LREC 2026

Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026

Workshop Proceedings

Editors
Nina Hosseini-Kivanani, Alessio Brutti, Marco Matassoni, Sandipana Dowerah, Davide Liga

11 May 2026
Palma, Mallorca (Spain)

Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026

ISBN 978-2-493814-83-8
EAN 9782493814838

Preface

Welcome to the SPEAKABLE 2026 Workshop on *Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis*.

SPEAKABLE 2026 is a full-day workshop co-located with LREC 2026. It brings together researchers and practitioners working on speech-native language models, with a particular focus on low-resource languages, dialects, and speaker communities. The workshop was created in response to a growing need in the speech and language technology community: while speech foundation models and Speech LLMs have advanced rapidly, their benefits remain unevenly distributed across languages, domains, devices, and user groups.

Low-resource speech settings continue to face persistent constraints related to limited training data, uneven annotation quality, scarce evaluation benchmarks, and restricted computational budgets. These difficulties are often intensified by real-world deployment conditions, including accent and dialectal variation, channel and microphone mismatch, spontaneous speech phenomena, code-switching, noisy environments, and limited availability of lexicons or grapheme-to-phoneme resources. As a result, systems that appear strong under standard benchmark conditions may still fail to provide reliable, fair, and useful performance for many underrepresented communities.

The goal of SPEAKABLE 2026 is to provide a focused forum for addressing these challenges through three closely connected strands. The first strand is **efficient adaptation**, including parameter-efficient fine-tuning, multilingual transfer, knowledge distillation, and edge- or streaming-constrained inference for low-resource speech tasks. The second strand is **meaningful evaluation**, with an emphasis on moving beyond aggregate scores such as WER toward task-appropriate metrics, calibration, robustness analysis, abstention, and slice-aware reporting by accent, dialect, channel, speaker group, and speaking style. The third strand is **responsible practice**, treating bias analysis, data documentation, synthetic-data disclosure, privacy, and safety considerations as routine parts of scientific reporting rather than optional additions.

The call for papers welcomed work on efficient adaptation of Speech LLMs for low-resource languages, evaluation methods for ASR, spoken language understanding, and speech generation, robustness under domain shift, cascaded versus end-to-end error propagation, low-resource corpus creation, lexicon and G2P development, and ethics-by-default reporting. We especially encouraged submissions that combine methodological innovation with strong empirical evidence, transparent documentation, calibrated uncertainty, and, where possible, openly released resources or code.

This first edition of SPEAKABLE reflects the increasing importance of speech technologies for the long tail of languages and communities. The accepted contributions address a broad range of topics, including multilingual and cross-lingual modelling, low-resource adaptation, evaluation design, robustness analysis, data-centric approaches, and practical deployment challenges. Together, they show that progress in speech technology cannot be measured only by performance on high-resource benchmarks. It must also be assessed by how reliably systems work under realistic constraints, how transparently they are evaluated, and how equitably they serve diverse speakers.

The workshop program includes oral and poster presentations, as well as an invited talk by Jordi Luque from Telefónica Research. We are grateful to our Program Committee, consisting of confirmed domain experts from academia and industry, for their careful and constructive reviews. Their work was essential in shaping a balanced and high-quality program. We also thank all authors for submitting their work, revising their papers, and contributing to the scientific scope of this first edition.

We hope that SPEAKABLE 2026 will serve not only as a venue for presenting current research, but also as a catalyst for collaboration, shared evaluation practices, and community-building around inclusive speech technologies. Our broader aim is captured by the workshop's guiding message: build strong models, measure what matters, and make bias analysis routine for speech in the long tail.

Finally, we thank the LREC 2026 workshop chairs and organizers for hosting SPEAKABLE as part of the LREC 2026 workshop program. We also thank our invited speaker, reviewers, authors, participants, and supporting institutions for helping make this first edition possible.

Further information about the workshop, including the program and updates, is available on the SPEAKABLE 2026 website: `https://speakable-2026.github.io/`.

The SPEAKABLE 2026 Organizing Committee

Organizing Committee

- Nina Hosseini-Kivanani, Radio Télévision Luxembourg & University of Luxembourg, Luxembourg
- Alessio Brutti, Fondazione Bruno Kessler, Italy
- Marco Matassoni, Fondazione Bruno Kessler, Italy
- Sandipana Dowerah, Tallinn University of Technology, Estonia
- Davide Liga, University of Luxembourg, Luxembourg
- Christoph Schommer, University of Luxembourg, Luxembourg

Program Committee

- Badr M. Abdullah, Saarland University, Germany
- Şeymanur Aktı, Karlsruhe Institute of Technology (KIT), Germany
- Tanel Alumäe, Tallinn University of Technology, Estonia
- Dimitra Anastasiou, Luxembourg Institute of Science and Technology, Luxembourg
- Leonardo Badino, Almawave, Italy
- Stefano Bannò, University of Cambridge, UK
- Carlos Carvalho, INESC-ID, Portugal
- Shammur Absar Chowdhury, Qatar Computing Research Institute (QCRI), Qatar
- Matt Coler, University of Groningen, Netherlands
- Lorenzo Concina, Fondazione Bruno Kessler (FBK), Italy
- Miguel Couceiro, Universidade de Lisboa, Portugal
- Rohan Kumar Das, Fortemedia, Singapore
- Ioannis Douros, Stavros Niarchos Foundation (SNF), Greece
- Yassine El Kheir, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
- Daniele Falavigna, Fondazione Bruno Kessler (FBK), Italy
- Saeed Farzi, Fondazione Bruno Kessler (FBK), Italy
- Maxime Fily, Inalco (Institut national des langues et civilisations orientales), France
- Peter Gilles, University of Luxembourg, Luxembourg
- Nabarun Goswami, University of Tokyo, Japan
- Felix Herron, Université Paris Dauphine – PSL, France
- Aditya Joshi, UNSW Sydney, Australia
- Joonas Kalda, Pyannote.ai, France
- Ajinkya Kulkarni, Idiap Research Institute, Switzerland
- Spyretta Leivaditi, University of Groningen, Netherlands
- Damien Lolive, IUT of Vannes (University of South Brittany), France
- Keerthana Murugaraj, University of Luxembourg, Luxembourg
- Maria Onoeva, Charles University, Czech Republic
- Ludovica Pannitto, University of Bologna, Italy
- Fernando Perez Tellez, Technological University Dublin, Ireland
- Ben Peters, INESC-ID & Instituto Superior Técnico, Portugal

- Fred Philippy, SnT, University of Luxembourg, Luxembourg
- Bornali Phukon, University of Illinois Urbana-Champaign, USA
- Tina Raissi, RWTH Aachen University, Germany
- Thomas Rolland, Orange, France
- Beatrice Savoldi, Fondazione Bruno Kessler (FBK), Italy
- Imran Sheikh, Vivoka, France
- Ravi Shekhar, University of Essex, UK
- Golshid Shekoufandeh, University of Amsterdam, Netherlands
- Francisco Teixeira, INESC-ID, Portugal
- Hawau Olamide Toyin, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
- Preben Vangberg, Bangor University, UK
- Jelena Vasić, Technological University Dublin, Ireland
- Martijn Wieling, University of Groningen, Netherlands
- Enrico Zovato, Almawave, Italy
- Juan Pablo Zuluaga-Gomez, AGIGO, Switzerland

Invited speaker

- Jordi Luque, Telefónica, Spain

Table of Contents

Workshop Program

May 11, 2026

Room: W11

09:00–09:20 **Introduction and general remarks**

09:20–10:20 **Invited speaker: Jordi Luque**

10:20–10:30 **Remote posters**

Adapting Foundational ASR Models to Efik: An Empirical Study of an Extremely Low-Resource Tonal Language
Offiong Bassey Edet, Stephen Orok Duke, Enoima Essien Umoh, Benjamin Okon Nyong and Andrew Asuquo Nkpanam

PAREDA: A Multi-Accent Speech Dataset of Natural Language Processing Research Discussions
Sicheng Jin, Dipankar Srirag and Aditya Joshi

10:30–12:00 **Coffee Break and Poster session**

Say Again? The Limits of Whisper with Conversation. A Case Study on the KIParla Corpus.
Martina Simonotti, Ludovica Pannitto, Caterina Mauri, Adriano Ferraresi and Gabriele Carioli

Not All Polar Questions Are the Same: ASR, Humans, and Russian
Maria Onoeva

Quantizing Whisper: How Design Choices Affect ASR Performance
Arthur Söhler, Julian Irigoyen and Andreas Søeborg Kirkedal

"OK Aura, Be Fair with Me": Demographics-Agnostic Training for Bias Mitigation in Wake-up Word Detection
Fernando López, Paula Delgado-Santos, Pablo Gómez, David Solans and Jordi Luque

Scalable Expansion of Multilingual Speech LLMs for ASR: A Continual Learning Approach
Lorenzo Concina, Marco Matassoni and Alessio Brutti

May 11, 2026 (continued)

Responsible Benchmarking of Fairness for Automatic Speech Recognition
Felix E. Herron, Ange Richard, François Portet, Alexandre Allauzen and Solange Rossato

Addressing Accent Disparities in Automatic Speech Recognition: A Comparative Study of Single and Two-Step Adaptation
Mykhailo Danilevskyi, Fernando Perez-Tellez and Jelena Vasic

Investigating Speaker Pronunciation Variability in Speech Embeddings: Speaker and L1 Effects on French as a Second Language
Maxime Fily, Martine Adda-Decker and Guillaume Wisniewski

What LID Systems Say About Dialectal Variation. The Case of Yiddish, Quechua and Mande
Johanna Cordova, Eric Jordan and Valentina Fedchenko

HARNESS: Lightweight Distilled Arabic Speech Foundation Models
Vrunda Nileshkumar Sukhadia and Shammur Absar Chowdhury

When Does OmniASR Fail? A Fine-Grained Human Evaluation on Saudi Arabic Dialects
Hend Al-Khalifa

12:00–13:00 **Spotlight papers**

SpeechLM for Automatic Speech Recognition in Low-resource Languages
Md Abdur Razzaq Riyadh, Eneko Agirre, Eva Navas and Claudia Borg

Improving Low-resource ASR Using Bilingual Fine-tuning with Language Identification: A Cross-linguistic Evaluation
Reihaneh Amooie, Yun Hao, Wietse de Vries, Jelske Dijkstra, Matt Coler and Martijn Wieling

May 11, 2026 (continued)

13:00–14:00 **Lunch break**

14:00–16:00 **Architectures and learning methods**

Leveraging Speech Models for Audio-based Lexical Retrieval in Dictionaries: The Case of the Teochew Language
Siman Chen, Ilaine Wang, Maxime Fily and Pierre Magistry

Stage-Aware Cross-Lingual Transfer for Faroese ASR: When and Which Languages Matter
Dávid í Lág, Barbara Scalvini, Carlos Daniel Mena and Jón Guðnason

Doing More with Less: Determining Optimal Pre-training Model for Irish Automatic Speech Recognition through Multi-step Fine-tuning
Caoilfhionn Ní Dheoráin, Ruth Holmes, Nicholas Evans, Thomas Laurent, Anthony Ventresque and Ellen Rushe

Blank-Aware Decoding for Transcript-Free Phoneme Alignment in Low-Resource Languages and Dialects
Domenico De Cristofaro, Barbara Plank and Alessandro Vietti

On the Role of Encoder Depth: Pruning Whisper and LoRA Fine-Tuning in SLAM-ASR
Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis and Ravi Shekhar

TaLK-Corpus: A Regionally Diverse Evaluation Set for Sri Lankan Tamil Speech
Adsajan Thillainathan, Nishanthini Kanthakumar, Nivethiga Rasan and Kengatharaiyer Sarveswaran

May 11, 2026 (continued)

16:00–16:30 **Coffee break**

16:30–17:00 **Best paper and closing remarks**

Say Again? The Limits of Whisper with Conversation. A Case Study on the KIParla Corpus.

Martina Simonotti, Ludovica Pannitto, Caterina Mauri
Adriano Ferraresi, Gabriele Carioli
University of Bologna, Bologna - Italy
martina.simonotti@studio.unibo.it,
{ludovica.pannitto, caterina.mauri, adriano.ferraresi, gabriele.carioli}@unibo.it

Abstract

This study investigates how Whisper handles interactional phenomena in spontaneous Italian conversation, focusing on backchannels, repairs, and filled pauses. We compare standard Word Error Rate (WER) optimization with a decoding strategy that explicitly rewards the preservation of interactional events. Results show that decoding choices have limited impact on overall accuracy, while recognition remains strongly phenomenon-dependent, suggesting structural limitations in the handling of interactional phenomena, with systematic linearization of repairs and frequent suppression of short conversational items.

Keywords: ASR, Interaction, Italian, Backchannels, Repairs, Filled Pauses

1. Introduction

End-to-end Automatic Speech Recognition (ASR) systems based on large neural models have reached high levels of performance. Models such as OpenAI's Whisper (Radford et al., 2023) are widely used thanks to their robustness across languages, recording conditions and speaker variability. However, most ASR systems are trained primarily on monologic, non-spontaneous speech and tend to normalize conversational input, treating interactional phenomena as noise (Lopez et al., 2022; Yamasaki et al., 2023). Previous studies report systematic limitations in the production of short conversational items, interjections, disfluencies, overlap and paralinguistic cues (Liesenfeld et al., 2023; Lopez et al., 2022; Umair et al., 2022; Zayats et al., 2019). All of these elements play a crucial role in the organization of spoken discourse and, from a linguistic perspective, constitute structurally relevant resources. This paper focuses on three specific phenomena: backchannels, conversational repairs and filled pauses. Backchannels signal attention or speaker alignment (Dideriksen et al., 2019; Mereu et al., 2024; Blomsma et al., 2024); repairs manage problems of speaking and understanding (Drew, 1997; Schegloff et al., 1977; Dingemanse and Enfield, 2015, 2024); and filled pauses reflect speech planning and turn management (Christenfeld et al., 1991; Spreafico, 2012; Cossavella and Cevasco, 2021). Because they are brief, prosodically subtle and often produced in overlap, they are especially vulnerable to omission or normalization in ASR output (Lopez et al., 2022). Studies show that limited representation in output transcripts does not necessarily imply their absence from the acoustic signal or from the model's internal representations. Interactional elements may be suppressed during decoding, when the system selects which hypotheses to render as text, implying that ASR output should therefore be treated as configuration-dependent (Vitale et al., 2024; Dinkar et al., 2023). This study adopts a linguistically oriented perspective to examine how different Whisper decoding configurations affect the representation of interactional phenomena in selected conversations from the KIParla corpus (Mauri et al., 2019)[1].

2. Backchannels, Repairs and Filled Pauses

Backchannels (Example (1)) are listener responses that display attention and support the speaker's ongoing turn (Blomsma et al., 2024; Dideriksen et al., 2019; Mereu et al., 2024). They may consist of short tokens or multi-unit sequences and contribute to maintaining interactional organization (Pernas and Borreguero Zuloaga, 2010). Following the definition proposed by Ward and Tsukahara (2000), a backchannel (i) responds directly to the other's utterance, (ii) is optional, and (iii) does not require acknowledgement.

(1) BOI115: ci sono esami sempre e
there are always exams and
si studia anche durante il
you have to study also during your
tirocinio anzi per forza
internship well actually you must
BOR031: madonna ok wow
oh my god ok wow

[1]Data and code are provided at https://github.com/KIParla/say-again.

Proceedings of SPEAKABLE @ LREC 2026, pages 16–30
11 May 2026.

BOI115: quindi non c'è un periodo di stop
so there isn't a break
che dici studio e basta
where you can just focus on studying
BOR031: **mhmhmh**
uh-huh, uh-huh
PBA030, ParlaBO (Mauri et al., 2024b).

Conversational Repair refers to practices through which speakers address problems of speaking, hearing or understanding (Schegloff et al., 1977; Fele, 2007; Dingemanse and Enfield, 2015; Clark, 2020). Repair preserves mutual understanding and conversational progressivity. It may be *self-initiated* (Example (3)), when speakers correct their own talk, or *other-initiated* (Example (2)), when recipients signal trouble and prompt clarification from the main speaker (Schegloff et al., 1977).

(2) PKP040: quando c'ha l'esame marco?
when does marco have his exam?
PKP041: il dieci
on the tenth
PKP126: quindi mo sta chiuso a studiare
so now he's locked up at home studying
[...]
PKP041: ma' tu te lo sei visto sherlock?
mum, did you watch sherlock?
PKP040: **prego?**
pardon?
PKP041: te lo sei visto sherlock? della bbc?
did you watch sherlock?
the bbc one?
PKP040: no cioè se è quello li che facevano
no I mean if that's
la serie sì
the series then yes
KPS008, KIPasti (Mauri et al., 2024a).

(3) PSB050: ti trovi meglio a bologna o pavia?
where do you feel most comfortable in Bologna or in Pavia?
PSB049: pavia
Pavia
eh no no scusa bologna scusa
no no sorry I mean Bologna
PSB050: ah era la risposta sbagliata
ah that was the wrong answer
PSB049: eh sì sì bologna
yeah yeah Bologna
SBIB006, StraParlaBO (Zucchini et al., 2026).

Filled Pauses (Example (4)) are hesitation markers that interrupt fluency without contributing to propositional content. They reflect speech planning and information retrieval processes, while often functioning as floor-holding devices (Christenfeld et al., 1991; Spreafico, 2012; Schettino and Cataldo, 2019; Cossavella and Cevasco, 2021).

(4) PST211: **ehm** ti vengono in mente dei casi
um *can you think of any situations*
in cui si mischiano le due
where the two
lingue in casa
languages get mixed at home
cioè proprio stereotipico
like something typical
che succede?
that happens?
PST036: **ehm** quando mia mamma è
um *when my mom gets*
arrabbiata
angry
STIR012, StraParlaTO (Bernasconi and Goria, 2026).

3. Whisper in Interactional Contexts

Whisper (Radford et al., 2023) represents a shift in ASR research. Trained on $680,000$ hours of audio, it shows strong zero-shot generalization across domains, speakers, and languages. Architecturally, it relies on a standard encoder–decoder Transformer architecture that formulates speech processing tasks as token prediction within a unified sequence-to-sequence framework. However, despite these recent advances, conversational speech remains challenging for ASR systems, including Whisper (Yamasaki et al., 2023). Elements central to interactional organization are underrepresented in ASR output, being disproportionately prone to be omitted or misrecognized, especially when utterances are very brief (Cumbal et al., 2021; Lopez et al., 2022). This issue is further compounded by the way transcription accuracy is typically evaluated, that is, through the Word Error Rate (WER, Klakow and Peters 2002). This metric has been criticized for conversational data because it oversimplifies performance and gives equal weight to all word-level errors, regardless of their interactional relevance (Liesenfeld et al., 2023; Gorisch and Schmidt, 2024). As a result, ASR systems may achieve acceptable global WER scores while still failing to capture interactional features that are crucial for Conversation Analysis. Therefore, in this work, we explicitly target two research questions, namely:

RQ1: How does decoding optimization affect the transcription accuracy of spontaneous conver-

sational speech, as measured by global WER and event-specific metrics?

RQ2: Are certain interactional phenomena structurally more vulnerable to omission or normalization, regardless of optimization strategy?

4. Methodology

The experimental workflow can be summarized as follows: (i) the three interactional phenomena described in Section 2 were manually annotated on selected conversations from the KIParla corpus (see Section 4.1); (ii) audio data were processed through a speaker diarization and segmentation pipeline to obtain speaker-attributed segments (see Section 4.2); (iii) Whisper decoding was performed by optimizing inference-time parameters while keeping the ASR model fixed, using standard WER as a baseline objective function and an Interaction-aware objective function designed to promote the retention of interactional phenomena (see Section 4.3); (iv) the resulting transcriptions were normalized and evaluated both quantitatively, through (a) global WER and (b) Mean Match Ratio for annotated phenomena (see Section 4.4), and qualitatively, through detailed inspection of match and mismatch patterns and representative examples. Results were compared across configurations to assess the impact of the two optimization pipelines on the representation of interactional phenomena.

4.1. Dataset

We selected 26 2-speaker conversations drawn from the KIParla Corpus of Spoken Italian, covering three macro types of interaction: semi-structured interviews, student-professor meetings (i.e., oral exams and office hours) and free conversations. The sample is heterogeneous in terms of speakers' metadata and is in line with the general composition of the KIParla corpus. Linguistic variation is also present, including one L2 speaker of Italian and several conversations featuring dialectal traits from different regions. The dataset in its entirety amounts to 15h48m of audio.

The KIParla Corpus is a resource of spoken Italian and is entirely transcribed following a manual pipeline, in both ortographical and Conversation Analysis format, following Jefferson notation (Jefferson, 2004). The resource is available in vertical, pseudo-tokenized format where each token bears information about its type (e.g., linguistic, metalinguistic), ID, corresponding speaker code, Jefferson notation, etc.[2]

The first 20 minutes of each recording were manually annotated in INCEpTION (Klie et al., 2018) using a simple multi-layer scheme, aimed at indentifying backchannels (BC), self- and other-initated repairs (SR and OR) and filled pauses (FP), while also maintaining Conversation Analysis information such as Jefferson notation, overlaps and intonation patterns. A screenshot of the INCEpTION annotation interface is provided in Appendix B, Figure 7.

The annotated portion of the dataset resulted in 8h40m of conversation. Annotation was carried out by an expert linguist while listening to the corresponding audio in ELAN (Max Planck Institute for Psycholinguistics, 2025) to accurately capture interactional dynamics. The remaining 6h48m were split into two groups, and employed for Whisper's optimization (Subset A, 3h25m) and subsequent control analysis (Subset B, 3h22). Annotation criteria is described in Appendix A.

4.2. Diarization

Segmentation, diarization, transcription and optimization were performed using DIT.DaT, a modular pipeline developed for the Department of Translation and Interpreting of the University of Bologna[3], combining PyAnnote (Bredin et al., 2019) for speaker diarization and Whisper for transcription. The pipeline produces speaker-aligned outputs that can be manually reviewed step-by-step. Compared to alternative solutions such as WhisperX (Bain et al., 2023), this approach offers greater flexibility, a more controlled workflow and more reliable speaker segmentation for the interactional phenomena under investigation.

Four types of segmentation were explored: while sharing the same processing pipeline, they only differ with respect to a limited set of parameters that manage diarization and segmentation sensitivity. Table 1 shows the four segmentation configurations, denominated A, B, C and D. Configuration A is the most restrictive: it enables exclusive mode (which assigns overlapping speech to a single, dominant speaker) and uses conservative segmentation thresholds (2s minimum pause, 0.25s minimum duration), favoring cleaner but less overlap-sensitive turns. Configuration B is identical to A but disables exclusive mode, allowing overlapping speech to be attributed to multiple speakers. Configuration C further increases segmentation sensitivity by reducing the minimum pause threshold to 1 second, enabling finer-grained turn segmentation. Configuration D is the most permissive setup: it maintains the 1-second pause threshold and further lowers the minimum segment duration to 0.20 seconds, maximizing retention of very short turns

[2]These files are available on Github for each module. For more information: https://github.com/KIParla/KIP?tab=readme-ov-file#verticalized-content.

[3]https://github.com/bilo1967/DIT.DaT

(e.g., minimal responses and hesitation markers) at the cost of greater fragmentation. All conversations were finally segmented according to each configuration.

ID	exclusive mode	min. pause (s)	min. duration (s)	sensitivity
A	yes	2.0	0.25	low
B	no	2.0	0.25	medium
C	no	1.0	0.25	high
D	no	1.0	0.20	very high

Table 1: Overview of the four processing configurations used in the optimization process.

4.3. Optimization

Before optimization, a pre-processing pipeline was implemented to ensure comparability between the output of each Optuna trial with the gold standard transcription used for reference (i.e., the manually transcribed version available for consultation). As far as the optimization is concerned, Whisper decoding parameters were automatically optimized on Subset A with Optuna (Akiba et al., 2019). Two objective functions were explored. The first relied on standard WER, computed through word-level alignment between ASR output and the gold standard. WER is defined as:

$$\mathrm{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions, C the number of correct tokens and N the total number tokens in the reference transcript. This standard WER-based optimization is used as a baseline condition against which the Interaction-aware objective described below is evaluated. Despite being widely criticized in the literature for disproportionately weighting the representation of interactional phenomena, WER was adopted as the primary optimization metric: at present, it remains the standard benchmark for evaluating ASR transcription accuracy and guiding parameter tuning. Also, the lack of widely established alternative metrics makes WER a suitable baseline objective function for this study.

A second, Interaction-aware objective function was introduced to better account for interactional items that are often suppressed or normalized in ASR output. Specifically, we introduce a new Loss function that weighs the WER according to the phenomena under investigation. Given the number of suppressed target items S_t and the number of correctly produced target items P_t, the objective function was defined as:

$$\mathcal{L} = \mathrm{WER} + \lambda \cdot S_t - \mu \cdot P_t$$

Suppressed target tokens are calculated as the proportion of backchannel, repair and filled pause tokens that are either deleted or substituted, and produced target tokens as the proportion of correctly transcribed tokens belonging to the same group. Target tokens were identified using a predefined form-based list of short conversational tokens provided in Appendix C. The coefficients were manually set to $\lambda = 0.3$ and $\mu = 0.1$: both weights were intentionally kept low to avoid distorting overall transcription quality or encouraging hallucinated output. This Interaction-aware optimization should be interpreted as an exploratory extension of the WER-based approach. It is important to underline that its aim is not maximizing the production of interactional events. Instead, it seeks to reduce their systematic suppression when they are present in the reference transcription, while preserving overall transcription accuracy. This choice reflects a conservative design aimed at testing whether small adjustments at decoding time are sufficient to influence the representation of interactional phenomena. For each segmentation configuration and for each objective function, the optimization process yielded the best trial, which represents the best-performing parameters. They are listed in Appendix D, Table 6.

4.4. Normalization and Error Analysis

Before evaluating the optimized outputs, a preliminary inspection of word-level alignments was conducted to identify systematic mismatches that were not related to genuine recognition errors. Raw substitution patterns and manual alignment checks revealed that several high-frequency errors were due to orthographic variation or tokenization inconsistencies. Normalization therefore addressed three main issues: (i) differences in token boundaries (e.g., apostrophe splitting: gold *all'* + *interno* vs. Whisper *all'interno*), (ii) truncated forms in repair sequences (e.g., *vennero anda-*), and (iii) deleted pause markers ([PAUSE]), which are not modeled by the ASR system. These adjustments aimed to prevent structural penalization of phenomena that Whisper cannot explicitly encode. Additionally, frequent orthographic variation in non-lexical vocalizations (e.g., nasal backchannels and filled pauses such as *m*, *mh*, *eh*) was observed. Tokens composed exclusively of *m* and *h* were treated as equivalent across gold and Whisper outputs. Similarly, *eh* $\rightarrow$ *e* ('and') substitutions were normalized only when functioning as filled pauses, based on their phonetic proximity and high occurrence (50 instances in the WER-based and 49 in the Interaction-aware output). In contrast, since no consistent pattern was found, substitutions such as *eh* $\rightarrow$ *è* ('is') were retained: in this case, normalizing them would have introduced interpretative assumptions rather than simply correcting orthographic variation. The

top ten raw substitutions are listed in Appendix F. All these normalization choices reflect a methodological trade-off between evaluation fairness and linguistic fidelity. On the one hand, normalization reduces structural mismatches that would otherwise inflate error rates due to tokenization differences, Whisper encoding limitations, or orthographic variation, thereby improving comparability between ASR output and the reference. On the other hand, it may partially obscure the fine-grained form and variability of interactional phenomena, especially in cases where orthographic variation carries interactional or phonetic significance.

A quantitative error analysis was subsequently conducted on all configurations. Word-level alignments were examined to compute the distribution of insertions (INS), deletions (DEL), substitutions (SUB), and correct matches (OK) for interactional tokens. This analysis allowed error patterns to be evaluated independently from the overall WER, providing an empirical basis for comparing the two optimization strategies. To verify that the observed behavior was not specific to Subset A, a complementary analysis was conducted on an additional held-out subset of conversational data not used during parameter tuning, Subset B. This control analysis was performed only for best-performing configuration for each optimization strategy, therefore Configuration A.

Performance was evaluated at two complementary levels: global transcription accuracy and interactional event preservation. First, overall accuracy was assessed through the WER, computed for each conversation and each decoding strategy. Conversation-level WER values were compared in a paired design, and descriptive statistics (mean, variance, standard deviation) were used to summarize central tendency and dispersion. Visualizations at both conversation and aggregate level facilitated comparison between optimization strategies. Second, to specifically evaluate the interactional phenomena under investigation, a personalized metric denominated Mean Match Ratio was introduced. Each annotated sequence was treated as a single event, including multi-token units. For each event, a match ratio was computed as the proportion of correctly transcribed tokens over evaluable tokens. For each phenomenon type, the Mean Match Ratio was then calculated as the simple average across events, ensuring equal weight regardless of event length. Analyses were conducted at two levels. At the conversation level, paired differences between optimization strategies were computed for each phenomenon. Since Shapiro–Wilk tests indicated non-normal distributions of paired differences, Wilcoxon signed-rank tests were employed. At the interaction level, events were aggregated to provide a comparison across conversational settings. However, due to the limited number of conversations per interaction type, normality testing within each interactional setting was deemed unreliable. Consequently, only inferential and descriptive analyses were conducted at the overall type of interaction level.

5. Results

5.1. Analysis on Optimal Parameters

All best-performing trials converged on the `large-v3` model, therefore the findings should be interpreted as model-specific. Decoding parameters varied considerably across configurations (see Table 6 in Appendix D), with no consistent optimal strategy emerging.

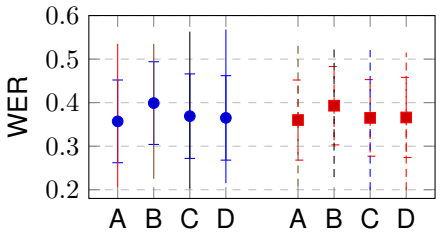

Figure 1: Mean WER with standard deviation (error bars) and min–max whiskers across configurations. Circles: WER-based optimization. Squares: Interaction-aware optimization.

Mean WER values remained comparable across objectives (Figure 1). In the WER-based optimization, mean values ranged from 0.357 (Configuration A) to 0.399 (Configuration B), with relatively similar variability across configurations. In the Interaction-aware optimization, Configurations A (0.360), C (0.365) and D (0.366) performed similarly, while B showed the highest WER mean (0.393). Variability was slightly reduced overall. Configuration A achieved both the lowest composite loss (0.563) and the lowest mean WER, and was therefore selected as the best Interaction-aware setup. Importantly, incorporating event-sensitive components did not degrade global WER.

Event-level error distributions (Table 2) show that deletions remain the dominant error under both objective functions (ranging from 50 to almost 55%), confirming the tendency of a structural suppression of interactional tokens. Introducing the Interaction-aware optimization does not substantially reduce deletion rates. However, minor shifts in distribution are observable: substitutions decrease slightly in some configurations (notably B), correct matches increase marginally, and insertions remain low. This means that the Interaction-aware objective function slightly redistributes errors without altering the overall omission tendency.

ID	type	DEL (%)	SUB (%)	OK (%)	INS (%)
A	WER	53.39	19.78	21.11	5.73
A	I-WER	54.67	20.44	21.39	3.50
B	WER	50.92	23.12	20.57	5.39
B	I-WER	53.04	20.45	20.91	5.61
C	WER	51.72	25.86	18.06	4.36
C	I-WER	50.89	25.13	19.67	4.32
D	WER	51.17	24.91	19.06	4.86
D	I-WER	51.18	24.83	20.12	3.87

Table 2: Distribution of alignment operations for event tokens under WER-based and Interaction-aware (I-WER) optimization.

The control analysis on Subset B (see Appendix E, Table 7) replicates this pattern. Interaction-aware optimization yields a slightly lower WER mean (33.24%) than the WER-based setup (34.23%), confirming that event-sensitive weighting does not harm global accuracy. Deletions, again, remain dominant (58%, approximately), with only minimal differences between strategies. Overall, Interaction-aware optimization introduces small, controlled shifts in error distribution while preserving transcription quality.

The contrast between the highest-WER configuration (Trial 0, value = 0.90) and the lowest-WER configuration (Trial 21, value = 0.35), which can be examined in Table 3, qualitatively illustrates the structural impact of decoding parameters. In the highest-WER output, lexical instability and hallucinations emerge: proper nouns are distorted (*San Luca* → *saluca/saluka*), non-existent forms appear (*l'equiline*, *videa*), and morphologically implausible variants are produced (*non cambiente*, *villana*). These errors may reflect decoding instability rather than simple substitution patterns. By contrast, the optimized configuration preserves referential consistency and syntactic coherence. *San Luca* remains stable, morphologically correct forms are produced (*non cambia niente*), and substitutions are semantically plausible (e.g., *nelle colline* instead of *l'equiline*). Minor infelicities persist (e.g., *sono tutte verde* instead of *sono tutte verdi*), but the transcript remains globally readable and interpretable. The difference is therefore not merely quantitative: suboptimal parameters produce cumulative lexical degradation and phonological drift, whereas optimized decoding preserves discourse continuity and referential stability.

5.2. Analysis on Annotated Data

Conversations from the annotated dataset were transcribed using the best parameters for each optimization strategy, that is, Configuration A for both. Across interaction types (Figure 2), relative rankings between the two types of optimizations remain stable, suggesting that conversation-specific factors (e.g., acoustic conditions, overlap density) may drive most variability rather than optimization strategy. Aggregated statistics confirm this pattern: the Interaction-aware configuration yields a slightly lower mean WER (0.361) compared to the WER-based setup (0.367), with marginally reduced variance (WER 0.0096 vs. IA 0.0071) and standard deviation (WER 0.0979 vs. IA 0.0846). Differences are small and distributional structures remain comparable. Overall, incorporating interaction-sensitive components does not substantially alter global transcription accuracy, but slightly stabilizes performance.

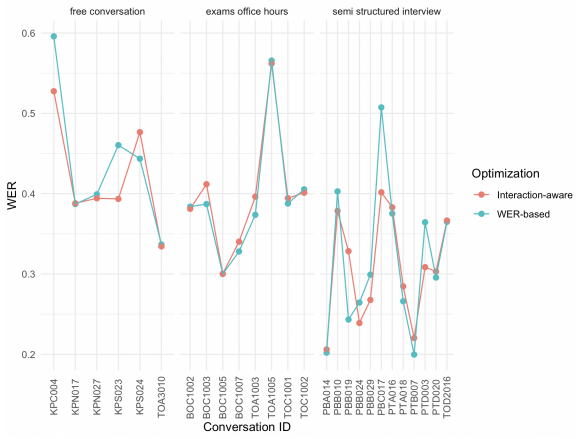

Figure 2: Comparison of conversation-level Word Error Rate (WER) between WER-based and Interaction-aware optimization strategies, grouped by interaction type. Each pair of points corresponds to the same conversation under the two decoding conditions.

Regarding the Mean Match Ratio, Wilcoxon signed-rank tests at the conversation-level revealed no statistically significant differences between the two configurations for any of the investigated phenomena, despite the variability observed in Figure 3. Indeed, all phenomena $p > 0.05$ (BC $= 0.207$, FP $= 0.780$, SR $= 0.197$, OR $= 0.371$). This indicates that the median difference in Mean Match Ratios between the two optimization strategies does not significantly deviate from zero: despite small conversation-level fluctuations, neither configuration demonstrates a systematic advantage over the other in the recognition of annotated interactional phenomena. This result suggests that decoding-time optimization alone may not be sufficient to substantially affect the recognition of interactional phenomena. Given the relatively low number of OR events per conversation, the Wilcoxon test for this phenomenon should be interpreted with caution, as limited sample size may reduce statistical power.

Clear differences emerge across phenomena when analyzed at the interaction-level (Table 4).

Speaker	High-WER output (Trial 0 – value 0.90)	Low-WER output (Trial 21 – value 0.35)
SPEAKER_B	San Luca, quando con una persona che mi ganna, magari sei innamorata, poi la scena quando vai lì, l'equiline, il San Luca è San Luca.	San Luca quando una persona che magari sei innamorata, poi la sera quando vai lì nelle colline, San Luca è San Luca.
SPEAKER_A	Il saluca è saluca. Cioè... Però avevi detto che hai fatto la camminata verso saluca che hai fatto. Sì, se mai non avevi avuto...	San Luca è San Luca. Però mi hai detto che hai fatto la camminata verso San Luca. Sì, sembra una mia vita.
SPEAKER_B	Saluka è sempre là, non cambiente perché tutti vanno alla fine, è sempre desiderata magari da tutti, vanno tutti là e quindi ci sono tutti tutti verdi, bellissimo, ci sono delle videa che rimani senza parole.	San Luca è sempre là, non cambia niente, perché tutti vanno alla fine, è sempre desiderata magari da tutti, vanno tutti là. E quindi sono tutte verde, bellissimo. Ci sono delle vite là che rimani senza parole.
SPEAKER_A	Chissà, se un giorno ormai...	Chissà se un giorno...
SPEAKER_B	per grattavinci compro una villana	Se non ci sono grattaventi comprerò una villa là.

Table 3: Comparison between the highest-WER configuration (Trial 0) and the lowest-WER configuration (Trial 21) of the WER-based optimization. Extracted from PBB010, ParlaBO (Mauri et al., 2024b).

Self-repairs show the highest Mean Match Ratios (≈ 45–56%), suggesting relative robustness. Backchannels display intermediate preservation (≈ 17–23%) and filled pauses remain extremely low across all contexts (≈ 1–4%), confirming their high susceptibility to omission. Other-initiated repairs show relatively high values, however they are based on small samples. That said, phenomenon type appears to be the primary determinant of recognition performance.

interaction type	phen.	n	WER (%)	Event (%)
free conversation	BC	524	16.92	16.89
	FP	174	3.45	4.02
	SR	48	51.80	52.49
	OR	13	46.15	53.85
exams/office hours	BC	636	22.10	21.53
	FP	761	1.38	1.45
	SR	133	46.64	45.12
	OR	18	55.56	58.15
semi-structured interview	BC	1354	21.71	23.01
	FP	776	4.23	4.17
	SR	62	56.11	53.33
	OR	0	–	–

Table 4: Mean match ratio (in percentage) by interaction type and optimization strategy.

5.3. Qualitative Analysis

Filled Pauses Across the few successful matches, filled pauses tend to be prosodically salient, segmentally isolated, and positioned at the onset of an intonational unit, preceding propositional content (e.g., PTD020, Figure 4). Nasal hesitations (e.g., *ehm*, *mh*) are almost systematically omitted, however *ehm* appears to be recognized more frequently than *mh*. The few matched instances of nasal forms, including the single normalization case (*mh* → *mmm* in KPS024),

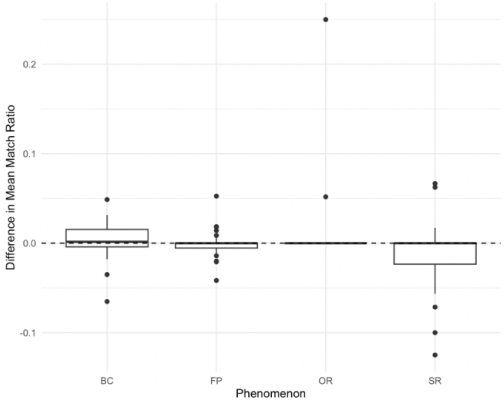

Figure 3: Distribution of differences in Mean Match Ratio, for each phenomenon.

share similar properties: they are acoustically isolated, not produced in overlap, and occupy structurally well-defined turn positions. The near absence of preserved nasal forms overall suggests that short, low-intensity hesitation markers are particularly vulnerable to deletion, especially when produced in turn-final or non-prominent positions.

```
TOR001: | [cosa fai durante il tuo tempo libero?]
TOI008: |

TOR001: |
TOI008: |          [e:::h è una bella domanda]
event:  |          [FP      ]

TOR001: |
TOI008: | [diciamo che::: mh ]
event:  |                [FP ]
```

Figure 4: A recognized FP from PTD020, ParlaTO (Cerruti and Ballarè, 2020).

Backchannels Qualitative inspection reveals a clear positional pattern in the preservation of backchannels. Successfully recognized cases tend

to occur outside overlap, either in the transition space between turns or at the onset of a new turn. In these contexts, backchannels such as *eh no ti capisco*, turn-initial *eh*, or nasal *mh* are sequentially and prosodically salient. Their placement at a recognizable boundary (i.e., immediately after a prior turn completion or as incipient speakership), makes them acoustically less masked and structurally easier to segment. Whisper therefore appears more likely to preserve backchannels when they are integrated into a clearly delimited turn rather than embedded within ongoing speech. Only a handful of preserved cases occur in overlap: however, these instances are relatively limited and typically involve multi-word backchannels (e.g., *eh immagino*, *eh vabbè*) that are longer and more acoustically robust than, for instance, an isolated interjection. The contrast becomes particularly evident in alignment sequences. In BOC1002 (Figure 5), the backchannel *bibliografia sì*, produced in overlap as an aligning response, is not preserved by Whisper. By contrast, the subsequent token *magistrale*, produced at the onset of a new turn and outside direct overlap, is correctly transcribed. This asymmetry suggests that overlap, rather than interactional function per se, constitutes the primary source of suppression.

```
BO087: |[[<bibliografia>],]     [magist..→
BO089: |[[bibliografia (.) sì].]
event: |[BC                ]

BO087: |rale.]
BO089: |      [magistrale (.) ho portato an..→
event: |      [BC         ]

BO087: |
BO089: |anche::: [il reperto..→
```

Figure 5: Example of both an unrecognized and a recognized BC alignment from BOC1002, KIP (Mauri et al., 2019).

Self-repairs Across both decoding strategies, self-repairs are predominantly partially preserved. Fully preserved cases are rare, and complete omissions remain limited but non-negligible. The dominant pattern is linearization, meaning that Whisper tends to retain the repair solution while deleting the repairable, truncations, and hesitation markers. In BOC1002, the sequence *delle opere pubbli- [PAUSE] delle traduzioni pubblicate* is reduced to *delle traduzioni pubblicate*: the initial formulation is removed, and only the corrected portion of the repair survives. Similarly, in TOC1002, *possono eh am- [PAUSE] agire* is simplified to *possono agire*, with the hesitation marker and truncated segment deleted. In TOC1001, in the sequence *che si è contratti [PAUSE] che si è contratto*, the incorrect plural agreement (*che si è contratti*) disappears, while the grammatically correct form in singular (*che si è contratto*) is directly preserved. In these cases, Whisper privileges grammatical coherence over the faithful reproduction of incremental speech production. Truncated forms are not preserved as such: in PTB007, the speaker first produces the truncated *dal millenovec-*, followed by the complete form *nel millenovecento*. In the ASR output, however, the truncated repairable is omitted, while the reformulated and phonologically complete segment (*nel millenovecento*) is retained. The repair trajectory therefore disappears, and only the resolved, well-formed formulation survives. More complex repairs are likewise compressed. In PBB029, the speaker first produces the incorrect clause *non ho mai pesato* and then reformulates it as *non mi è mai pesato*. In the ASR output, the entire first formulation is deleted, and only the second clause is preserved. Fully preserved cases are exceptional. In PBA024, both the initial formulation (*abito praticamente vicino al centro*) and the reformulation (*lavoro vicino al centro*), together with repair markers occurring between them (*cioè, scusa*), are retained. Unlike the previous examples, this sequence (i) contains no abrupt truncations (ii) does not contain a silenced pause and, finally, (iii) contains an explicit repair marker, which is the element that appears to have made the difference for the sequence to be fully preserved.

Other-initiated Repairs The qualitative analysis of other-initiated repairs (OR) reveals a recurrent pattern of structural simplification. Repair initiations are frequently partially preserved or entirely suppressed, especially when short, repeated, or produced in overlap. In BOC1005, the clarification request *che cosa vuol dire?* is produced twice in close succession, reinforcing the repair trajectory. However, Whisper only retains the first occurrence, while the second is omitted. From an interactional perspective, the repetition intensifies the request for clarification and signals persistent trouble. Its deletion attenuates this persistence, reducing the sequence to a single, non-reiterated question. This suggests that Whisper may suppress closely repeated repair initiators, particularly when they occur rapidly and in overlap. A more radical case of suppression is observed in TOA3010, shown in Figure 6. The repair initiation *ma chi?*, targeting referential ambiguity, and the subsequent specification (*Luca*) are both omitted in the ASR output. The entire repair trajectory disappears, leaving the referential problem unresolved. Here, both the initiator and the repair solution are removed, effectively erasing the interactional work performed to restore clarity. This pattern indicates that short, overlapping, and low-intensity turns are especially vulnerable to deletion. Minimal repair-relevant signals show similar behavior. In KPN027, the token *mh?*, functioning as a repair initiator or signal of

trouble, is not transcribed. As already observed for filled pauses and backchannels, nasal tokens such as *mh* are systematically suppressed, likely due to their brevity and low acoustic salience.

```
TO086: |[perché lui in genere gli aneddoti non li
TO085: |

TO086: | butta a caso li dice:: per]  [[li le~]   ]
TO085: |                              [[ma chi]?  ]
event: |                              [OR         ]

TO086: | [luca ]
TO085: |
```

Figure 6: Example of an other-initiated repair mismatched in TOA3010, KIP (Mauri et al., 2019).

By contrast, matched cases tend to share clear structural and prosodic properties. In KPS023, the repair initiator *che COSA?* is preserved. Unlike the previous examples, it is prosodically salient (as highlighted by the capital letters signaling high volume in CA format) and occurs outside direct overlap, forming a clearly bounded turn. This suggests that acoustic prominence and sequential independence increase the likelihood of preservation.

6. Conclusion and Future Work

Decoding optimization exerts only a limited influence on overall transcription accuracy (RQ1). Across configurations and subsets, Word Error Rate remains broadly comparable between the WER-based and the Interaction-aware strategies. Incorporating interaction-sensitive components into the objective function does not degrade global accuracy, but it does not substantially improve event preservation either. Variation appears to be driven more by conversation-specific factors (e.g., acoustic conditions, overlap density) than by decoding objective. The absence of statistically significant differences between optimization strategies further supports this interpretation, suggesting that decoding-time adjustments alone may not be sufficient to overcome the structural limitations of ASR systems in representing interactional phenomena.

Recognition patterns are strongly phenomenon-dependent (RQ2). Self-repairs are often preserved in linearized form, with the repair solution retained and the disfluent material deleted. Backchannels show intermediate robustness, especially when produced outside overlap, whereas filled pauses are systematically suppressed. Other-initiated repairs remain particularly vulnerable when short or acoustically weak. Structural and acoustic properties appear to be playing a more decisive role than decoding strategy in determining event survival.

Despite limitations, the study produced a linguistically annotated dataset aligned with ASR output, which provides a valuable resource for further empirical investigation of conversational phenomena and a basis upon which improved evaluation and modeling approaches can be developed. Future work should extend the annotated dataset and explore inter-annotator agreement to strengthen the annotation scheme robustness. Also, a more structurally sensitive computational modeling of interactional events, beyond form-based token lists, would allow evaluation to better reflect sequential function rather than surface preservation. Further research should also explore training-level adaptations, including fine-tuning on interactionally annotated conversational data, to assess whether normalization tendencies can be mitigated beyond inference-time adjustments. Integrating acoustic analysis (e.g., duration, intensity, prosodic prominence) would help clarify whether suppression patterns reflect measurable phonetic vulnerability. Finally, comparative evaluation across different ASR architectures would determine whether the linearization effects documented here are Whisper-specific or characteristic of contemporary end-to-end systems more broadly.

7. Limitations

Despite the methodological care adopted in this study, several limitations must be acknowledged. First, the size of the dataset remains relatively modest. The gold-annotated portion amounts to 8 hours and 40 minutes of speech, while the optimization procedure was conducted on approximately 3 hours of conversational data. Although sufficient for exploratory analysis, this size may limit statistical power, particularly for low-frequency phenomena such as other-initiated repairs. In addition, the data subsets were balanced primarily in terms of duration rather than interactional composition, which may affect the generalizability of the results. Only the first 20 minutes of each conversation were annotated, potentially underrepresenting phenomena that emerge in later stages of interaction. Second, annotation was performed by a single annotator. While consistent criteria were applied, the absence of inter-annotator agreement measures represents a limitation, particularly for perceptually subtle phenomena such as filled pauses and short vocalic items: due to time constraints, multi-annotator validation was not feasible. Third, the weighting parameters of the Interaction-aware objective function were manually defined and not optimized through a systematic search or ablation study. Finally, the analysis is restricted to a single ASR system (Whisper large-v3). Although this model provides strong baseline performance, the findings should be interpreted as model-specific. It remains an open question whether similar biases occur in other Whisper models or ASR systems, or whether they are amplified by large-scale models trained on predominantly non-conversational data.

8. Acknowledgements

We would like to thank Jaka Čibej, who kindly introduced us to INCEpTION and provided essential advice on organizing the annotation scheme.

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Peter Blomsma, Julija Vaitonyte, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16:1–24.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization.

N. Christenfeld, S. Schachter, and F. Bilous. 1991. Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20:1–10.

Eve V. Clark. 2020. Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Francisco Cossavella and Jazmín Cevasco. 2021. The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, 33(2):172–186.

Ronald Cumbal, Birger Moell, José Lopes, and Olov Engwall. 2021. "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. pages 4463–4467.

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse, and Morten H. Christiansen. 2019. Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In *Annual Meeting of the Cognitive Science Society*.

Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics*, 1(1).

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Tanvi Dinkar, Chloé Clavel, and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives.

Paul Drew. 1997. 'open' class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101.

Giolo Fele. 2007. *L'analisi della conversazione*. Il Mulino, Bologna.

Jan Gorisch and Thomas Schmidt. 2024. Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia. ELRA and ICCL.

Gail Jefferson. 2004. Glossary of transcript symbols with an introduction. In Gene H. Lerner, editor, *Conversation Analysis: Studies from the First Generation*, chapter 2, page 13–31. John Benjamins, Amsterdam / Philadelphia.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, pages 5–9, Santa Fe, NM.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems.

Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. 2022. Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing? In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 135–143, Potsdam, Germany. KONVENS 2022 Organizers.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. 2019.

KIParla corpus: A new resource for spoken Italian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 243–249, Bari, Italy. CEUR Workshop Proceedings.

Max Planck Institute for Psycholinguistics. 2025. *ELAN (Version 7.0)*. The Language Archive, Nijmegen. Computer software.

Daniela Mereu, Francesco Cangemi, and Martine Grice. 2024. Backchannels are not always very short utterances. the case of italian multi-unit backchannels. *Journal of Pragmatics*, 228:1–16.

Pilar Pernas and Margarita Borreguero Zuloaga. 2010. Cortesia e scortesia in un contesto di apprendimento linguistico: la gestione dei turni. In Marcello Pettorino, Antonietta Giannini, and Francescamaria Dovetto, editors, *La Comunicazione Parlata 3. Atti Del Congresso Internazionale (Napoli, 23-25 Febbraio 2009)*, volume I, pages 227–247. Università Degli Studi Napoli L'Orientale, Napoli.

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

Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382.

Loredana Schettino and Violetta Cataldo. 2019. Lexicalized pauses in italian. pages 189–192.

Lorenzo Spreafico. 2012. Le pause piene nel parlato plurilingue. In *Lessico e lessicologia: atti del XLIV Congresso internazionale di studi della Società di linguistica italiana (SLI)*, number 56 in Pubblicazioni della Società di linguistica italiana, pages 355–368, Roma. Bulzoni.

Muhammad Umair, Julia Mertens, Saul Albert, and J. Ruiter. 2022. Gailbot: An automatic transcription system for conversation analysis. *Dialogue & Discourse*, 13:63–95.

V.N. Vitale, L. Schettino, and F. Cutugno. 2024. Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model hesitation phenomena. In *Proc. Interspeech 2024*, pages 222–226.

Nigel Ward and Wataru Tsukahara. 2000. Tsukahara, w.: Prosodic features which cue backchannel responses in english and japanese. journal of pragmatics 23, 1177-1207. *Journal of Pragmatics*, 32:1177–1207.

Hiroyoshi Yamasaki, Jérôme Louradour, Julie Hunter, and Laurent Prévot. 2023. Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. In *Interspeech 2019*, pages 3088–3092.

10. Language Resource References

Bernasconi, Beatrice and Goria, Eugenio. 2026. *StraParlaTO*. Università degli Studi di Torino, 0.1.0.

Cerruti, Massimo and Ballarè, Silvia. 2020. *Modulo ParlaTO*.

Mauri, Caterina and Ballarè, Silvia and Zucchini, Eleonora. 2024a. *KIPasti*. distributed via ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", 1.1.0. PID http://hdl.handle.net/20.500.11752/OPEN-2124.

Mauri, Caterina and Ballarè, Silvia and Zucchini, Eleonora. 2024b. *ParlaBO*. distributed via ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", 1.1.0. PID http://hdl.handle.net/20.500.11752/OPEN-2126.

Mauri, Caterina and Goria, Eugenio and Ballarè, Silvia. 2019. *Modulo KIP*.

Zucchini, Eleonora and Ballarè, Silvia and Mauri, Caterina. 2026. *StraParlaBO*. Alma Mater Studiorum - Università di Bologna, 0.1.0.

A. Appendix: Annotation Criteria

Table 5 summarizes the annotated phenomena, tags and examples.

Phenomenon	TAG	Function	Example
Backchannels	BC	Continuers, assessments, agreement, incipient speakership	*mhmh*, *va bene*, *ho capito*
Filled pauses	FP	Hesitation markers	*eh*, *ehm*, *mh*
Self-repair	SR	Same turn, transition space, third position	*vennero anda- ... andarono*
Other-initiated repair	OR	Open and restricted initiators	*eh?*, *che cosa?*

Table 5: Overview of annotated phenomena.

Backchannels (BC) Both vocalic and lexical/multi-unit backchannels were annotated, including continuers, assessments, agreement tokens and incipient speakership. Particular attention was given to short and overlapping items, which are known to be vulnerable to omission in ASR output.

Other-initiated repair (OR) Annotation includes open-class initiators (e.g., *eh?*, *prego?*) and restricted requests (e.g., wh-questions). More elaborate repair sequences were excluded, as they are less prone to omission.

Self-repair (SR) Self-repair was annotated when speakers explicitly reformulated their own talk. This includes truncations followed by correction and repetitions reflecting reconstruction of prior material. Interruptions due to hesitation or overlap were excluded unless they resulted in clear reformulation.

Filled pauses (FP) Filled pauses include vocalic hesitation markers (e.g., *eh*, *ehm*, *mh*), identified through auditory inspection. Ambiguous cases were resolved based on prosodic and interactional cues. When occurring within self-repair, they were annotated both as FP and as part of SR.

General considerations It must be taken into consideration the fact that the KIParla corpus is a modular and incremental resource, and, as such, it was not always possible to control and uniform spelling variation with regard to these phenomena. Also, the project has grown over the years and not all transcription conventions were solidly defined from the start.

B. Appendix: INCEpTION Annotation Interface

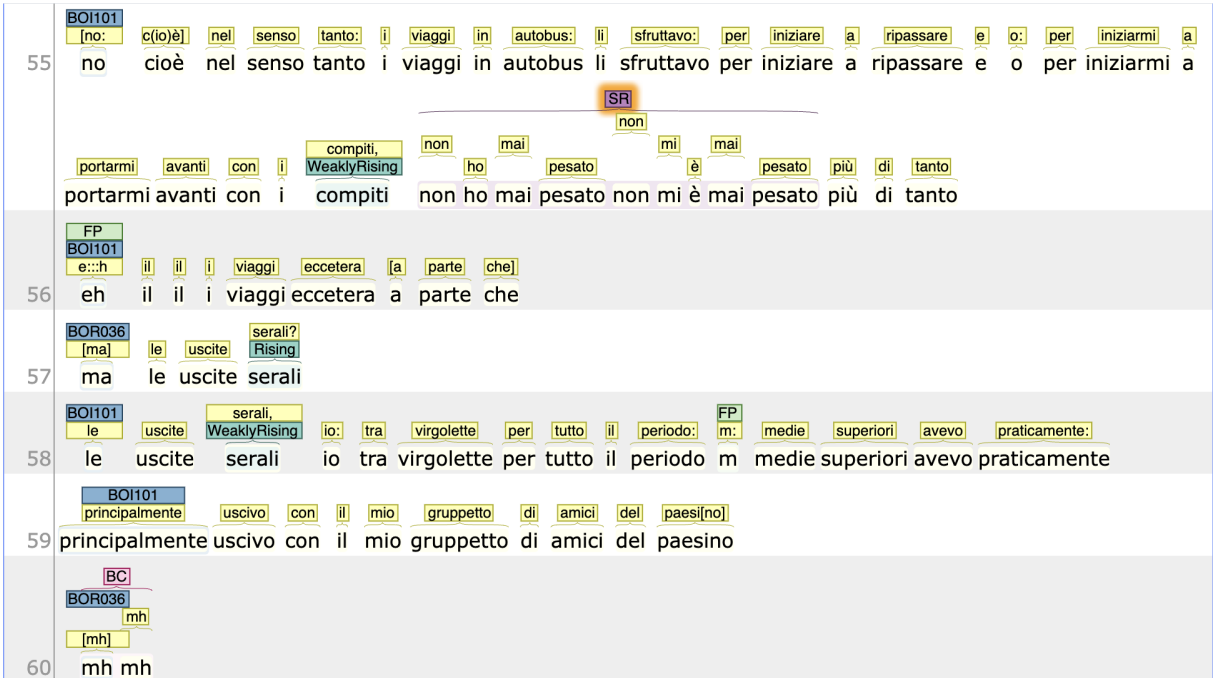

Figure 7: Interface of the INCEpTION annotation environment with annotated phenomena.

C. Appendix: Form-based Event Token List for Optimization

- *eh*
- *ehm*
- *ah*
- *oh*
- *mh*
- *mhmh*
- *mm*
- *okay*
- *sì*
- *bene*
- *esatto*
- *cioè*
- *diciamo*
- *insomma*

D. Appendix: Optimal Parameters

	WER				Interaction-aware WER			
Parameter	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
temperature	0.143	0.636	0.555	0.375	0.053	0.266	0.240	0.158
beam size	3	10	4	4	7	8	8	1
best-of	6	2	5	9	2	1	2	3
no speech threshold	0.203	0.415	0.600	0.249	0.539	0.354	0.617	0.465
compression ratio	2.121	2.501	2.240	2.61	1.734	1.630	2.227	2.492
model	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3	large-v3
condition on previous text	true	true	true	false	false	true	false	true
patience	0.592	0.626	0.344	0.217	0.506	0.308	0.705	0.316
length penalty	0.479	0.616	0.700	0.241	0.241	0.295	0.614	0.397

Table 6: Best-performing trials selected through Optuna optimization for each configuration under the WER-based and Interaction-aware objective functions.

E. Control Analysis on Subset B

type	**WER (%)**	**DEL (%)**	**SUB (%)**	**OK (%)**	**INS (%)**
WER	34.23	57.92	18.75	20.41	2.90
Interaction-aware	33.24	57.73	19.76	19.75	2.75

Table 7: Comparison summary of WER, deletions, substitutions, matches and insertions for the two decoding strategies on Subset B.

F. Appendix: Top 10 Raw Substitutions

WER-based (raw)			Interaction-aware (raw)		
Gold	**Whisper**	**Freq.**	**Gold**	**Whisper**	**Freq.**
eh	e	50	eh	e	49
mh	grazie	26	mh	grazie	28
eh	è	17	eh	è	17
ehm	e	16	mh	e	13
mh	e	15	ehm	e	12
mh	non	13	mh	non	11
mh	è	12	mh	che	10
sì	grazie	11	sì	grazie	10
mh	che	10	mh	è	9
eh	grazie	9	okay	e	9

Table 8: Top 10 raw substitution patterns across optimization strategies, before optimization.