Spoken Language Corpora at the Leibniz Institute for the German Language: Established Tools and New Developments

Mark-Christoph Müller, Elena Frick & Henrike Helmer
Leibniz-Institut für Deutsche Sprache, Mannheim
{mark-christoph.mueller, frick, helmer}@ids-mannheim.de

While corpora of **written text** in machine-readable form are readily available for researchers in (computational) linguistics, the situation for transcribed **spoken language** is radically different, because the peculiarities of spontaneous spoken language make the creation of high-quality, standardized, and versatile corpora a complex task. As the leading research data centre for German spoken language corpora, the **Archive for Spoken German** (*Archiv für Gesprochenes Deutsch* (AGD), agd.ids-mannheim.de), located at the IDS Mannheim, is dedicated to making corpora of spoken German available to the international research community in a sustainable way.

The **Database for Spoken German** (*Datenbank für Gesprochenes Deutsch* (DGD), Schmidt 2017, dgd.ids-mannheim.de), launched in 2012, is an established search and browsing platform providing access to (currently) 40 corpora from the AGD (approx. 4700 hours of audio and video recordings and manually created time-aligned transcripts with 20 mio. tokens). The DGD uses a rich data model which is specifically tailored to the requirements of spoken language. Each *transcribed* token has annotations for the corresponding *normalized* and *lemmatized* forms, which captures and allows to query e.g. dialectal pronunciation variants of the same lemma. Each token also has an entry from a set of POS-Tags specifically enhanced for spoken language, with tags for e.g. hesitation and response particles, discourse markers, and tag questions. These primary data are enriched by corpus *meta* data like e.g. speaker demographics, interactional settings, and conversation topics.

In addition, the **ZuMult** project (Fandrych et al. 2022, zumult.org) has developed a suite of web applications to complement the DGD with new features for searching and browsing transcripts. By extending the DGD data model with CQP and ISO/TEI standards, they allow to query time-based span annotations as well as typical spoken language phenomena, e.g. speaker changes and overlaps or paraverbal events like laughter, coughing, and pauses.

References: • Schmidt, T. (2017): DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. In *Zeitschrift für germanistische Linguistik* 45.3., Berlin/Boston: de Gruyter, 451–463. • Fandrych, C., Frick, E., Kaiser, J., Meißner, C., Portmann, A., Schmidt, T., Schwendemann, M., Wallner, F. & Wörner, K. (2022): ZuMult: Neue Zugangswege zu Korpora gesprochener Sprache. In Kämper, H. & Plewnia, A. (Hgg.): *Sprache in Politik und Gesellschaft. Perspektiven und Zugänge*. Berlin/Boston: de Gruyter (Jahrbuch des Instituts für Deutsche Sprache 2021), 305–312.