Linguistic diversity across languages and registers: A corpus-linguistic basis for investigating emerging grammars in language-contact situations

Martin Klotz
Humboldt-Universität zu Berlin

Annika Labrenz
Humboldt-Universität zu Berlin

Anke Lüdeling
Humboldt-Universität zu Berlin

Heike Wiese
Humboldt-Universität zu Berlin

Findings from language variation in language-contact situations can be difficult to interpret, since it is challenging to obtain suitable comparisons. For noncanonical patterns found in the language use of bilingual speakers particularly, it is important to take into account bilingual speakers' full repertoires, including informal and formal registers, and to use matching repertoire data from monolingual speakers for comparisons. We present methods of data elicitation and corpus-linguistic processing that allow for this, for the example of adolescent and adult heritage speakers of four languages (Greek, Russian, Turkish, German) in the context of two majority languages (English in the US, and German in Germany) and their monolingual counterparts in five countries (Greece, Russia, Turkey, US, Germany), and discuss methodological implications.

Our data has been obtained with the method described in Wiese (2017) ("Language Situations"), which yields naturalistic, yet controlled and comparable productions for informal and formal, written and spoken registers, covering key domains of speakers' repertoires (cf. also Biber and Conrad, 2009).

To allow for qualitative and quantitative linguistic analyses and systematic, broad-scale comparisons of potential new options and noncanonical patterns, all data is integrated into a single, unified, multi-layer corpus using the ANNIS corpus search engine (Krause, Leser, and Lüdeling, 2016). The corpus features multiple annotation layers of different kinds, multiple segmentations, and aligned multi-modal data. We present the corpus architecture and discuss challenges for the corpus-linguistic infrastructure posed by the integration of data from different languages, scripts, and registers including computer-mediated written language and informal spoken language. We illustrate potential conflicts between standardized representations and the explorative approach towards grammatical patterns. We show how the corpus supports the exploration and analysis of new grammatical options in cross-linguistic and within-language investigations across registers and speaker groups. By accessing rich metadata not only potential new dialects can be identified by grouping grammatical patterns, but also crucial extra-linguistical factors can be argued for as properties of the speaker communities.

References

Biber, Douglas and Susan Conrad (2009). Register, Genre, and Style. Cambridge: CUP.

Krause, Thomas, Ulf Leser, and Anke Lüdeling (2016). "graphANNIS: A Fast Query Engine for Deeply Annotated Linguistic Corpora". JLCL 31.1, pp. 1–25.

Wiese, Heike (2017). Language Situations: A method for capturing variation within speakers' repertoires. To appear in: Yoshiyuki Asahi (ed.), Methods in Dialectology XVI. Frankfurt a. M.: Peter Lang.