Datenbank für Gesprochenes Deutsch (DGD)

,
thomas.schmidt@ids-mannheim.de
IDS Mannheim, Deutschland

Eine Korpusplattform für die Arbeit mit mündlichen Daten

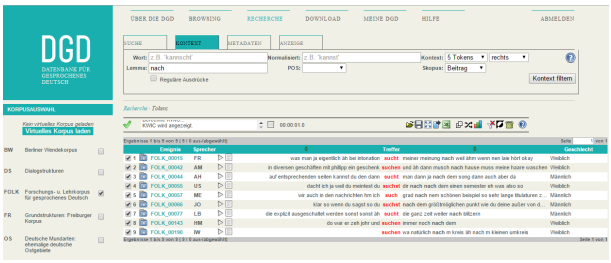
Die Datenbank für Gesprochenes Deutsch (DGD)(vgl. Institut für Deutsche Sprache; Schmidt 2014a) ist die zentrale Plattform für den Zugriff auf Daten des Archivs für Gesprochenes Deutsch (AGD). Über die DGD werden 23 mündliche Korpora des Deutschen im Gesamtumfang von mehr als 3000 Stunden Audio und 8 Millionen transkribierter Wort-Tokens angeboten.

Der Bestand umfasst erstens mehrere große Variationskorpora des Deutschen, insbesondere das Korpus „Deutsche Mundarten“ (Zwirner-Korpus) mit mehreren Satelliten-Korpora nach dem gleichen Design („Deutsche Mundarten: Ehemalige deutsche Ostgebiete“, „Deutsche Mundarten: DDR“ und mehrere kleinere, regional begrenzte Sammlungen von Dialektaufnahmen) sowie das Korpus „Deutsche Umgangssprachen“ (Pfeffer-Korpus). Diese Dokumentation binnendeutscher Mundarten wird komplementiert durch Korpora auslandsdeutscher Varietäten, z. B. das Korpus „Australiendeutsch“ und drei Korpora zum Emigrantendeutsch in Israel.

Zweitens bietet die DGD Zugriff auf verschiedene Gesprächskorpora, u. a. das Berliner Wendekorpus, die Korpora „Grundstrukturen“ (Freiburger Korpus) und „Dialogstrukturen“, sowie das Korpus „Elizitierte Konfliktgespräche“. Mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b) wird im AGD ein großes, breit stratifiziertes Gesprächskorpus des Deutschen aufgebaut, das technisch und methodisch auf dem Stand aktueller bester Praktiken ist und der wissenschaftlichen Gemeinschaft ebenfalls über die DGD zur Verfügung gestellt wird.

Die Bestände sind nach einem einheitlichen XML basierten Metadatenschema dokumentiert und durch Transkriptions- und Annotationsdaten, die ebenfalls auf einem gemeinsamen XML-Datenmodell basieren, für die wissenschaftliche Analyse erschlossen.

Die DGD erlaubt zum einen ein explorativ es Browsen auf diesen Daten. Korpus-, Sprecher- und Ereignisdokumentationen können eingesehen und die zugehörigen Audiodateien online abgespielt werden. Mit dem Audio alignierte Transkripte werden dem Benutzer in einer HTML5-basierten Darstellung präsentiert, die das Anspringen beliebiger Stellen im Transkript und das synchronisierte Abspielen der entsprechenden Segmente der Aufnahme ermöglicht. Diese Form des Zugangs dient sowohl dem Kennenlernen der Datenbestände als auch dem Einstieg in deren qualitative Analyse.



Für die gezielte Auswertung der Daten in quantitativer Hinsicht bietet die DGD zum anderen mehrere Recherchefunktionen. Über eine strukturierte Metadatensuche können nach flexibel spezifizierbaren Kriterien (z. B. Gespräche mit Sprechern aus dem norddeutschen Raum, älter als 40 Jahre) Teilmengen des Gesamtbestands ausgewählt und als virtuelle Korpora gespeichert werden. Die strukturierte Tokensuche erlaubt korpuslinguistische Anfragen über mehrere Annotationsebenen (Transkription, orthographische Normalisierung, Lemmatisierung, POS-Annotation), deren Ergebnisse in vielfältiger Hinsicht kontextualisiert (d. h. mit Metadaten korreliert, auf Transkript- und Aufnahmekontext rückbezogen) werden können.

Für die weitere Bearbeitung von Ausgangsdaten oder Analyseergebnissen bietet die DGD schließlich verschiedene Möglichkeiten zum Download von Datensätzen oder geeigneten Ausschnitten.

Bei allen Funktionen zum Browsen und Durchsuchen der Daten legt die DGD Wert darauf, korpusgesteuerte Analysemethoden zu ermöglichen, in denen Hypothesen aus den Daten selbst generiert und in einer interaktiven Auseinandersetzung mit selbigen schrittweise verfeinert werden können.

Die DGD ist seit Ende 2012 online und hat mittlerweile mehr als 4000 registrierte Nutzer_innen aus Forschung und Lehre. Datenbestände und Funktionalität werden kontinuierlich erweitert.

Bibliographie

Institut für Deutsche Sprache (IDS) (o. J.): *DGD.* Datenbank für Gesprochenes Deutsch http://dgd.ids-mannheim.de [letzter Zugriff 12. Februar 2016].

Schmidt, Thomas (2014a): "The Database for Spoken German – DGD2", in: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland 1451-1457 http://www.lrec-conf.org/proceedings/lrec2014/pdf/171_Paper.pdf [letzter Zugriff 12. Februar 2016].

Schmidt, Thomas (2014b): "The Research and Teaching Corpus of Spoken German – FOLK", in: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland 383-387 http://www.lrec-conf.org/proceedings/lrec2014/pdf/290_Paper.pdf [letzter Zugriff 12. Februar 2016].