Abteilung Pragmatik – Programmbereich Mündliche Korpora

INSTITUT FÜR
DEUTSCHE SPRACHE

DIE DATENBANK FÜR GESPROCHENES DEUTSCH: DGD

Thomas Schmidt

DATEN UND ANNOTATIONEN

Abbildung 1: Anzeige eines Transkripts aus dem Zwirner-Korpus in der DGD

Über die Datenbank für Gesprochenes Deutsch (DGD) stellt das Archiv für Gesprochenes Deutsch (AGD) am IDS Mannheim der wissenschaftlichen Community einen großen Teil seiner Datenbestände bereit. In der aktuellen Version bietet die DGD Zugriff auf **24 Korpora gesprochener Sprache**, die insgesamt über **9000 Interaktionen, 3000 Stunden Audio-Aufnahmen oder 8.5 Millionen transkribierter Wort-Tokens** umfassen.

Unter den Korpora befinden sich variationslinguistische Korpora wie das **Korpus „Deutsche Mundarten" (Zwirner-Korpus)** oder das **Korpus „Deutsche Umgangssprachen" (Pfeffer-Korpus)** sowie Gesprächskorpora wie das **Korpus „Dialogstrukturen"** oder das **„Freiburger Korpus"**. Aktuell baut das AGD außerdem das **Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)** auf, ein breit stratifiziertes Gesprächskorpus des Deutschen, das nach aktuellen texttechnologischen Standards erstellt und ebenfalls über die DGD bereitgestellt wird. Die Transkripte in FOLK werden nach den Vorgaben des Gesprächsanalytischen Transkriptionssystem (GAT) für Minimaltranskripte in literarischer Umschrift mit dem Editor **FOLKER** erstellt und in Abständen von ca. 3-5 Sekunden mit der zugehörigen Aufnahme aligniert. Zur eigentlichen Transkription kommen mit einer orthographischen Normalisierung sowie einer Lemmatisierung und einem POS-Tagging (STTS mit TreeTagger) drei weitere Annotationsebenen.

Transkription	da	gehst	de	jetzt	einfach	über	dem	bild
Normalisierung	da	gehst	du	jetzt	einfach	über	dem	Bild
Lemmatisierung	da	gehen	du	jetzt	einfach	über	d	Bild
POS	ADV	VFIN	PPER	ADV	ADJD	APPR	ART	NN

Abbildung 2: Annotationsebenen im FOLK-Korpus

BROWSING, QUERYING UND DOWNLOAD

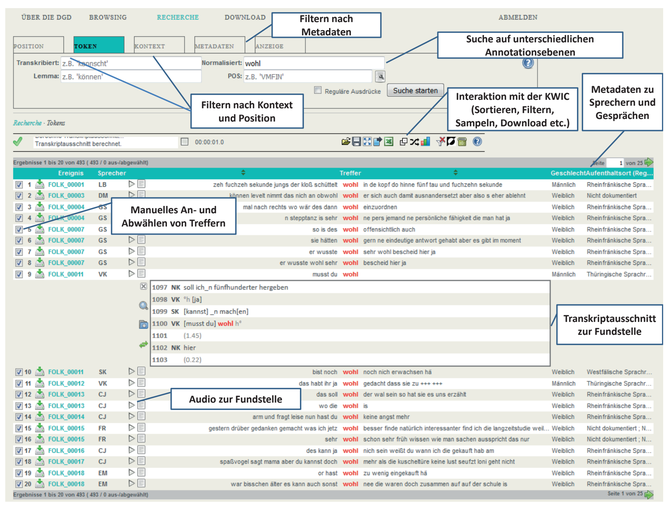

Abbildung 3: Anzeige eines Query-Ergebnisses als KWIC-Konkordanz

Die DGD ermöglicht ein Browsen auf **Metadaten, Audioaufnahmen und Transkriptionen** der verschiedenen Korpora, ein gezieltes Durchsuchen der Datenbestände sowie die Möglichkeit zum Download von Korpusausschnitten und ausgewählten kompletten Datensätzen.

Der **Menüpunkt „Browsing"** bietet die Möglichkeit, Metadaten, Transkripte und Zusatzmaterial anzusehen sowie Audioaufnahmen ausschnittsweise anzuhören. Für jeden Dokumenttyp werden Übersichtslisten mit einigen zentralen Angaben angeboten, die Nutzern ohne Erfahrung mit der Datenbasis eine erste Orientierung ermöglichen.

Im **Menüpunkt „Recherche"** stehen verschiedene Möglichkeiten zur systematischen Suche auf den Daten zur Verfügung:

Volltextsuchen bieten dem Benutzer bei einfacher Bedienung einen schnellen Überblick über die textuellen Inhalte großer Datenbestände. Dokumentstrukturen wie z.B. XML Markup, XML Attribute, hierarchische Beziehungen von Inhalten etc. werden ausgeblendet, wodurch dem Benutzer das Suchen in den Korpusdaten ohne detaillierte Kenntnisse der zugrunde liegenden Datenstrukturen ermöglicht wird.

Während die Volltextsuche auf dem reinen Text der Transkripte basiert, nutzt die **struktursensitive Suche** auch die in den XML-Daten kodierten Auszeichnungen und Annotationen (s.o.). Somit wird es möglich, auch normalisierte und lemmatisierte Formen in die Suchanfrage einzubeziehen und bei der Darstellung des Suchergebnisses gezielt auf zusätzliche zur Fundstelle gehörige Information, wie z.B. Metadaten zum betreffenden Sprecher, zuzugreifen. Verschiedene Filter (für Kontext und Position der *tokens*) und weitere Methoden erlauben ein schrittweises Verfeinern von Suchergebnissen.

Schließlich beinhaltet die DGD Funktionalität zum strukturierten **Durchsuchen von Metadaten**, deren Ergebnisse zum benutzerdefinierten **Erstellen virtueller Korpora** genutzt werden können.

Über den Menüpunkt **„Download"** können ausgewählte komplette Datensätze aus verschiedenen Korpora zur Weiterverarbeitung auf dem eigenen Rechner heruntergeladen werden. Nutzer haben die Möglichkeit, eigene **Kollektionen** von Ausschnitten für Ihre Fragestellungen zu bilden und beliebige solcher Ausschnitte ebenfalls auf den eigenen Rechner herunterzuladen.

AUSBLICKE

Die DGD bewegt sich vier Jahre nach der Veröffentlichung der ersten Beta-Version mittlerweile (Stand: Februar 2016) auf die Marke von **5000 registrierten Nutzern** zu. Dies zeigt den großen Bedarf, der in der sprachwissenschaftlichen Lehre und Forschung an Korpora gesprochener Sprache besteht.

Mit der aktuellen Version der DGD ist nun ein stabiler Standard für die Datenverarbeitung und Datenbereitstellung im Archiv für Gesprochenes Deutsch geschaffen, auf dessen Grundlage die derzeit angebotenen Datenbestände in Zukunft erweitert, vervollständigt und überarbeitet werden können, sowie die Funktionalität zum Einsehen und Durchsuchen der Daten ausgebaut werden kann.

Die wichtigsten **Erweiterungen der Datenbestände** in nächster Zukunft betreffen erstens **FOLK**, das weiterhin jährlich um mindestens 30 Stunden transkribierter Gesprächsaufnahmen wachsen wird. Zweitens wurde mit der finalen Aufbereitung des IDS-Korpus **Deutsch Heute** – einer Sammlung, die in über 1000 Stunden Aufnahmen systematisch die regionale Variation des Deutschen am Anfang des 21. Jahrhunderts dokumentiert – begonnen. Das Korpus wird ab 2018 über die DGD verfügbar sein. Drittens übernimmt das AGD kontinuierlich **Korpora aus abgeschlossenen Forschungsprojekten** (zuletzt etwa das Korpus „Mehrsprachige Kinder im Vorschulalter") die nach entsprechender Aufbereitung über die DGD verfügbar gemacht werden.

Im Hinblick auf die **Funktionalität der DGD** stehen die Erweiterung für die Arbeit mit **Video-Daten** sowie die Überarbeitung des **POS-Tagging** (beides zunächst in FOLK) an. Eine kürzlich abgeschlossene **Nutzerstudie** hat uns wertvolle Erkenntnisse zu Nutzerprofilen und Usability-Aspekten geliefert, die wir in bei der weiteren Entwicklung der Plattform berücksichtigen werden. Darüber hinaus ist eine schrittweise Integration der DGD in **digitale Infrastrukturen** (insbesondere CLARIN) geplant.

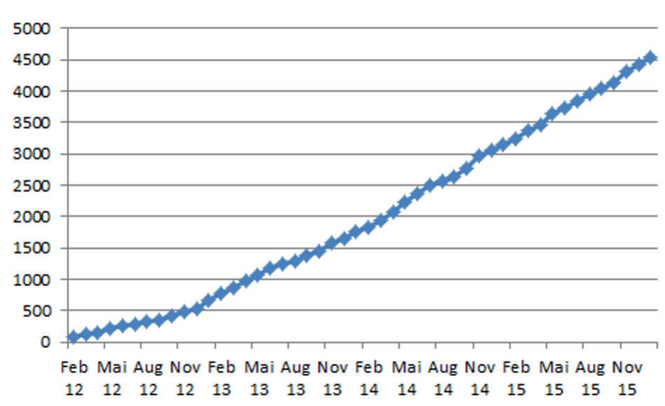

Abbildung 4: Entwicklung der DGD-Nutzerzahlen

DGD
DATENBANK FÜR GESPROCHENES DEUTSCH

http://dgd.ids-mannheim.de

Kontakt
Postadresse:
Dr. Thomas Schmidt
Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim

Tel.: +49 621 1581-239
Fax: +49 621 1581-200
thomas.schmidt@ids-mannheim.de

Hausadresse:
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
Deutschland
Tel.: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft