60. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 05. bis 07. März 2024

Gesprochenes Deutsch: Struktur, Variation, Interaktion

METHODENMESSE

Mittwoch, 06. März 2024, 16:15 Uhr bis 18:00 Uhr

## Mündliche Korpora und ihre Erschließung am Archiv für Gesprochenes Deutsch

*Siegwalt Lindenfelser, Silke Reineke (IDS)*

Das Archiv für Gesprochenes Deutsch (AGD) (Stift/Schmidt 2014) archiviert Korpora und Sammlungen, die mündlichen Sprachgebrauch in verschiedensten gesellschaftlichen Kontexten über einen Zeitraum von den 1950er Jahren bis heute dokumentieren. Es stellt damit die weltweit größte Datenbasis zur Erforschung von gesprochenem Deutsch bereit. Die Bestände gliedern sich im Wesentlichen in

- **Gesprächskorpora** (z. B. Gesprochene Wissenschaftssprache Kontrastiv (GWSS), das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), Belgische TV-Debatten (BETV), Jugendkommunikation (JK), Mitschreiben in Vorlesungen (MIKO)),
- **Interviewkorpora** (z. B. Bonner Längsschnittstudie des Alterns (BLSA), Deutsch von Türkeirückkehrern (DTRK), Emigrantendeutsch in Israel (IS)),
- **Sprachvarietätenkorpora** aus dem deutschsprachigen Kerngebiet (z. B. Deutsch Heute (DH), Deutsche Mundarten: Zwirner-Korpus (ZW)) sowie von extraterritorialen Varietäten (z. B. Deutsch in Namibia (DNAM), Mennonitenplautdietsch in Nord- und Südamerika (MEND)) sowie
- **Korpora zu Mehrsprachigkeit und Spracherwerb** (z. B. Hamburg Maptask Corpus (HMAT); Mehrsprachige Kinder im Vorschulalter (MEKI)).

Aktuell stellt das AGD 40 Korpora über die Datenbank für Gesprochenes Deutsch (DGD) für Forschungs- und Lehrzwecke zur Verfügung. Einen Teil der Daten stellt das AGD über einen persönlichen Service bereit. Die in der DGD zugänglichen und für eine Nachnutzung aufbereiteten Korpora haben einen Umfang von mehr als 4.700 Stunden Audioaufnahmen, rund 200 Stunden Videoaufnahmen und 20 Millionen transkribierten Tokens.

Die Korpora am AGD sind unterschiedlich tief erschlossen. **Metadaten** zu Sprecher/-innen und Sprechereignissen unterstützen Recherchen zum gezielten Auffinden einzelner Datensätze sowie die Erstellung virtueller Korpora. Die transkribierten Daten enthalten in der Regel **Annotationen** auf drei weiteren Ebenen: Normalisierung, Lemmatisierung und Part-of-Speech-Tagging mit einem für gesprochene Sprache adaptierten Tagset (Westpfahl/Schmidt 2013).

Zur **tieferen Erschließung** auf weiteren Annotationsebenen wurden kürzlich etwa die ortsbezogenen Metadaten der Sprachvarietätenkorpora am AGD mit der Wissensbasis Wikidata verlinkt und darauf aufbauend **geo-referenziert**. Der sukzessive Ausbau der Erschließung ermöglicht neue Analysewege im Rahmen der Erforschung von gesprochenem Deutsch.

Schließlich bietet das AGD als Forschungsdatenzentrum auch **Beratung und Unterstützung für prospektive Datengeber/-innen** bei der Planung und Durchführung von Datenerhebungen nach aktuellen Standards an und führt laufend die Integration neuer Korpora in die Bestände des Archivs durch, um sie so zur Nachnutzung in der Forschungsgemeinschaft verfügbar zu machen.

## Literatur


Stift, Ulf-Michael / Schmidt, Thomas (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (Hrsg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Redaktion: Melanie Steine, Franz Josef Berens. Mannheim: Institut für Deutsche Sprache, 360–375.

Westpfahl, Swantje / Schmidt, Thomas (2013): POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics 1, 139–156.