From Research Design to Reusable Data: Lessons from (Longitudinal) Learner Corpus Projects

Matthias Schwendemann
Friedrich-Alexander-Universität Erlangen-Nürnberg

Longitudinal L2-studies provide valuable insights into second-language development, but they also present a range of methodological and practical challenges. This talk addresses key decisions involved in designing and implementing longitudinal data collections with heterogeneous groups of language learners. Particular attention will be paid to the development of tasks that can be used across proficiency levels while producing comparable yet sufficiently rich data. Further topics include the collection of language-biographical metadata, participant management, data preparation, and the ways in which processing and annotating decisions may shape subsequent analyses.

Drawing on experiences from infrastructure initiatives like the DAKODA project and other corpus projects, including ZuMult and Expoko, the talk will discuss successful practices as well as challenges, mistakes, and lessons learned. Rather than presenting a single blueprint, it will highlight the trade-offs involved in creating longitudinal datasets that are methodologically sound, manageable in practice, and reusable for future research.