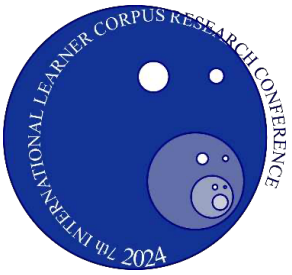
Book of abstracts
LCR2024 Tartu



UNIVERSITY OF TARTU
Institute of Foreign Languages and Cultures

Learner Corpus Association

JOHN BENJAMINS PUBLISHING COMPANY

Discussing the categorization of speakers' language background: implicit assumptions and methodological challenges for Learner Corpus Research

Lopopolo, Olga (Eurac Research, University For Foreigners of Perugia), Arianna Bienati (Eurac Research, Università di Modena e Reggio Emilia), Jennifer-Carmen Frey (Eurac Research), Aivars Glaznieks (Eurac Research) and Stefania Spina (University for Foreigners of Perugia)

The categorization of speakers' language background is one of the core types of information needed in Learner Corpus Research (LCR). The choice of terms like 'L1', 'native speaker' and 'mother-tongue' in the different corpus research projects leads to a complex and sometimes controversial web of linguistic inquiry. The adoption of specific terms, subsequently integrated in the form of speakers' metadata in learner corpora, may be rooted in underlying theoretical paradigms not consistently clarified in corpus description papers. Consequently, the adoption of specific terms often carries implicit meanings, leading to a lack of shared understanding among researchers. This lack of consensus regarding the intended meanings of these labels can result in diverse interpretations, affecting not only the coherence and reliability of learner corpus studies, but also the comparability of the results when adopting different categorization types.

In the present contribution we aim at addressing three key issues: (1) the impact of implicit assumptions and methodological choices taken in learner corpus design when categorizing speakers' language background, (2) the consequences that different conceptualizations of speakers' language background might have on results and (3) the integration of alternative perspectives on speakers' language categorization in LCR. Through a comprehensive review of prominent learner corpora, we (1) identify the most common operationalizations of the learner_L1 metadata (Paquot et al. 2023) and scrutinize the underlying theoretical assumptions. Our exploration then extends to (2) the Italian and German subsections of LEONIDE (Glaznieks et al. 2022). We analyzed the learner texts with a range of complexity measures obtained with CTAP (Chen & Meurers 2016; Okinina et al. 2020) and compared group effects on the results by varying the constitution of learner and reference groups according to the operationalization found in (1). Drawing on research on plurilingualism, we (3) problematize the practice of classifying speakers' language backgrounds in ways that may not align with the complex reality of multilingual societies. We therefore propose a re-evaluation of language classification systems for LEONIDE that might better reflect speakers' language experiences in multilingual environments.

References

Chen, X.B. & Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity. Osaka, Japan. The International Committee on Computational Linguistics.

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of Italian, German and English. International Journal of Learner Corpus Research, 8(1), 97–120.

Okinina, N., Frey, J.-C, & Weiss, Z. (2020). CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 7123–7131). European Language Resources Association.

Paquot, M., König, A., Stemle, E. & Frey, J.-C. (2023). Core Metadata Schema for Learner Corpora, https://doi.org/10.14428/DVN/4CDX3P