The Grunt Work of Conversational Grounding: Minimal Responses in Multimodal Dialogue

Richard Brutti (brutti@brandeis.edu) & James Pustejovsky (jamesp@brandeis.edu)
Department of Computer Science, Brandeis University

Abstract

Conversational grounding, the establishment of mutual understanding between participants, is fundamental to studies of dialogue. However, methods of grounding multimodal information remain understudied. We investigate how minimal vocalizations (conversational grunts like *hm* and *ooh*) encode grounding of verbal information versus observed events in collaborative tasks. Acoustic analysis of 497 grunts from natural dialogues revealed that token selection is strongly specialized by information source. Backchannels like *mm-hm* predominantly acknowledge speech, while reactive tokens like *oh* preferentially responded to events. Event-following grunts also exhibit longer duration and more open vowel production at the corpus level, though these differences are largely a consequence of which tokens speakers select rather than independent prosodic modulation. These findings establish observable events as informational contributions requiring explicit acknowledgment, and identify lexical selection as the primary mechanism by which speakers differentiate grounding sources in co-situated task interaction.

Keywords: grunts; minimal responses; grounding; multimodal dialogue; prosody; collaborative tasks

Introduction

Human communication requires the seamless integration information from multiple modalities in real time. Let's imagine two people assembling furniture together: one tightens a bolt, and the other responds '*"mm-hm."* This minimal vocalization signals successful integration of the completed action along with the understanding of overall task progress. The second person proceeds immediately to the next step without explicit verbal confirmation, demonstrating that minimal responses enable sophisticated coordination in collaborative activity. *Conversational grunts* (vocalizations including *oh*, *mm-hm*, *uh-huh*) are common in human interaction. Prior research has focused primarily on their pragmatic functions as backchannels and acknowledgments (Gravano & Hirschberg, 2011; Ward, 2006; Yngve, 1970), but has not systematically investigated how speakers ground multimodal information through token selection and prosodic modulation. In a dialogue setting, where participants are located in the same space, where both verbal contributions and observable events update common ground (Hunter et al., 2015), speakers produce these kinds of minimal responses that vary both in form (*which* grunt token) and expression (*how* it is produced acoustically). Understanding the coordination of these axes could reveal cognitive processes underlying the integration of multimodal input.

We investigate three questions about how speakers use minimal responses to signal multimodal grounding. First, do speakers systematically select different grunt tokens based on the source of the information being grounded (verbal vs. perceptual)? Second, do prosodic features differ between grunts that respond to speech and those that respond to events, and to what extent do these differences reflect token selection rather than independent prosodic modulation? Third, what do these patterns reveal about how speakers acknowledge multimodal information in real time?

We use the term grunts to refer to non-lexical vocalizations produced as standalone or quasi-standalone communicative contributions (e.g., oh, mm-hm, ooh, hm). We distinguish these from filled-pause hesitations (uh, um) which serve speaker-internal planning functions (Clark & Tree, 2002), although the boundary is sometimes blurred and the same form can serve both functions in different contexts.

The remainder of this paper is organized as follows. We review literature on grounding in multimodal dialogue and prosodic encoding in minimal responses (Section 2), describe our corpus and acoustic analysis methods (Section 3), present results demonstrating token specialization and prosodic variation (Section 4), and discuss theoretical implications for models of grounding, prosody, and cognitive processing (Section 5). We conclude by considering limitations and future directions for investigating minimal vocalizations as windows into cognitive mechanisms of communication.

Background

Grounding in Multimodal Dialogue

Grounding is the process by which interlocutors establish mutual understanding of contributions (Clark & Brennan, 1991; Clark & Schaefer, 1989; Fernández, 2022). The Clark and Schaefer (1989) contribution model describes the coordination of speakers and addressees with a presentation-acceptance process. Acceptance is signaled through explicit acknowledgments, continued attention, or relevant next speech turns. This grounding process maintains the *common ground*, the mutually shared knowledge and beliefs that enable coordination.

While much research has focused on verbal grounding, human interaction is fundamentally multimodal. Speakers integrate information from speech, gesture (Kendon, 2004), gaze (Alikhani & Stone, 2020; Cassell, 2001), and co-observable

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

actions (Tam et al., 2023). In situated dialogue, physical events and actions function as contributions that require grounding just as verbal utterances do (Pustejovsky & Krishnaswamy, 2021). In our earlier furniture example, the completion of an action (placing a bolt, tightening a screw) conveys information that must be mutually acknowledged for coordination to proceed.

Some previous work has examined grounding in multimodal contexts. Tanenhaus et al. (1995) demonstrate that speakers rapidly integrate visual information with linguistic context in reference resolution. Brennan and Hanna (2009) show that speakers adapt their referring expressions based on speaking partner. These leave open questions about how speakers signal their *comprehension* and *acceptance* of multimodal information.

Grunt tokens can provide a particularly informative window into grounding processes. They are produced quickly and with minimal planning, and they may reveal automatic cognitive processes underlying information integration (Clark & Tree, 2002). The ways in which speakers adapt these minimal forms to signal grounding of different types of information types, e.g., verbal versus perceptual, remains poorly understood.

Theories of event cognition and working memory suggest that perceptual and verbal information engage distinct cognitive resources. Event Segmentation Theory proposes that observers automatically parse ongoing activity into discrete events (Radvansky & Zacks, 2014; Zacks & Swallow, 2007). Segmentation occurs when changes in goals, objects, or spatial relations increase prediction error, triggering updates to mental representations of ongoing activity. This perceptual segmentation recruits specialized neural mechanisms and engages visuospatial working memory. Processing verbal contributions, in contrast, relies on the phonological loop, which maintains linguistic information through articulatory rehearsal (Baddeley, 2012). These modality-specific systems can operate in parallel under executive control, suggesting that grounding observable events may impose different processing demands than grounding verbal contributions, potentially reflected in systematic differences in acknowledgment behavior.

Prosody and Minimal Responses

Minimal responses can carry communicative functions through multiple channels. At the "lexical" level, different tokens can signal different pragmatic meanings: *mm-hm* and *uh-huh* typically function as continuers or acknowledgments, while *oh* marks information state changes (Heritage, 1984, 2002). Clark and Tree (2002) propose that speakers use *uh* to indicate a minor delay in speaker, and *um* for a major one. Ward (2006) demonstrated that even sub-grunt components, like individual phonemes, creaky voice, and breathiness seemingly have sound-meaning correspondence.

In linguistics, the encoding of surprise or novelty is known as *mirativity*, or the grammatical marking that information is new, unexpected, or violates the speaker's expectations (Aikhenvald, 2012; DeLancey, 1997). Kraus (2019) demonstrates that English discourse particles like *oh* and *huh* are commonly used when speakers' expectations are violated, and that these mirative particles are systematically paired with heightened prosodic contours.

This linguistic analysis connects to well-established cognitive processes. Surprise can trigger arousal and attentional responses that may be reflected in vocal production (Meyer et al., 1997). If observable events in dialogue introduce perceptual surprise or violate expectations, they may elicit grunt tokens like *oh* with distinct prosodic marking.

Beyond reactive tokens, prosody can systematically encode pragmatic meaning in minimal response grunts. Ward and Tsukahara (2000) showed that Japanese backchannels vary in pitch, duration, and timing based on discourse context. Gravano and Hirschberg (2011) identified prosodic features in English that predict backchannel-inviting cues, demonstrating prosodic coordination between speakers and listeners. Hockey (1993) found that pitch contour of *okay* and *uh-huh* can reduce the possible interpretations in context. Other studies have shown variability across different speakers when producing backchannels in Dutch and Japanese speakers of English (Cutrone, 2010; de Kok & Heylen, 2011).

However, existing research has focused primarily on how minimal responses ground *verbal* contributions. Whether and how speakers prosodically differentiate responses to speech versus observable events remains unexplored. Moreover, if these patterns operate automatically, they would offer insight into cognitive mechanisms linking perception, surprise, and vocal production. Our current work also borrows from work on laughter, analyzing the preconditions for grunts along with their properties following Mazzocconi et al. (2020).

The present study addresses these gaps by examining both token selection and prosodic modulation in minimal responses to speech and observable events during collaborative tasks. We investigate how speakers coordinate lexical and prosodic encoding to signal comprehension of multimodal information, with particular attention to whether event grounding elicits patterns consistent with surprise-marking functions.

Data & Methods

Corpus

We analyze 497 conversational grunts from collaborative dialogue. Briefly, the data comprise approximately 3 hours of video of participants engaged in two types of collaborative tasks:

Weights Task Dataset (WTD): Nine videos of triads solving a problem involving colored blocks and a balance scale. Participants must determine block weights through collaborative reasoning and testing (Khebour et al., 2024). This task creates a context with frequent observable events (balance scale movements) and verbal hypotheses from participants. The subjects were primarily university students recorded in a laboratory setting.

Glider Video: One 40-minute long video of a dyad assembling a piece of furniture from instructions. It is an unpub-

lished recording collected by Candace Sidner in 1989 using a camcorder (Sidner, 2025). This provides a procedural collaboration context with physical actions as primary events and minimal verbal instruction, as the two adult male participants are reading instructions and assembling sequentially.

In both settings, participants were co-located with shared perceptual access to task materials, enabling multimodal grounding through both speech and observable events. The videos were annotated in ELAN (Brugman et al., 2004) for grunt tokens, their temporal boundaries, dialogue act functions (Ginzburg, 2012), and antecedent types (speech, observable event, or speech describing an event). All videos were recorded with a single microphone, resulting in variable acoustic quality, depending on speaker's proximity.

Figure 1: Collaborative task contexts. Top: Weights Task Dataset showing triadic interaction around balance scale. Bottom: Glider furniture assembly showing dyadic coordination over shared workspace. Both contexts provide co-located interaction with shared perceptual access to observable events.

A subset of 4 videos was independently annotated by a second annotator as a reliability check. Krippendorff's alpha ranged from 0.38 to 0.48 per video for grunt detection and 0.38 to 0.50 for dialogue act classification. We treat these values as a lower bound on agreement reflecting the difficulty of the annotation task for novice annotators; the analyses reported below use the first author's annotations as the reference. These values are comparable to those reported for other subjective audio annotation tasks such as emotion recognition (Lotfian & Busso, 2017). Category-specific agreement for the ACCEPT label could not be reliably estimated from the dual-annotated subset due to small per-category counts; we flag this as a limitation.

Our analysis primarily focuses on 150 ACCEPT responses when grunts are standalone complete dialog turns, which constitute the most frequent dialogue act, and allow for comparisons within the same function. This subset includes grunts following speech (n=101), and following observable events (n=50). The 101 grunts following speech can be further divided into speech that describes an immediately preceding event (n=23) and conversational speech (n=78).

Acoustic Analysis

Acoustic features were extracted from the manually annotated temporal boundaries standalone grunts (those constituting complete conversational turns, and had no further speech) using Parselmouth (Jadoul et al., 2018), a Python interface to Praat (Boersma & Weenink, 2021). We focused on features that may reflect cognitive processing demands or affective states:

Duration: Total vocalization length from onset to offset.

Mean F0: Average fundamental frequency, converted to semitones relative to each speaker's median F0 across all that speaker's annotated tokens. This speaker-relative semitone scale is appropriate because pitch perception is non-linear and absolute Hz values are dominated by between-speaker baseline differences (Levitan & Hirschberg, 2011).

Pitch Range: Difference between maximum and minimum F0 within the vocalization.

F1 (first formant): Mean frequency of the first formant. F1 reflects vowel openness and is therefore a property of vowel quality and lexical choice; we report it descriptively rather than as a paralinguistic prosodic feature.

Statistical Analysis

To account for repeated measures from individual speakers, we fit linear mixed-effects models of the form

$$\text{outcome}_{ij} = \beta_0 + \beta_1\,\text{event}_{ij} + \beta_2\,\text{Sp}{\rightarrow}\text{Ev}_{ij} + b_i + \epsilon_{ij} \quad (1)$$

where $b_i \sim \mathcal{N}(0, \sigma_b^2)$ is a speaker-specific random intercept and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. In other words, each speaker is allowed their own baseline level of the outcome (duration, F0, pitch range, or F1), and 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) after each speaker's baseline is accounted for. This prevents speakers who happen to contribute many grunts from disproportionately driving the estimated effects. Speaker identity was constructed as a composite of group and within-group ID, since participant labels (p1, p2, p3) are not unique across groups in the WTD corpus. Models were fit by REML; coefficient significance was assessed by t-tests on fixed effects with effective degrees of freedom approximated based on the within- and between-speaker variance components. We report Cohen's d 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, we used chi-square tests to evaluate whether token choice (e.g., *mm-hm* vs. *oh*) varied systematically by antecedent type.

Results

Functional Specialization in Token Selection

Speakers showed strong systematic preferences for which tokens they produced based on grounding source. Figure 2 presents the distribution of token types by antecedent. Backchannels like *mm-hm* predominantly followed speech contributions (75.8%), while reactive tokens like *ooh* and *hm* preferentially followed observable events (66.7% and 57.7% respectively). When Sp→Ev is folded into speech (treating both as verbally-delivered antecedents), *mm-hm* follows verbal contributions in 95.2% of cases. This pattern was highly significant ($\chi^2(10) = 48.95$, $p < .001$, $n = 190$ standalone grunts across the six most frequent tokens), indicating that token choice provides categorical encoding of information source.

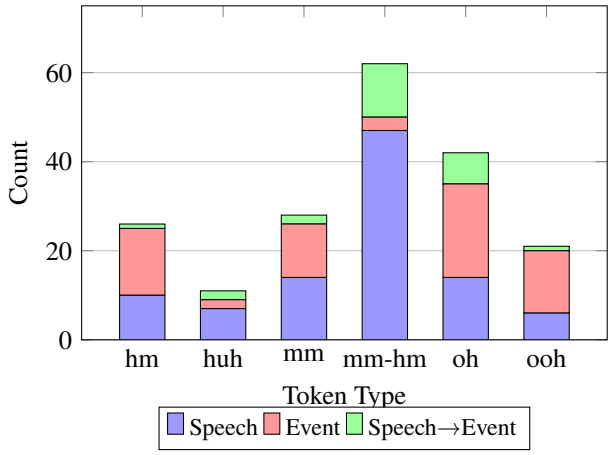

Figure 2: Distribution of grunt tokens by antecedent type ($n = 190$ standalone grunts across the six most frequent tokens). Backchannels (*mm-hm*, *mm*) predominantly follow speech contributions, while reactive tokens (*oh*, *ooh*, *hm*) preferentially follow observable events. Stacked bars show counts for each token–antecedent combination.

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.

Prosodic Encoding of Grounding Source

Beyond token selection, we examined whether prosodic features differed by what kind of information speakers were grounding. Table 1 presents acoustic features by antecedent type for the 150 ACCEPT responses with classifiable antecedents.

Table 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

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

Speech-describing-event responses (Sp→Ev) (grunts produced when the immediately preceding 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. We return to its theoretical interest in the Discussion.

Disentangling Token and Prosody

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. To distinguish these possibilities, we re-fit the prosodic models with grunt token included as a fixed effect (restricted to the six tokens 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 are 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 directionally 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.

A robustness check 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.

Discussion

Token Selection as Primary Grounding Signal

Our findings suggest that speakers use lexical selection as the principal mechanism for differentiating grounding sources in co-situated task interaction. At the lexical level, token choice provides strong categorical signals. This functional specialization is consistent with prior work characterizing *oh* as a change-of-state token (Heritage, 1984, 2002) and provides corpus-level support across two contexts.

At the population level, grunts following events are also longer and have higher F1 than grunts following speech. However, the token-controlled analysis indicates that these prosodic differences are largely a consequence of which tokens speakers selected rather than independent prosodic modulation. Once token identity is included in the model, the duration effect disappears and the F1 effect is reduced to a marginal trend. The simplest interpretation of our data is therefore that speakers do most of the differentiating through token choice, and the prosodic profiles of the resulting grunts follow primarily from the inherent acoustics of the selected vocalizations.

This interpretation does not rule out independent within-token prosodic modulation. Our exploratory analysis of *hm* showed large effect sizes for duration and F1 in the predicted direction, consistent with speakers modulating prosody beyond token choice. But these patterns require corpora with denser sampling to be tested rigorously, and the present data are underpowered for that question. The speech-describing-event (Sp→Ev) condition offers a useful probe of where the lexical-prosodic boundary falls. 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.

Implications for Grounding Theory

Our results make two contributions to theories of grounding in dialogue. First, we show that observable events function as informational contributions requiring situated grounding. Speakers in our corpus grunted to events and speech at comparable rates, showing the need to incorporate perceptual information into speech-centric models of dialogue. In collaborative tasks with shared perceptual access, what happens in the environment constitutes information that must be mutually acknowledged for coordination to proceed.

Second, the strong functional specialization of token choice by information source reveals that there is a small inventory of minimal forms whose lexical identity carries categorical information about what kind of contribution they are acknowledging. This complements rather than replaces verbal grounding: minimal vocalizations efficiently signal acknowledgment of perceptual events without requiring full verbal contributions, supporting fast-paced coordination in collaborative tasks.

Cognitive Mechanisms in Co-Situated Communication

The patterns we observe are consistent with the possibility that grounding observable events and speech engage different processes. For instance, perceptual state changes (Radvansky & Zacks, 2014; Zacks & Swallow, 2007) use additional cognitive methods besides linguistic (Baddeley, 2012). However, the present data do not adjudicate between cognitive accounts and simpler explanations of speakers' lexical inventories. Distinguishing automatic perception-driven token selection from learned lexical conventions would require experimental manipulation of expectedness or perceptual salience. We treat the cognitive interpretation as a hypothesis raised by the present data rather than an inference.

Limitations and Future Directions

This work has several limitations that can suggest directions for future research. First, our data comes entirely from task-based interactions. We have not tested whether these patterns might generalize to other contexts, such as casual conversation or computer-mediated interactions, or with speakers of other languages.

The sample size of our corpus is modest. The strong specialization of tokens by antecedent means that token choice and prosody are confounded in observational data. An experimental design constraining token choice would more directly test independent prosodic modulation. Larger samples, with cleaner audio recording methods, would enable more comprehensive review of prosodic variation.

A further limitation is that we have not modeled prosodic and lexical entrainment between interlocutors. Backchannels are known to entrain prosodically and lexically to the preceding speech (Levitan & Hirschberg, 2011), and convergence between partners may interact with the patterns we report, particularly in the WTD triadic context, where backchannels are produced in response to specific speakers. Future work could analyze partner-specific effects by including dyadic alignment measures or by analyzing speakers' baseline prosodic distributions independently of grunt tokens. Linguistic theories of mirativity suggest that tokens like *oh* mark information as new or unexpected (Aikhenvald, 2012; Kraus, 2019), which could partially explain their preference for event contexts. Testing such functional accounts directly would require behavioral or computational measures of antecedent expectedness. Notably, recent work on extended-reasoning language models has observed that such systems produce tokens like *wait*, *hmm*, and *oh* at points of internal state change during chain-of-thought reasoning.

Conclusion

We investigated how speakers signal grounding of verbal vs. perceptual information through minimal vocalizations in collaborative dialogue. Token selection shows strong functional specialization by information source: backchannels like *mm-hm* predominantly acknowledge speech, while *oh* and *ooh* preferentially follow events. Population-level prosodic differences (longer duration, higher F1 for event-following grunts) appear to be largely attributable to the inherent acoustics of the selected tokens rather than to independent prosodic modulation, suggesting that lexical choice is the primary mechanism by which speakers differentiate grounding sources. These findings extend grounding theory to multimodal contexts and reveal that even minimal, non-lexical vocalizations carry systematic information about what type of contribution speakers are acknowledging.

Acknowledgments

The authors used Claude (Anthropic) as a writing and analysis assistant during manuscript preparation. Claude helped draft initial versions of the Discussion and Limitation sections, as well as the code for the statistical analysis and effect-size reported in the Results. All research design, data collection, annotation, analysis decisions, theoretical framing, and final content were determined by the authors, who substantially reviewed and revised any AI-generated text. We would like to thank Candace Sidner for use of the Glider video. 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.

References

Aikhenvald, A. Y. (2012). The essence of mirativity. *Linguistic typology*, *16*(3), 435–485.

Alikhani, M., & Stone, M. (2020). Achieving common ground in multi-modal dialogue. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 10–15.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual review of psychology*, *63*(1), 1–29.

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [Computer program] [https://www.praat.org/].

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brugman, H., Russel, A., & Nijmegen, X. (2004). Annotating multi-media/multi-modal resources with elan. *LREC*, 2065–2068.

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI magazine*, *22*(4), 67–67.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259–294.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Cutrone, P. (2010). The backchannel norms of native english speakers: A target for japanese l2 english learners. *language Studies Working papers*, *2*, 28–37.

de Kok, I., & Heylen, D. (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* (pp. 362–375). Springer.

DeLancey, S. (1997). Mirativity: The grammatical marking of unexpected information.

Fernández, R. (2022, June). Dialogue. In *The oxford handbook of computational linguistics*. Oxford University Press.

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

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634.

Heritage, J. (1984). 13. a change-of-state token and aspects of its sequential placement.

Heritage, J. (2002). Oh-prefaced responses. *The language of turn and sequence*, 196.

Hockey, B. A. (1993). Prosody and the role of okay and uh-huh in discourse. *Proceedings of the eastern states conference on linguistics*, 128–136.

Hunter, J., Asher, N., & Lascarides, A. (2015). Integrating non-linguistic events into discourse structure. *Proceedings of the 11th international conference on computational semantics*, 184–194.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Khebour, I., Brutti, R., Dey, I., Dickler, R., Sikes, K., Lai, K., Bradford, M., Cates, B., Hansen, P., Jung, C., Wisniewski, B., Terpstra, C., Hirshfield, L., Puntambekar, S., Blanchard, N., Pustejovsky, J., & Krishnaswamy, N. (2024). When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*. https://doi.org/10.5334/johd.168

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

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proc. Interspeech 2011*, 3081–3084.

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

Mazzocconi, C., Tian, Y., & Ginzburg, J. (2020). What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, *13*(3), 1302–1321.

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251–274.

Pustejovsky, J., & Krishnaswamy, N. (2021). Embodied human computer interaction. *KI-Künstliche Intelligenz*, *35*(3), 307–327.

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford University Press.

Sidner, C. (2025). *Glider video*.

Tam, C., Brutti, R., Lai, K., & Pustejovsky, J. (2023, June). Annotating situated actions in dialogue. In J. Bonn & N. Xue (Eds.), *Proceedings of the fourth international workshop on designing meaning representations* (pp. 45–51). Association for Computational Linguistics. https://aclanthology.org/2023.dmr-1.5/

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

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

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

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

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current directions in psychological science*, *16*(2), 80–84.