Working with translingual spoken data: the NomadLingo Corpus

Novella Tedesco, Silvia Bernardini, Cristiana Cervini – University of Bologna

This contribution focuses on the main stages involved in building NomadLingo, a corpus of spoken translingual interactions collected in informal social settings at several digital nomad communities (Woldoff & Lichtfield 2021) around Europe — namely, data collection, transcription, annotation, and deposit. Developed as part of the PhD project Fluid Languages Observatory (FLO, thenomadlinguist.eu/flo), the corpus documents the communicative practices of highly mobile, plurilingual speakers in naturally occurring conversations.

NomadLingo comprises approximately 12 hours of recordings and transcripts involving 50 participants, gathered during ethnographic fieldwork in Madeira and the Canary Islands in 2024. Data collection followed ethical guidelines approved by the Bioethical Committee of the University of Bologna, ensuring participant consent, data protection, responsible storage and sharing practices. The project specifically addresses the challenges of working with informal, spontaneous spoken data in transcultural settings characterised by semiotic fluidity, unpredictability, intense variation, challenges to normalised standards and co-constructed discourse (Hepp 2015).

The linguistic environment of the corpus is best understood through the lens of translanguaging theories (García & Wei 2014), which conceptualize communication not as switching between separate languages but as drawing on an integrated semiotic repertoire. While English as a Lingua Franca (Centoze & Taronna 2019) is often a common ground, interaction frequently involves trans- and plurilingual practices, like intercomprehension (Capucho 2017), unveiling speakers' ability of drawing on a flexible repertoire to communicate effectively beyond cultural and linguistic boundaries.

For corpus compilation, transcription was carried out semi-automatically using OpenAI Whisper and refined manually through adapted Jeffersonian conventions to capture salient features of spoken interaction such as pauses, overlaps, and shifts in tone (Jefferson 1984). Contextual metadata (e.g. nationality, language repertoire, speakers present at each communicative event) are provided in accompanying CSV files and directly annotated in a version of the corpus. Metadata management was held through the software Lameta (Hatton et al. 2021); the metadata scheme was developed following standardized guidelines, inspired by, among the others, Burnard 2004, Windhouwer & Wright 2012, Liu & Russo 2024, Paquot et al. 2024. The XML-based format adheres to FAIR principles, promoting reusability and accessibility (Wilkinson et al. 2016). To ensure long-term preservation and interoperability, the corpus is being prepared for deposit in a certified CLARIN B-centre, allowing researchers to access, cite, and reuse the data through a trusted infrastructure.

The annotated version of the corpus also includes three further annotation layers describing language and communication practices. The *translanguaging* level is inspired by translanguaging theories (García & Wei 2014), tagging instances of code-switching, code-mixing, and on-the-fly translation. While we acknowledge that terms such as code-switching and code-mixing are theoretically problematic—since they may endevour a monolithic view of languages (Makoni & Pennycook, 2007)—we adopt them as operational categories. This pragmatic choice allows us to highlight and quantify instances of ouvert language fluidity for statistical and corpus-based analysis, while still remaining grounded in a translanguaging perspective where language – or *languaging* – is seen as dynamic, hybrid, and socially situated

(Love 2017, Wei 2018). Preliminary observations suggest that these translingual practices contribute positively to fluency in intercultural trancultural peer-group informal exchanges. Moreover, the *incomprehension* and *interaction* levels, adapted by Cervini and Paone 2024, highlight strategies of meaning negotiation and co-construction.

By combining corpus methodologies with ethnographic sensitivity (Tusting 2020) and a thoughtful attention to Open Science practices and FAIR principles, NomadLingo suggests a model for capturing and analysing linguistically rich, underrepresented forms of spoken interaction. This paper thus contributes to current discussions on how to document language in use beyond formal, standardized, monolingual contexts.

References

Burnard, L. (2005). Metadata for Corpus Work. *Developing Linguistic Corpora: a Guide to Good Practice* , ed. M. Wynne. Oxford: Oxbow Books: 30-46. Available online from http://ota.ox.ac.uk/documents/creating/dlc/.

Cervini, C., & Paone, E. (2024). COMUNICARE ALL'UNIVERSITÀ: QUANDO L'INTERAZIONE ORALE SI FA PLURILINGUE . Italiano LinguaDue, 16(2), 496–523. https://doi.org/10.54103/2037-3597/27861.

Hepp, A. (2015). Transcultural Communication. Wiley-Blackwell.

García, O., & Wei, L. (2014). Translanguaging: Language, Bilingualism and Education. Palgrave Macmillan.

Jefferson, G. (1984). Transcript notation. In J. M. Atkinson & J. Heritage (Eds.), Structures of Social Action: Studies in Conversation Analysis (pp. ix–xvi). Cambridge University Press.

Liu, L., & Russo, I. (2024). CLARIN Recommendations on Metadata for Spoken Corpora. In Proceedings of CLARIN Annual Conference 2024.

Love, N. (2017). On languaging and languages, *Language Sciences, Vol 61,* pages 113-147, https://doi.org/10.1016/j.langsci.2017.04.001.

Makoni, S., & Pennycook, A. (2012). Disinventing multilingualism: From monological multilingualism to multilingua francas. In The Routledge handbook of multilingualism (pp. 451-465). Routledge.

Paquot, M. König, A., Stemle, E., & Frey, J.-C. (2024). Making corpora FAIR: Metadata for spoken learner corpora. International Journal of Learner Corpus Research, 10(1), 1–28. https://doi.org/10.1075/ijlcr.24010.paq

Tusting, K. (2020). The Routledge Handbook of Linguistic Ethnography. Routledge.

Wei, Li. (2018). Translanguaging as a Practical Theory of Language, Applied Linguistics, Volume 39, Issue 1, Pages 9–30.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016).

The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, Article 160018. https://doi.org/10.1038/sdata.2016.18

Woldoff, R. A., & Litchfield, R. C. (2021). Digital Nomads: In Search of Meaningful Work in the New Economy. Oxford University Press.