From spoken multilingual data in intercomprehension settings to a searchable corpus: challenges and insights from the OIIC Corpus

Cristiana Cervini, Emanuela Paone – University of Bologna

This presentation focuses on the methodological workflow behind the construction of the *OIIC* corpus (*Oral Interactions in InterComprehension*), a resource of spoken academic peer interactions among plurilingual university students from STEM disciplines. More precisely, participants interact using intercomprehension strategies (De Carlo et al. 2015; Bonvino, Jamet 2016), each using their own language trying to negotiate meaning collaboratively (Varonis, Gass 1985).

Currently, the whole dataset is composed of 42 hours of videorecorded online spoken data from IC courses designed for Management Engineering and Veterinary Science students. So far, the first prototype of the OIIC corpus consists of 3h of peer interactions (approximately 20,000 tokens and around 7,000 annotations), where participants collaborate using Italian, Argentinian Spanish, and Brazilian Portuguese to carry out some tasks related to their field of study.

The initial transcription process is carried out using Whisper (OpenAI) for automatic speech recognition, followed by manual revision to ensure accuracy—especially for overlapping turns or for avoiding neutralization of typical oral traits such as disfluencies, filled pauses, discourse markers, etc. In this regard, we highlight the challenges of managing multilingual and multimodal oral conversational data, particularly in the transcription and annotation processes. A primary difficulty lies in selecting appropriate tools capable of handling the complexity of this type of interaction, characterized by the presence of more than one language in the same turn, by frequent translingual phenomena or by the presence of three different languages in three adjacent turns. ELAN (version 6.9) provides a robust framework for multimodal transcription, yet requires careful attention to the alignment of audio, video, and textual data across multiple languages.

The methodological framework incorporates a multilayered annotation system (Cervini, Zucchini 2024; Cervini, Paone 2024) that captures lexical, metadiscursive, and interactional features. Annotated data is then exported from ELAN and processed for integration into NoSketch Engine using Python scripts. While not originally designed for oral dialogic data, NoSketch Engine has been successfully used to manage corpora such as the KIParla corpus (Mauri et al. 2021; Ballarè, Goria, Mauri 2022) and the interpreting corpus EPTIC (Bernardini et al. 2016).

In our presentation, we aim to describe how the annotation scheme originally presented in Cervini & Zucchini (2024) has been expanded and revised, particularly with regard to the integration of conversational dominance (Itakura 2001) as an analytical dimension.

Publishing the corpus on NoSketch Engine offers the advantage of allowing users to query data by strategy type, opening up possibilities for observing which strategies could be effective in

facilitating conceptual and communicative mediation (Council of Europe, 2018), while also providing access to the corresponding audio and video recordings of peer interactions.

References

Ballarè S., Goria E., Mauri C. (2022), *Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla,* Bologna, Pàtron Editore.

Bernardini S., Ferraresi A. and Miličević M. (2016), "From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective", in *Target* 28, pp. 61-86.

Bonvino E., Jamet M. (eds.) (2016), Intercomprensione: lingue, processi e percorsi. Venezia, Edizioni Ca' Foscari – Digital Publishing. https://edizionicafoscari.unive.it/media/pdf/books/978-88-6969-135-5/978-88-6969-135-5.pdf.

Cervini C., Paone E. (2024), "Comunicare all'università: quando l'interazione orale si fa plurilingue", in *Italiano LinguaDue*, 16(2), pp. 496–523. https://doi.org/10.54103/2037-3597/27861

Cervini C., Zucchini E. (2024), "Caso di studio sull'interazione orale plurilingue in contesto di intercomprensione: dai dati all'analisi", in *Club Working Papers*, Università di Bologna, pp. 207-229.

Council of Europe (2018), *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume with New Descriptors,* Strasbourg, Council of Europe Publishing.

De Carlo M. (coord.) (2015), *Un Référentiel de compétences de communication plurilingue en intercompréhension* (REFIC), MIRIADI - Mutualisation et Innovation pour un Réseau de l'Intercompréhension à Distance, Lyon, APICAD: Association internationale pour la promotion de l'intercompréhension à distance.

ELAN (Version 6.9) [Computer software], (2024), Nijmegen, Max Planck Institute for Psycholinguistics. Retrieved from https://archive.mpi.nl/tla/elan

Itakura H. (2001), "Describing conversational dominance", in *Journal of Pragmatics* 33/12, pp. 1859–80.

Mauri C., Ballarè S., Goria E., Cerruti M. (2021), "Il corpus KIParla", in Cresti, E., M. Moneglia (eds.), *Corpora e Studi Linguistici. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana (Online, 8-10 settembre 2021),* Milano, Officinaventuno, pp. 109-118.

Varonis E. M., Gass S. (1985), "Non-native/Non-native Conversations: A Model for Negotiation of Meaning", in *Applied Linguistics* 6/1, pp. 71–90.