Thomas Schmidt/Joachim Gasch/Julia Kaiser (Mannheim)

DGD – die Datenbank für Gesprochenes Deutsch

Die Datenbank für Gesprochenes Deutsch (DGD; <https://dgd.ids-mannheim.de>; Schmidt 2014a) stellt der wissenschaftlichen Gemeinschaft einen großen Teil des Bestandes des Archivs für Gesprochenes Deutsch (AGD; <http://agd.ids-mannheim.de>; Stift/Schmidt 2014) zur Verfügung. Das AGD ist die zentrale Sammelstelle für Korpora des gesprochenen Deutsch am IDS. Gegründet als Deutsches Spracharchiv (DSAv) im Jahre 1932 hat es über Eigenprojekte, Kooperationen und Übernahmen von Daten aus abgeschlossenen Forschungsprojekten einen Bestand von über 80 mündlichen Korpora aufgebaut.

Den Grundbestand des Archivs stellt das unter Leitung des Archivgründers Eberhard Zwirner zwischen 1955 und 1970 aufgebaute Varietäten-Korpus „Deutsche Mundarten“ (ZW in der DGD) dar. Die Korpora „Deutsche Mundarten: ehemalige deutsche Ostgebiete“ (OS), „Deutsche Mundarten: DDR“ (DR), „Deutsche Mundarten: Südwestdeutschland und Vorarlberg“ (SV) sowie „Deutsche Mundarten: Schwarzwald“ (SW) folgen einem ähnlichen Design und können als Ergänzungen bzw. „Nachverdichtungen“ des Zwirner-Korpus aufgefasst werden. Die Korpora „Deutsche Umgangssprache“ („Pfeffer-Korpus“, PF) und „Deutsche Standardsprache“ („König-Korpus“, KN) komplementieren die Mundartenkorpora um Daten zur umgangssprachlichen und standardnahen Variation des Deutschen. Weitere Variationskorpora – wie „Australiendeutsch“ (AD) oder „Russlanddeutsche Dialekte“ (RUDI) – nehmen wie auch drei Korpora zum Emigrantendeutsch in Israel (IS, ISW, ISZ) auslandsdeutsche („Sprachinsel-“)Varietäten in den Blick.

Mit den Korpora „Grundstrukturen“ („Freiburger Korpus“, FR) und „Dialogstrukturen“ (DS) enthält die DGD zwei der frühesten Korpora des Deutschen mit interaktionslinguistischer Ausrichtung. Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK; Schmidt 2014b) wird seit 2008 am IDS mit dem Ziel aufgebaut, der Forschungsgemeinschaft ein großes, breit diversifiziertes Gesprächskorpus zur Verfügung zu stellen, das aktuellen methodischen, technischen und rechtlichen Best Practices genügt. FOLK umfasst aktuell (April 2018) etwa 230 Stunden Audio- und Videoaufnahmen sowie gut 2,3 Millionen transkribierter Tokens. Die aufgezeichneten Gesprächsereignisse umfassen verschiedenste Bereiche privater (z.B. Tischgespräche, Telefongespräche, Gesellschaftsspiele, Interaktion beim gemeinsamen Kochen), institutioneller (z.B. Schulunterricht, Prüfungsgespräche, berufliche Meetings, Service-Interaktionen) und öffentlicher (Podiumsdiskussionen, öffentliches Schlichtungsgespräch) Interaktion.

<https://doi.org/10.1515/9783110622591-022>

Über die DGD werden diese Daten sowie einige weitere mündliche Korpora der wissenschaftlichen Öffentlichkeit für die Nutzung in Forschung und Lehre zugänglich gemacht. Die DGD bietet zwei Zugangsmodi für die Korpora an:

1. Der Browsing-Modus gibt einen Überblick über den Datenbestand und ermöglicht das explorative Ansehen bzw. Anhören von Transkripten und Aufnahmen sowie Metadaten zu (Sprech-)Ereignissen und Sprecher(inne)n. In diesem Modus können Nutzer(innen) sich mit den Daten vertraut machen und Einzelanalysen anhand von Transkripten vornehmen.

Abb. 1: Browsing: Video mit Transkript

2. Der Recherche-Modus ermöglicht das gezielte Suchen in den Datenbeständen und die Analyse der resultierenden Suchergebnisse. Über eine Recherche auf Metadaten können Nutzer(innen) nach Sprechereignissen mit bestimmten Eigenschaften suchen. Auf diese Weise können virtuelle Teilkorpora gebildet werden – in FOLK etwa mit Gesprächen im privaten Umfeld, an denen männliche Sprecher zwischen 20 und 40 beteiligt sind, aber z.B. auch im Zwirner-Korpus mit Datensätzen, in denen Sprecher(innen) aus einer bestimmten Region (z.B. „Schleswig-Holstein, Hamburg oder Niedersachsen“) aufgezeichnet wurden.

Das Kernstück der DGD bildet die Token-Suche. Diese erlaubt eine gezielte Abfrage der Daten über vier Annotationsebenen (transkribierte Form, orthografisch normalisierte Form, Lemma, POS-Tag), wobei auch Kombinationen (z.B. norma-

lisierte Form „meine“ mit Lemma „meinen“ oder weitere Formen im Kontext des Suchausdrucks) und reguläre Ausdrücke (z.B. „ver.+“ für Formen, die mit dem Präfix „ver-“ beginnen) möglich sind. Das Ergebnis einer solchen Recherche wird als Keyword-in-Context-(KWIC-)Konkordanz angezeigt, aus der heraus bei Bedarf zu jedem Treffer auch das zugehörige Audio abgespielt oder der betreffende Transkriptausschnitt angezeigt werden kann (siehe Abb. 2).

The screenshot shows the DGD search interface. At the top, there are tabs for POSITION, TOKEN, KONTEXT, METADATEN, and ANZEIGER. Below these, there are input fields for 'Transkribiert:' (containing 'z.B. 'kannscht''), 'Normalisiert:' (containing 'meine'), 'Lemma:' (containing 'meinen'), and 'POS:' (containing 'z.B. 'VMFIN''). There is also a checkbox for 'Reguläre Ausdrücke' and a 'Suche starten' button. Below the search bar, there is a section titled 'Recherche - Token:' with a status bar indicating 'KWIC wird angezeigt.' and a timer '00:00:00'. The main area displays a table of search results, showing 'Ergebnisse 1 bis 29 von 1284 (1284 / 0 aus- (abgewählt))'. The table has columns for 'Ereignis', 'Sprecher', and 'Treffer'. The first few rows are visible, showing results for 'FOLK_00003' and 'FOLK_00004'.

Ereignis	Sprecher	Treffer
1	FOLK_00003 JS	weil es is ja ich meine mh es is schon interessant aber ich kann einiges ablesen
2	FOLK_00003 DM	und eigentlich also das findet man ich meine sie ham ja sehr viel gemacht bei tarlara oder phonetik
3	FOLK_00003 JS	würden sie den ausweg sehen also ich mein jetz nich indem sie neu selbst eigene modelle machen äh
4	FOLK_00003 JS	dass man unterschiedliche möglichkeiten nutzt ich mein ja jetz ham wa das sehr schön ausgedrückt man kann
5	FOLK_00004 TH	mein wenn der zum beispiel den ganzen tuch nur gut gelaunt
6	FOLK_00004 TF	ja gut ich mein ich war da verwöhnt ich hab
7	FOLK_00004 MS	ich mein weil
8	FOLK_00006 RZ	nee ich mein von de von de höhe her

Abb. 2: Tokensuche nach normalisierter Form „meine“ mit Lemma „meinen“, Ergebnis als KWIC-Konkordanz

Wie mittlerweile gut 8.000 Registrierungen für die DGD (seit Ende 2012, Stand: 6.4.2018) bezeugen, besteht seitens Studierender, Lehrender und Forschender ein großes Interesse an den Möglichkeiten, die die DGD zum Arbeiten mit mündlichen Korpora bietet.

Durch regelmäßige Updates werden die Korpusbestände in der DGD kontinuierlich erweitert. So wächst etwa FOLK jährlich um neue Aufnahmen im Umfang von etwa 30 Stunden, und seit Ende 2017 sind auch die deutschsprachigen Daten des Korpus „Gesprochene Wissenschaftssprache“ (GeWiss) in der DGD verfügbar. Aktuell wird im AGD das Korpus „Deutsch heute“ (DH) für eine Veröffentlichung in der DGD vorbereitet, womit den klassischen älteren Variationskorpora des IDS dann eine sehr umfangreiche (mehr als 1.000 Stunden Aufnahmen umfassende) systematische Datensammlung zur regionalsprachlichen Variation des Deutschen im 21. Jahrhundert zur Seite gestellt wird. Mit jedem Update wird außerdem auch die Funktionalität der Plattform erweitert und verbessert.

Literatur

- Schmidt, Thomas (2014a): The database for spoken German – DGD2. In: Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC 2014). Reykjavik, S. 1451–1457.
- Schmidt, Thomas (2014b): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 15, S. 196–233. Internet: www.gespraechsforschung-online.de/fileadmin/dateien/heft2014/px-schmidt.pdf (Stand: 9.7.2018).
- Stift, Ulf-Michael/Schmidt, Thomas (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim, S. 360–375.