antecedent Type	n	Duration (s)	Mean Pitch (Hz)	Pitch Range (Hz)	F1 (Hz)
Speech	106	0.453 (0.231)	187.1 (116.1)	102.9 (132.3)	582
Event	83	0.626 (0.515)	193.9 (108.8)	146.3 (158.7)	652
Speech DE	28	0.485 (0.185)	233.7 (141.5)	155.0 (187.6)	686

Note: Event vs. Speech differences were significant for duration, $t(107.8) = 2.84$, $p < .01$, $d = 0.45$; pitch range, $t(158.8) = 2.00$, $p < .05$, $d = 0.30$; and F1, $t(176.3) = 4.78$, $p < .001$, $d = 0.70$ (Welch's t-tests).

Table 4: Variation of *mm-hm* by dialogue act

Dialog Act	n	Dur (s)	Pitch (Hz)	Range (Hz)
ACCEPT	48	0.424	188.9	96.6
ANSWER	11	0.502	216.9	151.9

Note: Effect sizes (Cohen's d) indicate small-to-medium prosodic differences by function: duration $d = 0.43$, pitch range $d = 0.46$.

discourse context beyond the immediate communicative act.

8.2. Prosodic Variation by Dialogue Act

Analysis of individual tokens reveals systematic prosodic variation as a function of dialogue act, demonstrating that speakers modulate identical lexical forms to signal different communicative functions. The token *mm-hm* showed different acoustic profiles when functioning as ANSWER versus ACCEPT (Table 4). ANSWER tokens were 18% longer in duration and exhibited 57% greater pitch range than ACCEPT tokens. These prosodic differences occurred despite ANS tokens being produced almost exclusively in response to speech (10 of 11), suggesting that dialogue act function rather than antecedent context drives the prosodic modulation of *mm-hm*. This pattern extended across other high-frequency tokens, with *hm* showing even more dramatic flexibility across five different dialogue acts, ranging from brief, high-pitched question forms (0.245s, 281 Hz) to extended, low-pitched doubt expressions (0.934s, 118 Hz)—a 281% duration difference.

Despite small sample sizes for individual token-function pairs, these findings demonstrate that prosody disambiguates minimal response particles where the token alone provides insufficient information. Speakers draw on prosodic resources to map identical grunt forms onto distinct communicative functions, supporting the view that grunts are sophisticated communicative tools integrating lexical and suprasegmental information, not merely reduced forms of speech.

9. Conclusion

Consider the cooking scenario introduced earlier: water begins to boil, one person says *ooh*, and the other adds pasta. This interaction illustrates how grunts ground information from observable events, enabling coordinated action without explicit verbal confirmation. Through systematic annotation and analysis of 497 grunts across approximately 3 hours of multimodal collaborative tasks, this work demonstrates that such minimal vocalizations play systematic roles in dialogue management and common ground construction.

9.1. Key Results

Our work provides both empirical findings and methodological resources for studying grunts in situated dialogue:

Annotation scheme and corpus resource. This work presents a systematic annotation scheme for non-lexical vocalizations in multimodal collaborative tasks, capturing grunts, their antecedents (speech and observable events), and dialogue act functions. The resulting corpus comprises 497 annotated grunts across 10 videos (approximately 3 hours) with 32 speakers, representing the first resource enabling quantitative analysis of how minimal vocalizations ground information from both verbal and non-verbal sources. The annotation scheme is designed for multimodal contexts and provides a replicable framework for future corpus development. Both the annotation guidelines and annotated data provide foundations for developing automatic grunt detection systems and for comparative studies across languages, tasks, and populations.

Functional specialization by antecedent type Grunt tokens exhibit systematic specialization based on what they respond to. Backchannels like *mm-hm* predominantly acknowledge verbal speech, while reactive tokens such as *oh*, *ooh*, and *hm* preferentially acknowledge observed events. Critically, grunts respond to speech and observable events at nearly equal rates, demonstrating that non-verbal

Table 5: Acoustic features of ACCEPT responses by antecedent. Mean (SD)

ACC Following	n	Duration (s)	Mean Pitch (Hz)	Pitch Range (Hz)
Speech	77	0.428 (0.201)	190.1 (128.6)	100.9 (139.4)
Event	50	0.619 (0.565)	180.9 (99.1)	140.2 (161.8)
Speech DE	23	0.492 (0.185)	223.2 (143.1)	134.5 (183.6)
Effect size (Event vs. Speech):				
Difference		+44.7%	−4.8%	+39.0%

Note: ACCEPT after Event showed significantly longer duration than ACCEPT after Speech, $t(57.1) = 2.30$, $p < .05$, $d = 0.49$, and marginally greater pitch range, $t(93.6) = 1.41$, $p = .16$, $d = 0.26$ (Welch's t-tests).

events function as conversational contributions on par with utterances. This supports multimodal information state tracking that accounts for how minimal vocalizations ground both verbal and non-verbal sources.

Prosodic encoding of grounding source Responses to non-verbal events showed markedly different prosodic profiles than responses to speech, with substantially longer duration and greater pitch range. These differences persisted when controlling for dialogue act function, indicating that prosody encodes grounding source beyond immediate communicative function. Individual tokens showed systematic prosodic variation across dialogue acts, revealing how speakers use acoustic-phonetic resources to map identical grunt forms onto distinct communicative functions.

Grunts as complete conversational turns Approximately 80% of grunts in our corpus functioned as standalone conversational turns rather than embedded elements within larger utterances. These standalone grunts often constituted sufficient responses that advanced the dialogue state—acknowledging observations, accepting proposals, or signaling understanding—without requiring verbal elaboration. For example, after an observable event like a scale balancing, a participant's *oh* can serve as a complete turn that grounds the information and enables coordinated next actions. This challenges dialogue models that treat minimal responses primarily as backchannels that hold the floor without contributing propositional content.

9.2. Implications for Dialogue Systems

Our findings suggest opportunities for improving dialogue system design. In dialogue state tracking, grunts could serve as explicit signals that information has been grounded, particularly in multimodal contexts where actions and observations update common ground non-verbally. The near-equal distribution of grunts responding to speech versus events suggests potential benefits from integrating visual perception with language understanding to detect relevant environmental events and recognize grounding signals. Finally, the high proportion of standalone grunts indicates that systems could benefit from treating these vocalizations as complete turns, and from generating contextually appropriate grunts with prosodic variation matching dialogue act function and antecedent type.

9.3. Limitations and Future Directions

Corpus and Generalization Our corpus represents a specific type task with co-present participants working toward shared goals. The repeated nature of the WTD task is a design choice of the original corpus, and not an artifact of our data collection, constraining generalization. However the controlled structure also makes the WTD well-suited to a pilot. The predictable even structure allowed for cleaner annotation of grunt-event relationships. While we hypothesize that grunt patterns extend to everyday multimodal interaction, generalization to other dialogue contexts (remote communication, non-collaborative settings, spontaneous conversation) requires further investigation. The corpus comprises only 3 hours of interaction; larger-scale studies would enable more robust statistical analysis and examination of individual variation. Our prosodic analysis focused on standalone grunts; embedded grunts may show different patterns and warrant separate analysis.

Annotation Challenges The moderate inter-annotator agreement for grunt detection reflects challenges in annotating these subtle signals, particularly with single-microphone recordings. Turn-taking patterns show substantial individual variation (Skantze and Cavalcanti, 2025), making consistent annotation difficult. Gruntable identification proved even more challenging, requiring annotators to infer temporal and causal relationships between grunts and their triggers. Future data efforts should explore: (1) multi-microphone recording setups, (2)focus on more easily identifiable gruntable types, and

(3) better annotator training.

Future Research Directions This work opens several avenues for future investigation. First, automatic detection and classification systems for grunts in multimodal dialogue could build on our annotation scheme and findings, potentially using acoustic-prosodic features to predict dialogue act function and antecedent type with the aim of improving naturalness. Second, investigating how grunts interact with gesture and gaze could illuminate their role in establishing reference and managing joint attention in embodied interaction. Third, extending annotation to additional task types, social contexts, and languages would assess the generalizability of functional specialization patterns. Finally, other non-lexical vocalizations beyond grunts warrant systematic study, including sounds like whistling or humming, and celebratory vocalizations that signal problem-solving success.

10. Ethics Statement

The Glider video was collected with participants' knowledge and consent for research purposes. The Weights Task data was released in an open access journal. To protect privacy, we use participant IDs and do not include unnecessary identifying information. Annotators were graduate students with appropriate training and the ability to decline content.

While this research aims to improve dialogue systems, we acknowledge potential privacy concerns with voice data and encourage privacy-preserving approaches. The corpus represents limited demographics (primarily US university students) and should not be assumed to generalize without validation.

11. Acknowledgements

We would like to thank Candace Sidner for use of the Glider video, and helpful comments on an earlier draft of this work. We would also like to thank Professor Noor Abo Mokh and the students of CS 230B at Brandeis for their help with the annotations. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grants DRL 2019805 and DRL 2454151. The opinions expressed are those of the authors and do not represent views of the NSF.

12. Bibliographical References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating multi-media/multi-modal resources with elan. In *LREC*, pages 2065–2068. Lisbon.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Justine Cassell, Timothy Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. 2000. Designing embodied conversational agents. *Embodied conversational agents*, 29.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7521–7528.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Daniel Erker 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. *Estudios del Observatorio Cervantes en Harvard*, 80:1–31.

Raquel Fernández. 2022. Dialogue. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. In *International conference on*

universal access in human-computer interaction, pages 828–837. Springer.

Rod Gardner. 1997. The conversation object mm: A weak and variable acknowledging token. *Research on Language and Social Interaction*, 30(2):131–156.

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

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics (2021-...)*, 5(1).

Susan Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Matthew Henderson, 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)*, pages 292–299.

John Heritage. 2002. Oh-prefaced responses. *The language of turn and sequence*, page 196.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In *10th edition of the Language Resources and Evaluation Conference*.

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

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*.

Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, et al. 2024b. Common ground tracking in multimodal dialogue. *arXiv preprint arXiv:2403.17284*.

Dimosthenis Kontogiorgos, 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*, pages 743–749, Marseille, France. European Language Resources Association.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Stephanie M. Lukin, Claire Bonial, Matthew Marge, Taylor A. Hudson, Cory J. Hayes, Kimberly Pollard, Anthony Baker, Ashley N. Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. 2024. SCOUT: A situated and multi-modal human-robot dialogue corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458, Torino, Italia. ELRA and ICCL.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2020. What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, 13(3):1302–1321.

Eric Pacuit. 2017. *Neighborhood semantics for modal logic*. Springer.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3):307–327.

Candance Sidner. 2025. Glider video.

Gabriel Skantze and Julio Cesar Cavalcanti. 2025. " dyadosyncrasy", idiosyncrasy and demographic factors in turn-taking. *Pre-print: accepted to Interspeech 2025*.

Thorvald Sorensen. 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. *Biologiske skrifter*, 5:1–34.

Christopher Tam, Richard Brutti, Kenneth Lai, and James Pustejovsky. 2023. Annotating situated actions in dialogue. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 45–51, Nancy, France. Association for Computational Linguistics.

Johan van Benthem, David Fernández-Duque, and Eric Pacuit. 2014. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133.

Nigel Ward. 2006. Non-lexical conversational sounds in american english. *Pragmatics & Cognition*, 14(1):129–182.

David P Wilkins. 1992. Interjections as deictics. *Journal of Pragmatics*, 18(2-3):119–158.

Victor H Yngve. 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*, pages 567–578.