"Mm, Wat?" Detecting Other-initiated Repair Requests in Dialogue

Anh Ngo[1,5], Nicolas Rollet[1,2], Catherine Pelachaud[4], Chloé Clavel[1,3],

[1]ALMAnaCH, INRIA Paris , [2]Télécom Paris, SES, Institut Polytechnique de Paris, I3-CNRS,
[3]Télécom Paris, LTCI, Institut Polytechnique de Paris, [4]CNRS, ISIR, Sorbonne University,
[5]ISIR, Sorbonne University

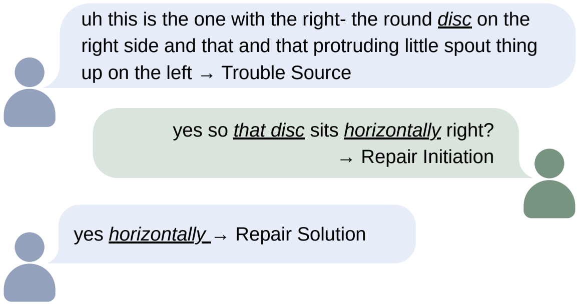
{anh.ngo-ha,nicolas.rollet,chloe.clavel}@inria.fr,catherine.pelachaud@upmc.fr

Abstract

Maintaining mutual understanding is a key component in human-human conversation to avoid conversation breakdowns, in which repair, particularly Other-Initiated Repair (OIR, when one speaker signals trouble and prompts the other to resolve), plays a vital role. However, Conversational Agents (CAs) still fail to recognize user repair initiation, leading to breakdowns or disengagement. This work proposes a multimodal model to automatically detect repair initiation in Dutch dialogues by integrating linguistic and prosodic features grounded in Conversation Analysis. The results show that prosodic cues complement linguistic features and significantly improve the results of pretrained text and audio embeddings, offering insights into how different features interact. Future directions include incorporating visual cues, exploring multilingual and cross-context corpora to assess the robustness and generalizability.

1 Introduction

Conversational agents (CAs), software systems that interact with users using natural language in written or spoken form, are increasingly being used in multiple domains such as commerce, healthcare, and education (Allouch et al., 2021). While maintaining smooth communication is crucial in these settings, current state-of-the-art (SOTA) CAs still struggle to handle conversational breakdowns. Unlike humans, who rely on conversational repair to resolve issues like mishearing or misunderstanding (Schegloff et al., 1977; Schegloff, 2000), CAs' repair capabilities remain limited and incomplete. Repair refers to the interactional effort by which participants suspend the ongoing talk to address potential trouble, which can be categorized by who initiates and who resolves it: the speaker of the trouble (self) or the co-participant (other) (Schegloff, 2000). This work focuses on **Other-initiated Self-repair**, in short, **Other-initiated Repair (OIR)**, where the **talk of a speaker is treated as problematic by a co-participant via repair initiation, and the original speaker resolves it, as illustrated in Figure 1.** Current CAs handle repairs in a limited fashion that mainly support self-initiated repair by the agent (e.g., the agent asks users to repeat what they said) (Li et al., 2020; Cuadra et al., 2021; Ashktorab et al., 2019) or rely on user self-correction when users realize troubles and clarify their own intent (e.g., saying "no, I mean. . . ") (Balaraman et al., 2023). However, CAs struggle to recognize when users signal trouble with the agent's utterances (other-initiated) and fail to provide appropriate repair (self-repaired), while effective communication requires bidirectional repair capabilities (Moore et al., 2024). Supporting this, Gehle et al. (2014) found that robots failing to resolve communication issues quickly caused user disengagement, while van Arkel et al. (2020) showed that basic OIR mechanisms improve communication success and reduce computational and interaction costs compared to relying on pragmatic reasoning.



uh this is the one with the right- the round *disc* on the right side and that and that protruding little spout thing up on the left → Trouble Source

yes so *that disc* sits *horizontally* right? → Repair Initiation

yes *horizontally* → Repair Solution

Figure 1: Other-initiated Repair (OIR) sequence example from Rasenberg et al. (2022), English translated: repair initiation (green) signals trouble of ambiguous object reference *disc* with candidate understanding *horizontally*, confirmed by repair solution *yes horizontally*.

Modeling OIR strategies on CAs that recognize user-initiated repair first requires robust automatic

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22926–22939
November 4-9, 2025 ©2025 Association for Computational Linguistics

repair initiation detection in human-human interaction. Previous work has established foundations for text-based approaches, training with English corpora, and relying on lexical cues (Höhn, 2017; Purver et al., 2018; Alloatti et al., 2024). However, prosodic cues tend to be more cross-linguistically stable than surface forms (Dingemanse and Enfield, 2015; Benjamin, 2013; Walker and Benjamin, 2017), and can provide valuable insight into the pragmatic functions of expressions like the interjection "huh". Building upon text-based methods, this work focuses on spoken dialogue interaction, where prosodic cues provide additional signals for repair initiation detection that may be missed by text-only models trained on transcriptions. Finally, understanding the OIR sequence also requires examining the local sequential environment of the surrounding turns, which we term "dialogue micro context", based on Schegloff (1987)'s work on local interactional organization.

These gaps motivate our main research question: **What are the verbal and prosodic indicators of repair initiation in OIR sequences and how can we model them?** To address this, we analyze OIR sequences in a Dutch task-oriented corpus, focusing on text and audio patterns where one speaker initiates repair. Drawing on Conversation Analysis literature, we introduce feature sets and a computational model to detect such requests. Our contributions are in two folds: (1) a novel multimodal model for detecting repair initiations in OIR sequences that integrates linguistic and prosodic features extracted automatically based on the literature, advancing beyond text- or audio-only approaches; (2) provide insights into how linguistic and prosodic features interact and contribute in detection performance, grounded in Conversation Analysis, and what causes model misclassifications. The remainings of this paper is structured as follows: Section 2 reviews SOTA computational models for OIR detection and related dialogue understanding tasks. Section 3 provides the used OIR coding schema and typology, and Section 4 details our approach, including linguistic and prosodic feature design. Section 5 presents our experiment details and results, followed by error analysis in Section 6.

2 Related Work

An early approach to automatic OIR detection was proposed by Höhn (2017), with a pattern-based chatbot handling user-initiated repair in text chats between native and non-native German speakers. Purver et al. (2018) extended this by training a supervised classifier using turn-level features in English, including lexical, syntactic, and semantic parallelism between turns. More recently, Alloatti et al. (2024) introduced a hierarchical tag-based system for annotating repair strategies in Italian task-oriented dialogue, distinguishing between utterance-specific and context-dependent functions. Closely related, Garí Soler et al. (2025)'s recent work introduced and investigated the task of automatically detecting word meaning negotiation indicators, where speakers signal a need to clarify or challenge word meanings, a phenomenon that can be seen as a specific form of repair initiation.

Although direct research on OIR detection is still limited, advances in related dialogue understanding tasks provide promising foundations for our work. Miah et al. (2024) combined pretrained audio (Wav2Vec2) and text (RoBERTa) embeddings to detect dialogue breakdowns in healthcare calls. Similarly, Huang et al. (2023) used BERT, Wav2Vec2.0, and Faster R-CNN for intent classification, introducing multimodal fusion with attention-based gating to balance modality contributions and reduce noise. Saha et al. (2020) proposed a multimodal, multitask network jointly modeling dialogue acts and emotions using attention mechanisms. More recently, high-performing but more opaque and resource-intensive approaches have emerged, such as Mohapatra et al. (2024) showed that larger LLMs outperform smaller ones on tasks like repair and anaphora resolution, albeit with higher computational cost and latency.

Despite robust performance, recent largest models remain difficult to interpret due to their black-box nature and multimodal fusion complexity (Jain et al., 2024). To address this gap, we propose a computational model for repair initiation detection in Dutch spoken dialogue that fuses pretrained text and audio embeddings with linguistic and prosodic features **grounded in Conversation Analysis**. The model also integrates a multihead attention mechanism to weigh and capture nonlinear relationships across modalities, allowing our model to keep the strengths of multimodal deep learning while **offering insight from linguistic and prosodic features to understand their interaction and impact** towards model's decision.

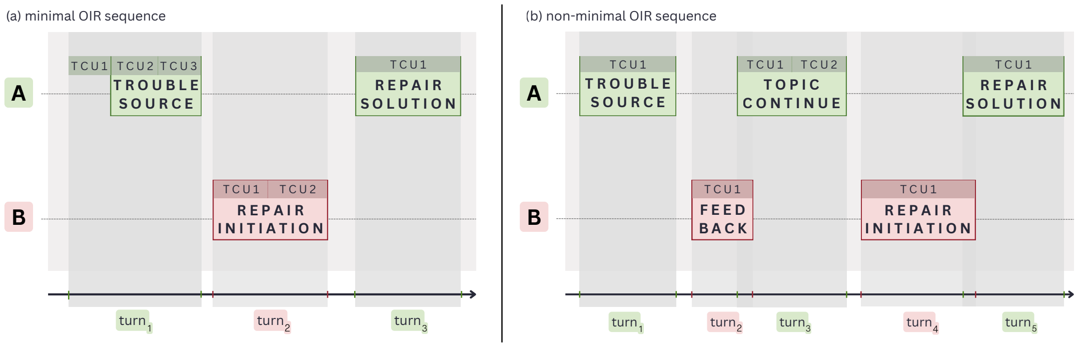


Figure 2: OIR sequence organization between 2 speakers A (green) and B (red): (a) Minimal; (b) Non-minimal

3 OIR Coding Schema and Typology

We follow Dingemanse and Enfield (2015)'s coding schema, which structures OIR sequences into three components: trouble source, repair initiation, and repair solution segments, with repair initiation categorized as *open request* (the least specific, not giving clues of trouble), *restricted request* (implied trouble source location), or *restricted offer* (the most specific, proposing a candidate understanding). Throughout this work, repair initiation refers specifically to this component within OIR sequences. We use the corpus and the OIR sequences annotation from Rasenberg et al. (2022), where dialogues were manually transcribed and segmented into Turn Construction Units (TCUs), the smallest meaningful elements of speech that can potentially complete a speaker turn. They align OIR component boundaries with these pre-existing TCU boundaries. Following the conversational analysis practice, such as in (Mondada, 2018), we adopt the "segment" as our unit of analysis, defined as: stretches of talk corresponding to annotated OIR components (e.g., repair initiation) that may span one or more TCUs within larger speaker turns (illustrated in Figure 2). This allows us to target only the stretch of talk relevant to the OIR component, avoiding the overinclusiveness of full turns. Figure 2 illustrates two organizational scenarios of OIR sequences described in Dingemanse and Enfield (2015), including: *minimal* (repair initiation produced immediately after the turn containing the trouble source) and *non-minimal* (repair initiation delayed by a few turns).

4 Proposed Approach

4.1 Overview

Task Formulation. We formulate the repair initiation detection task as a binary classification problem. Given a segment (x_i), corresponding to one or several TCUs within a speaker turn, the task is to predict whether it is an OIR repair initiation or a regular dialogue (RD) segment (*i.e.*, not belonging to an OIR sequence). In this initial study, we limit the scope to detecting repair initiations only, without classifying other OIR components such as trouble sources or repair solutions. This simplification allows us to establish a baseline for the most critical component in the OIR sequence, the moment when repair is initiated by another speaker.

Architecture Overview. Figure 3 shows our proposed multimodal approach for repair initiation detection. We incorporate the handcrafted linguistic and prosodic features, automatically computed based on literature reviews, with embeddings from pretrained models (RobBERT for text, Whisper for audio). For a given segment (x_i), we first extract both handcrafted features and pretrained embeddings from text and audio modalities. All features are then projected to a shared dimensionality to ensure consistency across modalities. To capture the complex interactions between text and audio embeddings with handcrafted features, a multihead attention mechanism was employed to weigh and capture nonlinear relationships. Finally, the whole representation is obtained by concatenating the text embedding and the fused representation from multihead attention.

4.2 Pretrained Models

Language model. Our proposed approach utilizes BERT (Devlin et al., 2019), a transformer-based language model, to obtain text embedding of the current given segment. As our corpus is in Dutch, we use the pretrained RobBERT (Delobelle et al., 2020) model, which is based on the BERT architecture, pretrained with a Dutch tokenizer, and

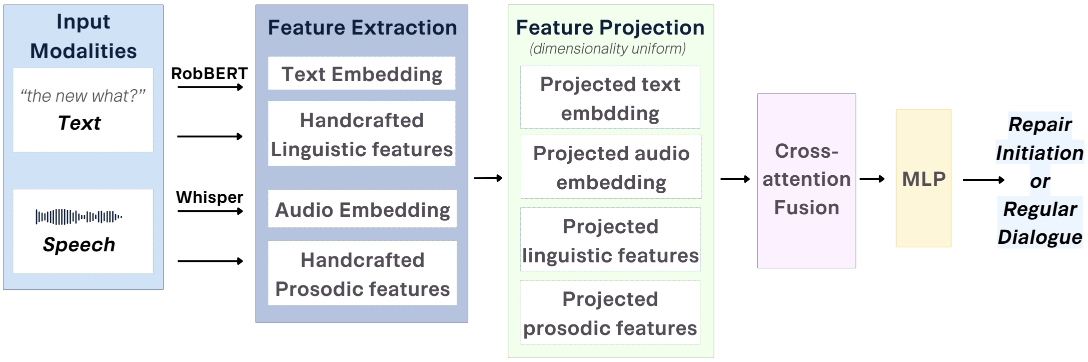


Figure 3: Multimodal architecture for repair initiation detection

39 GB of training data. We use the latest release of *RobBERT-v2-base* model which pretrained on the Dutch corpus OSCAR 2023 version, which outperforms other BERT-based language models for several different Dutch language tasks.

Audio model. For audio representations, we utilize Whisper (Radford et al., 2023), an encoder-decoder transformer-based model trained on 680,000 hours of multilingual and multitask speech data, to extract audio embeddings from our dialogue segments. Whisper model stands out for its robustness in handling diverse and complex linguistic structures, a feature that is crucial when dealing with Dutch, a language known for its intricate syntax. Besides, Whisper was trained on large datasets including Dutch and demonstrated good performance in zero shot learning, making it ideal serving as a naive baseline for task with small corpus like ours.

4.3 Dialogue Micro Context

Schegloff (1987) demonstrated that the OIR sequence is systematically associated with multiple organizational aspects of conversation, and understanding an OIR repair initiation requires examining the local sequential environment, which he terms the micro context, that we adopt in this work. Therefore, for each given target segment x_i, to capture the micro context, we iteratively concatenate the previous (x_{i-j}) and following (x_{i+j}) segment within a window of size (j), using special separator token of transformers (e.g. [SEP] for BERT-based models) until reaching the maximum token limit (excluding [CLS] and [EOS]), inspired by similar ideas in (Wu et al., 2020). If the sequence exceeds the limit, we truncate the most recently added segments. The final sequence is enclosed with [CLS] and [EOS], as shown in Figure 9 (Appendix D).

4.4 Linguistic Feature Extraction

Figure 4(a) outlines our linguistic feature set for the representation of the target segment, capturing local properties such as part-of-speech (POS) tagging patterns, question formats, transcribed non-verbal actions (target segment features), and features, which quantify repetition and coreference across turns to reflect backward and forward relations around the repair initiation (cross-segment feature to capture micro context). The detailed description is in the Appendix E.

4.4.1 Target Segment Features

We automatically extracted the linguistic features proposed by (Ngo et al., 2024) at the intra-segment level to capture grammatical and pragmatic patterns related to the repair initiation. For instance, (Ngo et al., 2024) shows that restricted requests often show a POS tag sequence pattern of interrogative pronouns followed by verbs, while OIR open requests and regular dialogue segments differ in key lemmas used of the same tag: modal auxiliary verb kunnen ("can") vs. primary auxiliary verb zijn ("to be"). We also include question mark usage, derived from original transcription, which is marked with a question mark if the annotator detected question prosody. It implicitly reflects prosodic cues as interpreted by the human annotators, which are relevant to repair initiation, as described in Schegloff (2000) regarding interrogative and polar question formats. A complete list of features is fully provided in Appendix E.

4.4.2 Cross-Segment Features

Grounded on the literature (Schegloff, 2000; Ngo et al., 2024), we define inter-segment features that capture the sequential dynamics of the repair initiation, including repetitions and the use of coreferences referring to entities in prior turns containing

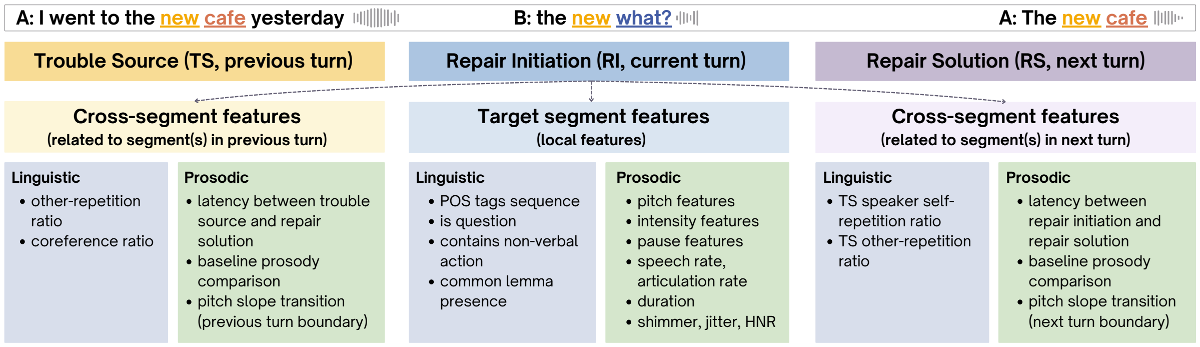

A: I went to the new cafe yesterday
B: the new what?
A: The new cafe
Trouble Source (TS, previous turn)
Repair Initiation (RI, current turn)
Repair Solution (RS, next turn)
Cross-segment features
(related to segment(s) in previous turn)
Target segment features
(local features)
Cross-segment features
(related to segment(s) in next turn)
Linguistic
• other-repetition ratio
• coreference ratio
Prosodic
• latency between trouble source and repair solution
• baseline prosody comparison
• pitch slope transition (previous turn boundary)
Linguistic
• POS tags sequence
• is question
• contains non-verbal action
• common lemma presence
Prosodic
• pitch features
• intensity features
• pause features
• speech rate, articulation rate
• duration
• shimmer, jitter, HNR
Linguistic
• TS speaker self-repetition ratio
• TS other-repetition ratio
Prosodic
• latency between repair initiation and repair solution
• baseline prosody comparison
• pitch slope transition (next turn boundary)

Figure 4: Handcrafted linguistic and prosodic features design

the trouble source segment. We also compute self and other-repetition in the subsequent turn containing the repair solution segment, to capture how the trouble source speaker responds. These features reflect the global dynamics of OIR sequences.

4.5 Prosodic Features Extraction

Prosody plays a crucial role in signaling repair initiation. Previous studies in Conversation Analysis show that pitch, loudness, and contour shape can indicate whether repair initiation is perceived as "normal" or expresses "astonishment"(Selting, 1996), and that Dutch question types differ in pitch height, final rises, and F0 register (Haan et al., 1997). Building upon these characteristics, we design a prosodic feature set that includes both local features within the target segment, such as pitch, intensity, pauses, duration, and word-level prosody, and global features across segments of the OIR sequence, such as latency between OIR sequence segments, pitch slope transitions at boundaries, and comparison to speaker-specific prosodic baselines. The features are detailed in Figure 4(b) and in the Appendix F.

4.5.1 Target Segment Features

We use Praat (Boersma, 2000) to extract prosodic features at the segment level, including: pitch features (e.g., min, max, mean, standard deviation, range, number of peaks) which are computed from voiced frames after smoothing and outlier removal, with pitch floor/ceiling set between 60–500 Hz and adapted to each speaker range (van Bezooijen, 1995; Theelen, 2017; Verhoeven and Connell, 2024); first (mean and variability of pitch slope change) and second derivatives (pitch acceleration) of pitch contour, capturing pitch dynamics. Additional features are intensity (e.g., min, max, mean, range, standard deviation), and voice quality measures (jitter, shimmer, and harmonics-to-noise ratio). We also model pause-related features by detecting silent pauses over 200 ms and categorizing them by duration and position in the utterance, reflecting their conversational function associated with repair possibilities (van Donzel and Beinum, 1996; Hoey, 2018). Inspired by findings about prosody of other-repetition in OIR sequences (Dingemanse et al., 2015; Walker and Benjamin, 2017), we extract pitch and intensity features for repeated words from the trouble source segment, and for the specific repair marker "wat" (what/which/any), as indicators of repair initiation type and speaker perspective (Huhtamäki, 2015).

4.5.2 Cross-Segment Features

To model the speaker-specific prosodic variation (van Bezooijen, 1995; Theelen, 2017; Verhoeven and Connell, 2024), we normalize pitch and intensity using z-scores, relative percentage change, and position within the speakers' range. These features capture how far the current segment deviates from the speaker's typical behaviour across previous turns and the normalized range position of the current segment within the speaker's baseline. Inspired by work on prosodic entrainment (Levitan and Hirschberg, 2011), we also compute pitch and intensity slope transitions across segment boundaries (e.g., TS→OIR, OIR→RS), both within and across speakers, to assess prosodic alignment. We normalized slopes to semitones per second for consistency across speakers.

5 Experiments & Results

To answer the main research question mentioned in Section 1, we design the experiments to answer the following research sub-questions: *i)* **RQ1**: To what extent do audio-based features complement text-based features in identifying repair initiation?

Model	Modal & Features	Precision	Recall	F1-score
Text_{Emb}	U & T	72.0 ± 4.0	87.6 ± 7.5	78.9 ± 4.7
$\text{Audio}_{\text{Emb}}$	U & A	72.6 ± 9.7	76.3 ± 13.1	70.6 ± 8.1
$\text{Multi}_{\text{Emb}}$	M & T+A	79.1 ± 5.4	82.2 ± 3.8	82.1 ± 0.9
$\text{Text}_{\text{Ling}}$	U & L	82.2 ± 3.6	80.4 ± 6.1	80.4 ± 3.8
$\text{Audio}_{\text{Pros}}$	U & P	81.7 ± 4.2	77.4 ± 5.4	77.3 ± 2.7
$\text{Multi}_{\text{LingPros}}$	M & L+P	81.7 ± 7.6	82.2 ± 1.5	81.8 ± 3.4
$\text{Multi}_{\text{Ours}}$	M & T+A+L+P	$\mathbf{93.2 \pm 2.8}$	$\mathbf{96.1 \pm 2.6}$	$\mathbf{94.6 \pm 2.3}$

U: Unimodal, **M**: Multimodal, **T**: Text, **A**: Audio, **P**: Prosodic features, **L**: Linguistic features

Table 1: Overall results across modalities for repair initiation detection. The table groups models by research question: **RQ1** compares unimodal vs. multimodal combinations of audio and text; **RQ2** compares handcrafted features with pretrained embeddings.

ii) **RQ2**: Do our proposed linguistic and prosodic features (see Figures 4(a) and 4(b)) perform better than pretrained embeddings? *iii)* **RQ3**: Which prosodic and linguistic features contribute the most to repair initiation detection? *iv)* **RQ4**: How does the involvement of dialogue micro context affect detection performance?

5.1 Implementation Details

Dataset. Based on (Colman and Healey, 2011)'s finding that repair occurs more frequently in task-oriented dialogues, we selected a Dutch multimodal task-oriented corpus (Rasenberg et al., 2022; Eijk et al., 2022), containing 19 dyads collaborating on referential communication tasks in a standing face-to-face setting. For each round, participants alternated roles to describe (Director) or identify (Matcher) a geometric object (called "Fribbles") displayed on screens, in which the unconstrained design encouraged natural modality use and OIR sequences. Rasenberg et al. (2022) annotated OIR sequences using Dingemanse and Enfield, 2015's schema, resulting in 10 open requests, 31 restricted requests, and 252 restricted offers. While we acknowledge that OIR sequences are rarer in natural dialogue, our goal in this paper is to study detection performance with sufficient examples of both classes. Therefore, we balanced the dataset with 306 randomly selected regular dialogue segments, stratified across all dyads, resulting in 712 samples overall. The data were split 70:15:15 for training, validation, and testing. Limitations regarding the generalizability of the artificial balancing are discussed in Section 7. Examples of Fribbles objects and repair initiation types are provided in the Appendix A and B.

Training Details. We fine-tuned our models using 10-fold cross-validation, in which the optimal learning rate was 2e-5. We employed AdamW optimizer with weight decay of 0.01 and a learning rate scheduler with 10% warm-up steps. Training ran for up to 20 epochs with 3-epoch early stopping patience, and batch size 16. The source code is publicly available [1].

Evaluation Metrics. We evaluated model performance using binary classification metrics including precision, recall, and macro F1-score.

5.2 Experiment Scenarios & Results Analysis

RQ1: Audio vs. Text Complementarity. To address RQ1, we compare the performance of unimodal against multimodal models, including: *i)* Single $\textbf{Text}_{\textbf{Emb}}$ or $\textbf{Audio}_{\textbf{Emb}}$ vs. $\textbf{Multi}_{\textbf{Emb}}$; *ii)* Single $\textbf{Text}_{\textbf{Ling}}$ or $\textbf{Audio}_{\textbf{Pros}}$ *vs.* $\textbf{Multi}_{\textbf{LingPros}}$. We examine whether integrating the audio-based features, either by pretrained embeddings or by using handcrafted prosodic features, will improve the performance of the text-based models. The multimodal models include $\textbf{Multi}_{\textbf{Emb}}$, which fuses pretrained text and audio embeddings, and $\textbf{Multi}_{\textbf{LingPros}}$, which combines handcrafted linguistic and prosodic features, using cross-attention fusion as illustrated in Figure 3.

From Table 1, we observe that multimodal models consistently outperform unimodal ones across all metrics. For both pretrained embeddings and handcrafted features, text-based models outperform audio-based ones individually. However, incorporating audio improves performance in both settings. Specifically, in the pretrained setting,

[1] https://github.com/haanh764/other_initiated_repair_detection

the multimodal model **Multi$_{\mathbf{Emb}}$** achieves an F1-score of 82.1, improving over **Text$_{\mathbf{Emb}}$** by 3.2 percentage points (pp) and over **Audio$_{\mathbf{Emb}}$** by 11.5 pp. Similarly, in the handcrafted feature setting, combining linguistic and prosodic features **Multi$_{\mathbf{LingPros}}$** yields an F1 of 81.8, outperforming Text$_{\text{Ling}}$ by 1.4 pp and **Audio$_{\mathbf{Pros}}$** by 4.5 pp. Interestingly, the unimodal handcrafted models **Text$_{\mathbf{Ling}}$**, **Audio$_{\mathbf{Pros}}$** show higher precision than recall, whereas **Multi$_{\mathbf{LingPros}}$** shows slightly higher recall, suggesting a tendency to favor detection over omission. This is potentially beneficial in interactive systems where missing an repair initiation could be more disruptive than a false alarm. For embedding-based models, recall exceeds precision in all cases, but the multimodal model shows a notable gain in precision, indicating a better trade-off between identifying true repair initiation and minimizing false positives.

RQ2: Handcrafted Features vs. Pretrained Embeddings. To address RQ2, we compare the performance of models using handcrafted features against the models using embeddings from pretrained models. We thus compare: *i)* Text representations: text embeddings (**Text$_{\mathbf{Emb}}$**) *vs.* handcrafted linguistic features (**Text$_{\mathbf{Ling}}$**); *ii)* Audio representations: audio embeddings (**Audio$_{\mathbf{Emb}}$**) *vs.* handcrafted prosodic features (**Audio$_{\mathbf{Pros}}$**); *iii)* Combined approaches: multimodal models using pretrained embeddings (**Multi$_{\mathbf{Emb}}$**) *vs.* using handcrafted linguistic and prosodic features (**Multi$_{\mathbf{LingPros}}$**) and *vs.* our proposed approach leveraging both of them **Multi$_{\mathbf{Ours}}$**.

Table 1 shows that handcrafted feature models are comparable to embedding-based approaches. In unimodal settings, Text$_{\text{Ling}}$ achieves higher precision (+10 pp) with comparable F1-score (+1.5 pp) to **Text$_{\mathbf{Emb}}$**, despite lower recall (-7.2 pp). Likewise, **Audio$_{\mathbf{Pros}}$** outperforms **Audio$_{\mathbf{Emb}}$** across all metrics (precision +9.1 pp, recall +1.1 pp, F1-score +6.7 pp). In multimodal settings, **Multi$_{\mathbf{Emb}}$** and **Multi$_{\mathbf{LingPros}}$** perform nearly identically (F1-score difference of 0.3 pp). Overall, we observe a general trend emerges: embedding-based approaches tend to achieve higher recall but lower precision, likely because they can learn more complex representation that captures more subtle patterns, whereas handcrafted feature models target specific repair initiation markers, such as question forms, repetition, and pause patterns, resulting in better balanced precision-recall trade-offs. The embedding models may also overgeneralize in the case of our small, task-specific corpus.

RQ3: Handcrafted Feature Importance Analysis. Although the linguistic and prosodic features could not solely outperform pretrained text and audio embeddings, they are useful in interpreting the model's behaviours, especially to see if they are aligned with the Conversation Analysis findings. To answer RQ3, we used SHAP (SHapley Additive exPlanations) analysis to analyze the contribution and behaviours of linguistic and prosodic features towards the model's decision. Figure 5 illustrates the top 10 features by SHAP value, which measures how much each single feature pushed the model's prediction compared to the average prediction. The pausing behaviours (positions and durations), intensity measures (max, mean, and relative change), and harmonic-to-noise ratio (HNR) appear particularly important among prosodic features. For linguistic features, the grammatical structure linking to coreference used, some POS tags, and various word type ratios rank highly, which align well with systematic linguistic patterns, as demonstrated by Ngo et al. (2024). The most important features include the number of long and medium pauses, the relative position of the longest pause, and the verb-followed-by-coref structure, all scoring near 1.0 on the importance scale, which aligned with the works in (Hoey, 2018; Ngo et al., 2024) about pauses in repair initiation and its structure, respectively.

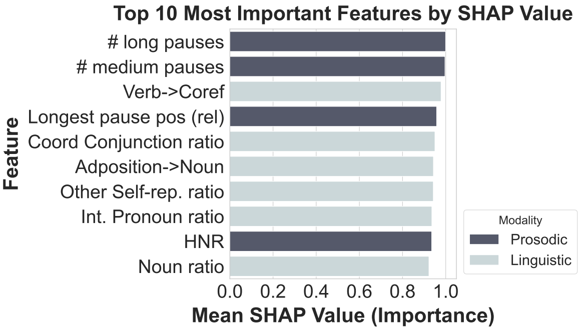


Figure 5: The top 10 most important handcrafted features ranked by SHAP value. Appendix C provides the full list of the 20 most contributed features.

Figure 6 displays the synergy (Ittner et al., 2021) between linguistic and prosodic features, computed based on the SHAP interaction values. It reflects how complementary a pair of linguistic and prosodic features is in improving model performance, in which high synergy means that combining both features adds more value than what each

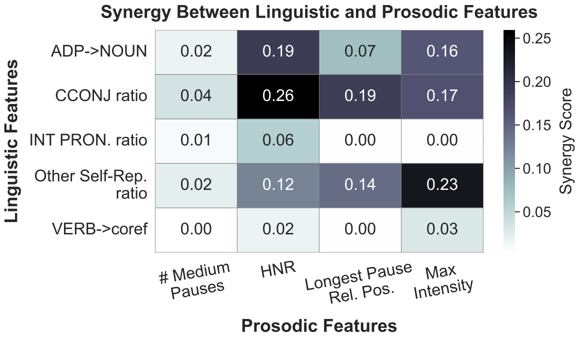


Figure 6: Handcrafted feature interaction analysis: Linguistic vs Prosodic

of them contributes individually. These features do not always need to co-vary, but their combination brings useful information for the model. Coordinating conjunction ratio (CCONJ ratio) shows the strongest synergy (0.26) with harmonics-to-noise ratio (HNR), while other speaker self-repetition ratio has strong synergy (0.23) with maximum intensity. This suggests that certain grammatical patterns work closely with specific voice qualities, particularly how conjunctions interact with voice clarity and how self-repetition correlates with voice intensity. The results indicate that conversation involves a complex interplay between what we say (linguistic elements) and how we say it (prosodic elements), which is aligned with the Conversation Analysis work.

RQ4: Dialogue Micro Context Analysis. To address RQ4, we experimented 4 scenarios of concatenating micro context, including: (1) **Past$_{\text{Context}}$** - concatenated current input segment with the segments in the prior turns and cross-segment handcrafted features (past-related, Figure 4); (2) **Future$_{\text{Context}}$** - concatenated current input segment with the segments in the subsequent turns and handcrafted cross-segment features (future-related, Figure 4); (3) **Current$_{\text{Context}}$** - no context concatenation and used only current input segment features (Figure 4); (4) **Multi$_{\text{Ours}}$** - the full context scenario, where we concatenate current input segment with both the prior and subsequent segments and use full handcrafted feature set. For (1) and (4), we experimented with window_length of 2 and max (the micro context are concatenated as much as possible until it reach maximum token limit) based on results from corpus analysis; for (3) only max was used, as repair solutions typically occur immediately within maximum 2 turns in this corpus.

Context	Win. len	Precision	Recall	F1-score
(1) Past$_{\text{Context}}$	2	86.0 ± 3.0	78.4 ± 5.4	82.0 ± 4.1
(1) Past$_{\text{Context}}$	max	86.6 ± 5.2	81.0 ± 6.1	83.5 ± 4.3
(2) Current$_{\text{Context}}$	-	84.6 ± 3.8	82.9 ± 6.0	83.6 ± 4.4
(3) Future$_{\text{Context}}$	max	84.00 ± 1.53	78.20 ± 5.78	80.18 ± 2.52
(4) Multi$_{\text{Ours}}$	2	**93.2 ± 2.8**	**96.1 ± 2.6**	**94.6 ± 2.3**
(4) Multi$_{\text{Ours}}$	max	87.7 ± 3.5	89.1 ± 5.3	88.3 ± 3.7

Table 2: Performance comparison across different micro context configurations

Table 2 highlights the impact of different micro context configurations, in which incorporating surrounding segments from prior, and subsequent segments combining with the whole handcrafted feature set leads to the best overall performance, as also stated in Table 1. Notably, our full context setting with smaller window_length=2 achieves the highest results across all metrics, while concatenating to the maximum allowed token limits degrades the performance, with a drop of approximately 6.3 pp of F1-score, 9 pp of precision, and 4.1 pp of recall. It suggests that while surrounding context of input segment is helpful, overly long concatenation may introduce noise and irrelevant information to the model. In addition, integrating past or solely current segments yields moderate performance, with F1-scores ranging from approximately 80.2% to 83.6%, while future context integration results in the lowest scores, indicating that the upcoming dialogue can offer informative cues but less relevant than the prior and current input segments, which aligned with the nature of OIR sequence.

6 Error Analysis

To better interpret model performance, we analyze the False Negative (FN) instances, which are repair initiations that were misclassified as regular dialogue, to identify whether there are common patterns in these instances that our models struggle to predict, illustrated in Table 3. We compare these FN instances across our proposed multimodal model with the unimodal baselines by extracting representative dialogue samples for each model from test set and identifying their common linguistic and prosodic characteristics.

Our proposed model shows the lowest FN rate (3.8%) of the test set, compared to 15% and 24% on **Text$_{\text{Ling}}$** and **Audio$_{\text{Pros}}$**, respectively. **Text$_{\text{Ling}}$** seems to struggle in detecting samples with vague references, especially in restricted offers, even when OIR syntactic forms like *question mark* is present. Besides, **Audio$_{\text{Pros}}$** tends to over-rely on pause structure and pitch contour even though important prosodic cues were presented. Short declar-

Model	%Error	Samples	Patterns	OIR Type
Text$_{\text{Ling}}$	15%	*(or a) triangle*	Vague, elliptical reference	RO
		yes uh yes on the right side right? or ascending yes	Disfluencies, vague interrogative	RO
		yes the one with the protrusion	Referential expression, lacks direct marker	RO
Audio$_{\text{Pros}}$	24%	*with a sunshade*	Short declarative, flat prosody	RO
		uh but the platform sits that cuts the	Flat intonation, short pauses in beginning	RO
		Is it vertical?	Question intonation, few short pauses	RO
		ah and is his arm uh round but also a bit with angles?	High pitch, question intonation, pauses mid-turn	RO
		but what did you say at the beginning?	Rising intonation, wide pitch range	RR
Multi$_{\text{Ours}}$	3.8%	*with a sunshade*	Short, declarative structure	RO
		oh who so	Declarative, high but flat pitch	RO
		sorry again?	Clear OIR but subtle prosodic signal	OR

Table 3: Samples of False Negative (FN) instances from unimodal and multimodal models with qualitative patterns. OR: open request; RR: restricted request; RO: restricted offer. The Dutch samples are translated to English by DeepL.

atives with flat intonation were often misclassified, suggesting the impact of missing syntactic form information in this model. Finally, our proposed multimodal failed with mostly short phrases and subtle prosodic signals, which are not strongly marked as an repair initiation. Considering the error across 3 types of repair initiations, it seems that only **Audio$_{\text{Pros}}$** struggled with various types of repair initiations; the other 2 models misclassified on restricted offer and open request instances only. However, as this corpus is imbalanced between the 3 types of repair initiation, with a majority of restricted offers, it could be the potential reason.

7 Conclusion & Future Works

This work presents a novel approach for detecting repair initiation in Other-Initiated Repair (OIR) sequences within human-human conversation. It leverages automatically extracted linguistic and prosodic features grounded in Conversation Analysis theories. Our results demonstrate that incorporating handcrafted features significantly enhances detection performance compared to using only pretrained embedding models. Additionally, audio modality complements textual modality, improving detection performance across both pretrained embeddings and handcrafted features. Handcrafted feature analysis revealed both individual impact and complementary contributions between modalities. Key prosodic indicators include pause-related features, intensity, and harmonic-to-noise ratio (HNR), while important linguistic features involve grammatical patterns, POS tags, and lemma ratios. Synergy analysis demonstrates that features do not act independently; for example, coordinating conjunction usage shows strong synergy with HNR, and trouble source speaker self-repetition leads significantly to maximum intensity presence. These patterns highlight the nature of OIR sequences, in which how something is said modulates what is being said.

Our results also highlight the importance of dialogue micro context in repair initiation detection: models using both prior and subsequent segments outperform those relying only on the target segment, reflecting the interactional structure crucial for OIR interpretation. However, overusing context can add noise and degrade performance.

Finally, error analysis revealed that while the text-based model failed with vague references and disfluencies, the audio-based model was prone to misclassifying flat or subtle prosodic cues, which raised the need for a multimodal model. The proposed multimodal model mitigates these weaknesses, but it still struggles with short, minimally marked repair initiation that lacks both strong syntactic and prosodic cues. This work establishes foundations for conversational agents capable of detecting human repair initiation to avoid communication breakdowns.

Building on these insights, future work will explore the integration of visual features to more accurately model the embodied aspects of OIR sequences, as well as the development of multilingual and cross-context corpora to assess the robustness and generalizability of the detection approach.

Limitations

Dataset Limitations and Generalizability. Due to the limited multimodal OIR-labeled corpora, our study utilized the only available multimodal OIR-labeled corpus, which is specific to Dutch language and referential object matching tasks. This specificity could limit the generalizability of our model across different OIR categories, languages, and conversation settings. Future works should test the model on more diverse datasets to validate its robustness and establish broader applicability.

Dataset Balancing and Class Distribution. In natural conversation, repair initiation instances are much less frequent than regular dialogue. To enable robust model training and evaluation, we balanced repair initiation and regular dialogue samples across dyads. However, this balancing approach may affect the model's performance in real-world settings where OIR sequences are rare, and therefore, the results should be interpreted with caution. Future work should evaluate the performance of models while maintaining the natural class distribution to assess practical applicability.

Adaptability in Real-time Processing. Despite the computational efficiency of our approach using handcrafted features compared to Large Language Models, several limitations remain for real-time adaptation. The feature extraction of some linguistic and prosodic features, such as coreference chains, requires additional computation with pre-trained models, potentially introducing latency. Future work should explore real-time feature extraction pipelines and incremental processing architectures, while evaluating potential trade-offs between model complexity and real-time performance to make the system practical for CA systems.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. Data were provided (in part) by the Radboud University, Nijmegen, The Netherlands. This work has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. It was also partially funded by the ANR-23-CE23-0033-01 SINNet project.

References

Francesca Alloatti, Francesca Grasso, Roger Ferrod, Giovanni Siragusa, Luigi Di Caro, and Federica Cena. 2024. A tag-based methodology for the detection of user repair strategies in task-oriented conversational agents. *Computer Speech & Language*, 86:101603.

Merav Allouch, A. Azaria, and Rina Azoulay-Schwartz. 2021. Conversational agents: Goals, technologies, vision and challenges. *Sensors (Basel, Switzerland)*, 21.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Vevake Balaraman, Arash Eshghi, Ioannis Konstas, and Ioannis Papaioannou. 2023. No that‘s not what I meant: Handling third position repair in conversational question answering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 562–571, Prague, Czechia. Association for Computational Linguistics.

Trevor Michael Benjamin. 2013. *Signaling trouble: on the linguistic design of other-initiation of repair in English conversation*. Ph.D. thesis. Relation: http://www.rug.nl/ Rights: University of Groningen.

Paul Boersma. 2000. A system for doing phonetics by computer. 5.

Marcus Colman and Patrick G. T. Healey. 2011. The distribution of repair in dialogue. *Cognitive Science*, 33.

Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: Towards a typology of conversational structures.

Mark Dingemanse, Seán G Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S Gisladottir, Kobin H Kendrick, Stephen C Levinson, Elizabeth Manrique, and 1 others. 2015. Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F. Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris van Rooij, Herbert Schriefers, Ivan Toni, James Trujillo, and Sara Bögels. 2022. The cabb dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2025. Toward the automatic detection of word meaning negotiation indicators in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. To appear.

Raphaela Gehle, Karola Pitsch, and Sebastian Benjamin Wrede. 2014. Signaling trouble in robot-to-group interaction.emerging visitor dynamics with a museum guide robot. *Proceedings of the second international conference on Human-agent interaction*.

Judith Haan, Vincent Van Heuven, Jos Pacilly, and R.L. Bezooijen. 1997. An anatomy of dutch question intonation. *J. Coerts & H. de Hoop (eds.), Linguistics in the Netherlands 1997, 97 - 108 (1997)*, 14.

Elliott Hoey. 2018. How speakers continue with talk after a lapse in conversation. *Research on Language and Social Interaction*, 51.

Sviatlana Höhn. 2017. A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 395–405, Saarbrücken, Germany. Association for Computational Linguistics.

Xuejian Huang, Tinghuai Ma, Li Jia, Yuanjian Zhang, Huan Rong, and Najla Alnabhan. 2023. An effective multimodal representation and fusion method for multimodal intent recognition. *Neurocomputing*, 548:126373.

Martina Huhtamäki. 2015. The interactional function of prosody in repair initiation: Pitch height and timing of va 'what' in helsinki swedish. *Journal of Pragmatics*, 90:48–66.

Jan Ittner, Lukasz Bolikowski, Konstantin Hemker, and Ricardo Kennedy. 2021. Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition. *ArXiv*, abs/2107.12436.

D. Jain, Anil Rahate, Gargi Joshi, Rahee Walambe, and K. Kotecha. 2024. Employing co-learning to evaluate the explainability of multimodal sentiment analysis. *IEEE Transactions on Computational Social Systems*, 11:4673–4680.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. pages 3081–3084.

Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1094–1107, New York, NY, USA. Association for Computing Machinery.

Md Messal Monem Miah, Ulie Schnaithmann, Arushi Raghuvanshi, and Youngseo Son. 2024. Multimodal contextual dialogue breakdown detection for conversational ai models. *ArXiv*, abs/2404.08156.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Lorenza Mondada. 2018. Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction*, 51(1):85–106.

Robert J. Moore, Sungeun An, and Olivia H. Marrese. 2024. Understanding is a two-way street: User-initiated repair on agent responses and hearing in conversational interfaces. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Anh Ngo, Dirk Heylen, Nicolas Rollet, Catherine Pelachaud, and Chloé Clavel. 2024. Exploration of human repair initiation in task-oriented dialogue: A linguistic feature-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 603–609, Kyoto, Japan. Association for Computational Linguistics.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Marlou Rasenberg, Wim Pouw, Asli Özyürek, and Mark Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, 12.

Tulika Saha, Aditya Patra, S. Saha, and P. Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. pages 4361–4372.

Emanuel A. Schegloff. 1987. Between micro and macro: contexts and other connections. In Richard Munch Jeffrey C. Alexander, Bernhard Giesen and Neil J. Smelser, editors, *The Micro-Macro Link*, page 207–234. University of California Press, Berkeley.

Emanuel A. Schegloff. 2000. When 'others' initiate repair. *Applied Linguistics*, 21:205–243.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361.

Margret Selting. 1996. *Prosody as an activity-type distinctive cue in conversation: the case of so-called 'astonished' questions in repair initiation*, page 231–270. Studies in Interactional Sociolinguistics. Cambridge University Press.

Mathilde Theelen. 2017. Fundamental frequency differences including language effects. *Junctions: Graduate Journal of the Humanities*, 2:9.

Jacqueline van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 177–194, Online. Association for Computational Linguistics.

Reneé van Bezooijen. 1995. Sociocultural aspects of pitch differences between japanese and dutch women. *Language and Speech*, 38(3):253–265. PMID: 8816084.

Monique van Donzel and Florien Beinum. 1996. Pausing strategies in discourse in dutch. pages 1029 – 1032 vol.2.

Jo Verhoeven and Bruce Connell. 2024. Intrinsic vowel pitch in hamont dutch: Evidence for if0 reduction in the lower pitch range. *Journal of the International Phonetic Association*, 54(1):108–125.

Traci Walker and Trevor Benjamin. 2017. Phonetic and sequential differences of other-repetitions in repair initiation. *Research on Language and Social Interaction*, 50(4):330–347.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

A Dataset Details

Figure 7 presents samples of 16 geometrical objects called "Fribbles" displayed on the participants' screens. Each dyad completed 6 rounds per session, resulting in 96 trials total. In each trial, participants alternated between Director and Matcher roles: the Director described a highlighted Fribble while the Matcher identified and confirmed the corresponding object by naming it loudly before proceeding to the next trial.

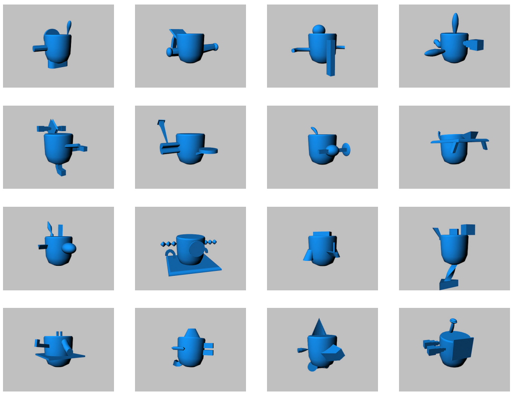


Figure 7: 16 "Fribbles" were used in the object matching task (Rasenberg et al., 2022; Eijk et al., 2022).

B OIR Types Examples

Example 1. Open request sample

TS SPEAKER: op dat driehoek (TS)
(on that triangle)
REPAIR INITIATOR: wat zei je? (RI)
(what did you say?)
TS SPEAKER: op die driehoek (RS)
(on that triangle)

Example 2. Restricted request sample

TS SPEAKER: deze heeft twee oren die aan de onderkant breder worden en een soort hanekam op zijn hoofd een kleintje (TS)
(this one has two ears that widen at the bottom and a sort of cock's comb on its head a little one)
REPAIR INITIATOR: maar wat zei wat zei je in het begin? (RI)
(but what did you say at the beginning?)
TS SPEAKER: een soort oren die aan de onderkant breder worden (RS)
(a kind of ears that widen at the bottom)

Example 3. Restricted offer sample

TS SPEAKER: waarbij je dus op de bovenkant zo'n zo'n mini uh kegeltje hebt (TS)
(where you have one of those mini uh cones on the top)
REPAIR INITIATOR: oh ja die zo scheef naar achter staat? (RI)
(oh yes which is so slanted backwards?)
TS SPEAKER: ja precies (RS)
(yes exactly)

C Top 20 Important Features

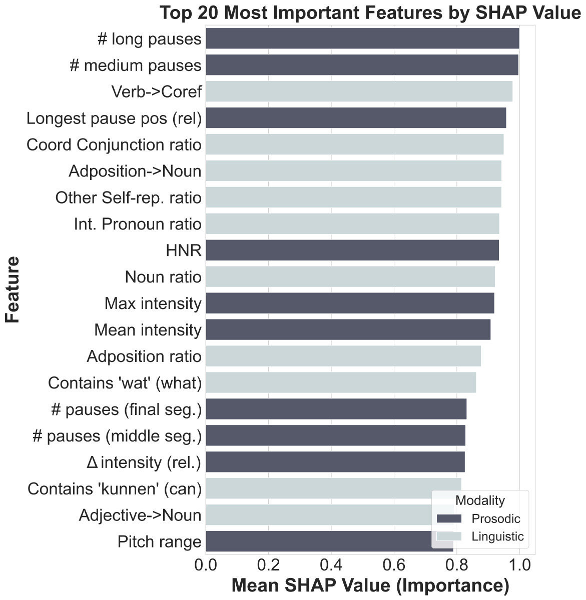
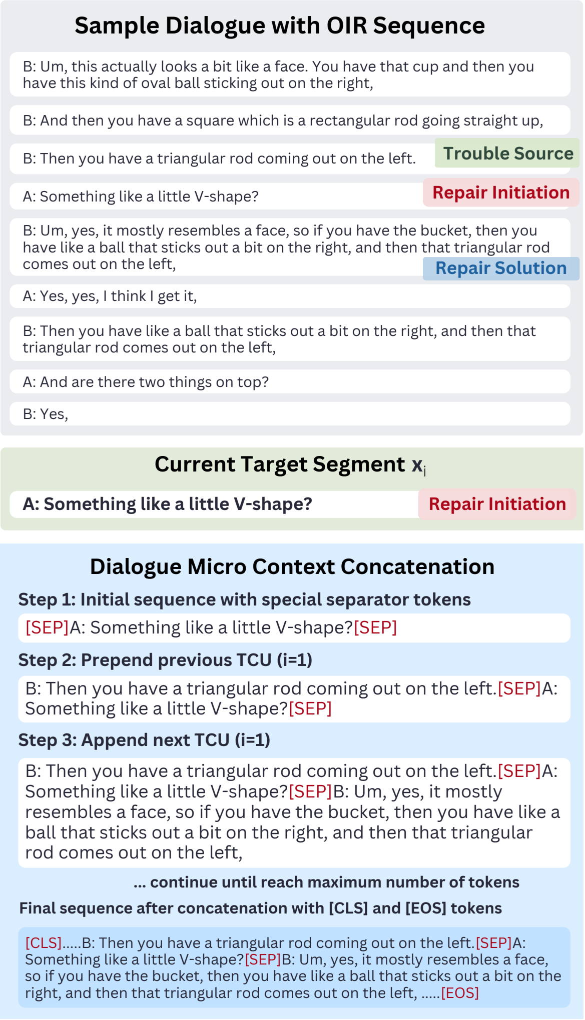


Figure 8: Top 20 most contributed features by SHAP values.

D Dialogue Micro Context



Sample Dialogue with OIR Sequence

B: Um, this actually looks a bit like a face. You have that cup and then you have this kind of oval ball sticking out on the right,

B: And then you have a square which is a rectangular rod going straight up,

B: Then you have a triangular rod coming out on the left. Trouble Source

A: Something like a little V-shape? Repair Initiation

B: Um, yes, it mostly resembles a face, so if you have the bucket, then you have like a ball that sticks out a bit on the right, and then that triangular rod comes out on the left, Repair Solution

A: Yes, yes, I think I get it,

B: Then you have like a ball that sticks out a bit on the right, and then that triangular rod comes out on the left,

A: And are there two things on top?

B: Yes,

Current Target Segment x_i

A: Something like a little V-shape? Repair Initiation

Dialogue Micro Context Concatenation

Step 1: Initial sequence with special separator tokens

[SEP]A: Something like a little V-shape?[SEP]

Step 2: Prepend previous TCU (i=1)

B: Then you have a triangular rod coming out on the left.[SEP]A: Something like a little V-shape?[SEP]

Step 3: Append next TCU (i=1)

B: Then you have a triangular rod coming out on the left.[SEP]A: Something like a little V-shape?[SEP]B: Um, yes, it mostly resembles a face, so if you have the bucket, then you have like a ball that sticks out a bit on the right, and then that triangular rod comes out on the left,

... continue until reach maximum number of tokens

Final sequence after concatenation with [CLS] and [EOS] tokens

[CLS].....B: Then you have a triangular rod coming out on the left.[SEP]A: Something like a little V-shape?[SEP]B: Um, yes, it mostly resembles a face, so if you have the bucket, then you have like a ball that sticks out a bit on the right, and then that triangular rod comes out on the left, ...[EOS]

Figure 9: Dialogue micro context concatenation approach. *Micro context* refers to the immediate conversational environment, including the prior and the subsequent segments of the current target segment in dialogue (Schegloff, 1987).

E Detailed Linguistic Features

Table 4 summarizes the handcrafted feature set that were automatically extracted using the approach proposed in Ngo et al. (2024)'s work.

Level	Feature Group	Feature Type(s)	Description
Segment-level	POS tags sequence	`POS tag bigrams`, `POS tag ratios`	Binary features for frequent POS tag bigrams (e.g., `PRON_Prs`→`VERB`, `VERB`→`COREF`); POS tags frequency ratios computed per segment.
	Lemma	`contains_lemma` (e.g., nog, kunnen)	Binary indicators for presence of high-frequency lemmas relevant to different type of repair initiation.
	Question form	`ends_with_question_mark`	Binary feature indicating whether the segment ends with a question mark.
	Non-verbal action	`contains_laugh`, `contains_sigh`, etc.	Binary features for transcribed non-verbal actions like #laugh#, #sigh#, etc.
Cross-segment level (prior turns related)	Repetition from previous turn	`other_repetition_ratio`	Ratio of tokens in the current segment that are repeated from the other speaker's previous turn relative to total segment length.
	Coreference from previous turn	`coref_used_ratio`	Ratio of coreference phrases (e.g., pronouns or noun phrases referring to previous turn) relative to total segment length.
Cross-segment level (subsequent turns related)	Repair solution TSS self-repetition	`other_speaker_self_rep_ratio`	Ratio of self-repetition in the turn following the repair initiation.
	Repair solution TSS other-repetition	`other_speaker_other_rep_ratio`	Ratio of other-repetition in the turn following the repair initiation

Table 4: Summary of linguistic feature set used for modeling repair initiation. The full POS tag list includes: ADJ (adjectives), ADP (prepositions and postpositions), ADV (adverbs), AUX (auxiliaries, including perfect tense auxiliaries "hebben" (*to have*), "zijn" (*to be*); passive tense auxiliaries "worden" (*to become*), "zijn" (*to be*), "krijgen" (*to get*); and modal verbs "kunnen" (*to be able, can*), "zullen" (*shall*), "moeten" (*must*), "mogen" (*to be allow*)), CCONJ (coordinating conjunctions such as "en" (*and*), "of" (*or*)), DET (determiners), INTJ (interjections), NOUN (nouns), PRON_Dem (demonstrative pronouns), PRON_Int (interrogative pronouns), PRON_Prs (personal pronouns), PUNCT (punctuation), SYM (symbols), and VERB (verbs). The considered common lemma includes: *wat* (what), *kunnen* (can), *zitten* (to sit/set), *zijn* (to be), *nog* (yet/still), *wachten* (to wait), *aan* (on/to/at/in/by/beside/upon). And the transcribed non-verbal actions includes: *laughs*, *sighs*, *breath*, and *mouth noise*.

F Detailed Prosodic Features

Level	Feature Group	Feature Type	Description
Segment-level	Pitch features	`min`, `max`, `mean`, `std`, `range`, `num_peaks`	Extracted from voiced frames; outliers removed; peaks from smoothed contour
	Pitch dynamics	`slope`	Captures pitch variation within segment.
	Intensity features	`min`, `max`, `mean`, `std`, `range`	Computed from nonzero intensity frames; reflects loudness.
	Voice quality	`jitter`, `shimmer`, `hnr`	Reflects vocal fold irregularity and breathiness.
	Pause features	`num`, `durations`, `short/med/long`, `positional counts`, `rel_longest`	Pause detection using adaptive thresholds; categorized by duration and position.
	Speech timing	`rate`, `articulation_rate`, `duration`	Segment length and estimated speech rate (e.g., syllables/sec).
Cross-segment level (both prior and subsequent related)	Transition features	`end_slope`, `start_slope`, `transition`	Pitch slope difference across segment boundaries (prev→cur, cur→next); in semitones/sec.
	Baseline comparison	`z_score`, `rel_change`, `range_pos`	Comparison to speaker's pitch/intensity baseline.
	Latency	`TS`→`RI`, `RI`→`RS`	Silence duration between trouble source and repair initiation, repair initiation and repair solution.

Table 5: Summary of prosodic feature set used for modeling repair initiation.