Originalveröffentlichung in: Beißwenger, Michael/Lemnitzer, Lothar/Müller-Spitzer, Carolin (Hrsg.): Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium. – Paderborn: Wilhelm Fink, 2022. S. 421-430. (UTB 5711)

POSTPRINT

Thomas Schmidt

25. Korpora gesprochener Sprache

Korpora gesprochener Sprache bestehen aus Audio- oder Videoaufnahmen sprachlicher Produktionen, die über eine Transkription einer linguistischen Analyse zugänglich gemacht werden. Sie kommen zur Untersuchung unterschiedlichster sprachwissenschaftlicher Fragestellungen unter anderem in der Gesprächsforschung, der Dialektologie und der Phonetik zum Einsatz. Dieser Beitrag diskutiert die wichtigsten Eigenschaften von Korpora gesprochener Sprache und stellt einige Vertreter der verschiedenen Kategorien vor.

1. Einleitung

Ein großer Teil der empirischen Arbeit in den Sprachwissenschaften ist noch immer an der Schriftsprache ausgerichtet. So bestehen etwa die größten Referenzkorpora des Deutschen (*DeReKo* und *DWDS-Kernkorpus*, → Kapitel 24 [Korpora geschriebener Sprache] in diesem Band) fast ausschließlich aus schriftsprachlichem Material, und die Korpuslinguistik setzt nicht selten die Begriffe „Sprache" und „Text(e)" synonym. Es gibt allerdings eine ganze Reihe sprachlicher Phänomene, die sich gar nicht, oder zumindest nicht alleine, anhand der Schriftsprache empirisch untersuchen lassen. Dazu gehören:

- alle Aspekte der sprachlichen Face-to-Face- und medial vermittelter Interaktion, die vor allem in verschiedenen Richtungen der Gesprächsforschung (Konversationsanalyse, Interaktionale Linguistik, Funktionale Pragmatik u. W.) untersucht werden,
- Aspekte der dialektalen, regionalen und soziolektalen Variation von Sprache, die zwar nicht ausschließlich, aber doch vornehmlich in der Mündlichkeit zu beobachten sind,
- Spracherwerbs- oder -lernprozesse beim kindlichen Erstspracherwerb und beim Zweitsprachenerwerb von Kindern und Erwachsenen,
- alle akustischen/lautlichen Aspekte von Sprache, also insbesondere der Gegenstandsbereich von Phonetik und Phonologie,
- Multimodalität im Sinne des Zusammenspiels zwischen Sprache, Gestik und Mimik,
- Bestimmte Klassen lexikalischer Elemente, die bevorzugt oder gar ausschließlich in der gesprochenen Sprache oder der mündlichen Interaktion vorkommen wie Gesprächspartikel, Modalpartikel, Interjektion, Diskursmarker,
- die Dokumentation von Sprachen ohne Schriftkultur.

Um diese Aspekte von Sprache empirisch zu untersuchen, braucht es Korpora gesprochener Sprache – also Aufzeichnungen mündlicher Sprachproduktionen auf Audio oder Video, die durch eine Transkription und weitere Maßnahmen für eine Auswertung erschlossen werden.

Entsprechend den vielfältigen Fragestellungen, die mit ihrer Hilfe untersucht werden, fallen unter den Überbegriff „Korpora gesprochener Sprache" ganz verschiedenartige Sammlungen von Sprachdaten. Dieser Beitrag gibt nach der Diskussion einiger grundlegender Begriffe einen Überblick über die wichtigsten Typen von Korpora des gesprochenen Deutsch und stellt einige ausgewählte Ressourcen kurz vor.

2. Grundlegendes

Die Primärdaten von Korpora gesprochener Sprache sind Audio- oder Videoaufnahmen entweder von natürlichen sprachlichen Inter-

aktionen oder von sprachlichen Produktionen, die vom Forscher elizitiert, also gezielt hervorgerufen, werden. Recherchier- und analysierbar werden die Primärdaten erst durch eine Verschriftung – die Transkription. Was und wie transkribiert wird, ist abhängig von Forschungsinteressen und theoretischen Vorannahmen (vgl. den Begriff der „Transcription as Theory" bei Ochs 1979). Für die Arbeit mit Korpora gesprochener Sprache spielt die Transkriptionsweise daher eine zentrale Rolle, denn die Recherchemöglichkeiten sind notwendigerweise durch das verwendete Transkriptionssystem (→ Kapitel 23 [Gesprächsanalytische Transkription] in diesem Band) vorbestimmt.

Wie schriftsprachliche Korpora können Korpora gesprochener Sprache über Korpusportale, die oft mehrere verschiedene Korpora beinhalten, zugänglich gemacht werden. Für Korpora des gesprochenen Deutsch ist die *Datenbank für Gesprochenes Deutsch* (Schmidt 2017), das wohl meistgenutzte Portal. Neben Recherchemöglichkeiten, die auch für Korpora geschriebener Sprache üblich sind, insbesondere Suchfunktionen und die Darstellung der Ergebnisse in Konkordanzen, beinhalten solche Portale zusätzliche Funktionalität zur Darstellung vollständiger Transkripte und zum Rückgriff auf die zugrunde liegenden Audio- oder Videodaten.

Die Erschließung von Daten gesprochener Sprache – also die Erhebung von Aufnahmen, deren Transkription und Dokumentation – ist im Vergleich zu schriftsprachlichen Korpora äußerst aufwändig. Korpora gesprochener Sprache sind daher um mehrere Größenordnungen kleiner als ihre schriftsprachlichen Pendants: Während Korpora wie DeReKo mittlerweile Tokenzahlen im höheren zweistelligen Milliardenbereich aufweisen, umfassen selbst die größten Korpora gesprochener Sprache nicht mehr als 10 Millionen Tokens.

3. Gesprächskorpora

Bei Gesprächskorpora stehen die Natürlichkeit und Authentizität der aufgezeichneten Sprechsituationen im Vordergrund. Es interessieren nur solche Daten, bei denen die sprachliche Interaktion nicht eigens für den Forschungsanlass inszeniert wurde (wie z. B. bei einem Interview) und bei denen die Teilnehmer*innen weitestgehend so agieren, wie sie es auch ohne den Anlass der Gesprächsaufzeichnung tun würden. Wichtig ist außerdem der Charakter der Interaktivität, also der Umstand, dass mehrere Sprecher aufeinander abgestimmt sprachlich handeln. Das Design von Gesprächskorpora ist zuerst an Eigenschaften der Gespräche (wie: Interaktionsdomäne, Lebensbereich, Zwei- vs. Mehrpersonengespräche, → Kapitel 13 [Daten und Metadaten] in diesem Band) ausgerichtet, erst in zweiter Linie können auch Eigenschaften der daran beteiligten Sprecher*innen (wie: Geschlecht, Alter) eine Rolle spielen.

3.1 Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

Das *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (*FOLK*, Deppermann / Schmidt 2014) wird am Leibniz-Institut für Deutsche Sprache (IDS) in Mannheim aufgebaut. Es ist konzipiert als ein großes Korpus von authentischen, spontanen Gesprächsdaten des Deutschen, die mündliche Interaktionen in möglichst vielfältigen gesellschaftlichen Zusammenhängen und von möglichst unterschiedlichen Gesprächstypen umfassen. Dazu gehören:

- verschiedenste Formen privater Kommunikation wie Tischgespräche, Unterhaltungen bei privaten Aktivitäten (z. B. Kochen, Spaziergang), private Telefongespräche, Spielinteraktionen oder Vorlesen für Kinder;
- Interaktionen aus institutionellen Bereichen wie dem Bildungswesen (z. B. schulische Unterrichtskommunikation, universitäre Beratungsgespräche), dem Servicebereich (z. B. Verkaufsgespräche in der Apotheke oder im Baumarkt, Fahrstunden, s. Abb. 1) oder dem Beruf (z. B. Meeting, Schichtübergabe im Krankenhaus, Bewerbungstraining);

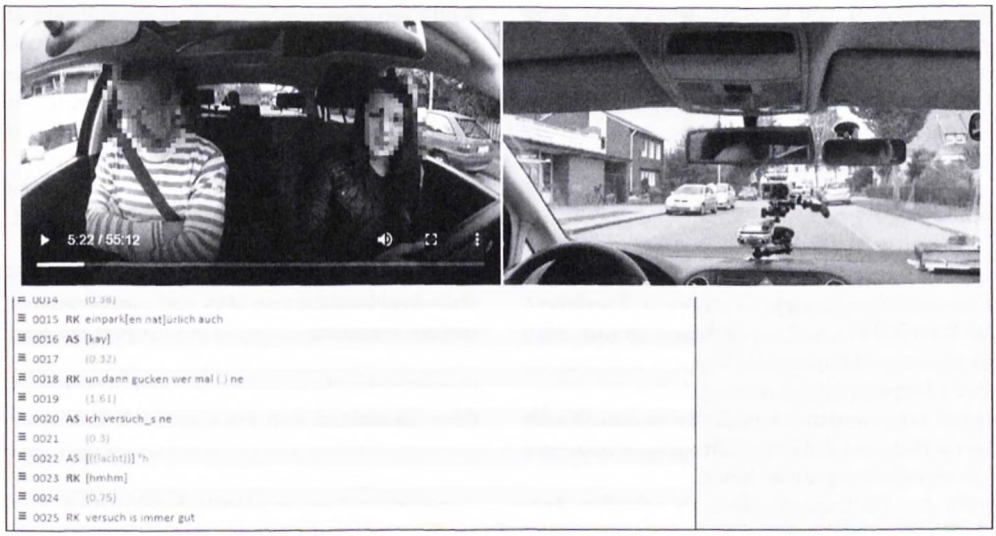

Abb. 1: Videoaufnahme einer Fahrschulstunde mit zugehörigem Transkript in der DGD

- Gespräche aus dem öffentlichen Raum wie Bundestagsdebatten, Podiumsdiskussionen oder Ausschnitte aus den Schlichtungsgesprächen zu Stuttgart 21.

Zu allen in *FOLK* enthaltenen Interaktionen liegen Audio-Aufnahmen vor, zu einem großen Teil auch eine oder mehrere Videoaufzeichnungen. Transkripte in *FOLK* folgen den Vorgaben des cGAT-Minimaltranskripts (Schmidt/Schütte/Winterscheid 2015) und verwenden eine literarische Umschrift, um Besonderheiten des Mündlichen (wie Verschleifungen: „haste", dialektale Formen: „hasch du" und Ähnliches) wiederzugeben.

Für die korpuslinguistische Auswertung wird der transkribierte Text mit Annotationen auf drei Ebenen angereichert: 1. einer orthographischen *Normalisierung*, die jedes transkribierte Wort auf seine standardorthographische Entsprechung abbildet, 2. einer *Lemmatisierung*, die jedem Token seine Grundform zuordnet, sowie 3. einem Part-of-Speech-Tagging, bei dem jeder Wortform eine Angabe zur Wortart nach dem „Stuttgart-Tübingen-Tagset" (STTS) zugeordnet wird (vgl. Westpfahl 2020 und → Kapitel 14 [Annotationen] in diesem Band).

FOLK versteht sich als „Referenzkorpus", das als empirische Grundlage für unterschiedlichste Fragestellungen eingesetzt werden kann. Es ist ein kontinuierlich wachsendes („dynamisches") Korpus. Die erste Vollversion aus dem Jahr 2013 umfasste ca. 100 Stunden Aufnahmen und etwa eine Million transkribierter Tokens. Seitdem wird das Korpus jährlich um neue Daten im Umfang von etwa 30 Stunden erweitert, sodass FOLK in der aktuellen Version (vom Mai 2021) auf gut 314 Stunden Aufnahmen oder 3 Millionen Tokens an-

Tab. 1: Annotationsebenen in FOLK

Transkription	**da**	**gehst**	**de**	**jetz**	**einfach**	**über**	**dem**	**bild**
Normalisierung	da	gehst	du	jetzt	einfach	über	dem	Bild
Lemmatisierung	da	gehen	du	jetzt	einfach	über	d	Bild
POS (STTS)	ADV	VVFIN	PPER	ADV	ADJD	APPR	ART	NN

gewachsen ist. FOLK wird – wie viele weitere der im Folgenden vorgestellten Korpora – über die *Datenbank für Gesprochenes Deutsch* (DGD, https://dgd.ids-mannheim.de) verfügbar gemacht.

3.2 Gesprochene Wissenschaftssprache Kontrastiv (GeWiss)

Das *GeWiss-Korpus* (Fandrych/Meißner/Wallner 2017) wurde zwischen 2009 und 2013 in einem gemeinsamen Projekt der Universitäten Leipzig, Wrocław und Aston (Birmingham) aufgebaut mit dem Ziel, eine empirische Grundlage für die Erforschung gesprochener Wissenschaftssprache (auch) unter dem Aspekt der Mehrsprachigkeit zu schaffen. Wesentlich für das Design des Korpus ist erstens die Unterscheidung dreier Gesprächstypen, die als typisch für die Kommunikation an der Hochschule erachtet werden können:

- universitäre Prüfungsgespräche als abschließender Teil der Qualifikation zum Erwerb eines universitären Abschlusses;
- Expertenvorträge, wie sie üblicherweise im Rahmen von Kolloquien, Workshops oder Konferenzen von Wissenschaftler*innen gehalten werden;
- studentische Vorträge, typischerweise Referate im Rahmen einer Lehrveranstaltung.

Diese drei Gesprächstypen wurden zweitens an allen Standorten, also an Universitäten in den drei Ländern (und zusätzlich in geringerem Umfang in Italien, Bulgarien und Finnland), erhoben, und zwar drittens einmal mit Studierenden bzw. Expert*innen in den jeweiligen Muttersprachen (d.h. auf Deutsch, Polnisch und Englisch) und an den Standorten im Ausland zusätzlich auf Deutsch mit Sprecher*innen, für die Deutsch nicht die Muttersprache ist. Nicht-muttersprachliche Daten liegen außerdem für das Englische (Aufnahmen vom Standort Großbritannien) vor.

GeWiss ist damit ein gesprochensprachliches *Vergleichskorpus*, das auf einige ausgewählte Interaktionstypen beschränkt ist. Sein Design erlaubt systematische Vergleiche ähnlicher Daten über vier verschiedene Dimensionen, genauer:

- den *Gesprächstyp*: Prüfungsgespräch, Expertenvortrag, studentischer Vortrag
- die *Sprache*: Deutsch, Englisch, Polnisch, (Italienisch)
- den *Sprachstatus*: Deutsch (und Englisch) als Erstsprache (=L1) vs. Deutsch (und Englisch) als Zweit- oder Fremdsprache (=L2)
- den *akademischen Kontext*: Deutschland, Großbritannien, Polen, (Bulgarien, Italien, Finnland)

Die Aufnahmen in *GeWiss* wurden nach GAT transkribiert. In seiner ursprünglichen Form ist das vollständige Korpus über eine eigene Plattform an der Universität Leipzig (https://gewiss.uni-leipzig.de/index.php) zugänglich.

Die deutschsprachigen Bestandteile wurden außerdem 2017 in die Datenbank für Gespro-

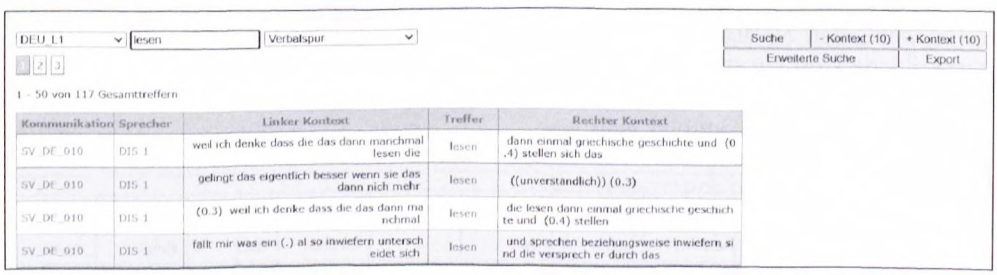

Abb. 2: Suche auf dem GeWiss-Korpus über das Portal der Uni Leipzig

chenes Deutsch integriert,[1] analog zu *FOLK* mit Lemma- und Part-of-Speech-Annotationen versehen und später durch weitere deutschsprachige Daten von einer finnischen Universität erweitert.

3.3 Kiezdeutsch (KidKo)

Das *Kiezdeutsch-Korpus* (Rehbein et al. 2014) entstand zwischen 2008 und 2015 in einem Teilprojekt des Sonderforschungsbereichs Informationsstruktur an der Universität Potsdam. Es hat einen Umfang von knapp 230.000 Tokens und dokumentiert den Sprachgebrauch von Jugendlichen in einem multiethnischen Wohngebiet in Berlin-Kreuzberg, also die Verwendung deutscher Spontansprache unter dem Einfluss von und im Zusammenspiel mit verschiedenen anderen („Herkunfts"-)Sprachen. Als Vergleichsgröße wurde ein Ergänzungskorpus im Umfang von gut 100.000 Tokens mit jugendlichen Sprecher*innen aus einem mono-ethnischen Wohngebiet (Berlin-Hellersdorf) erhoben.

Ausgangspunkt der Erhebung sind 17 bzw. sechs sogenannte „Anker"-Sprecher*innen zwischen 14 und 17 Jahren, deren Interaktion in der „Peer Group" auf Audio aufgezeichnet wurde. Um eine größtmögliche Natürlichkeit der Daten zu erzielen, wurden diese Aufzeichnungen als Selbstaufnahmen, also von

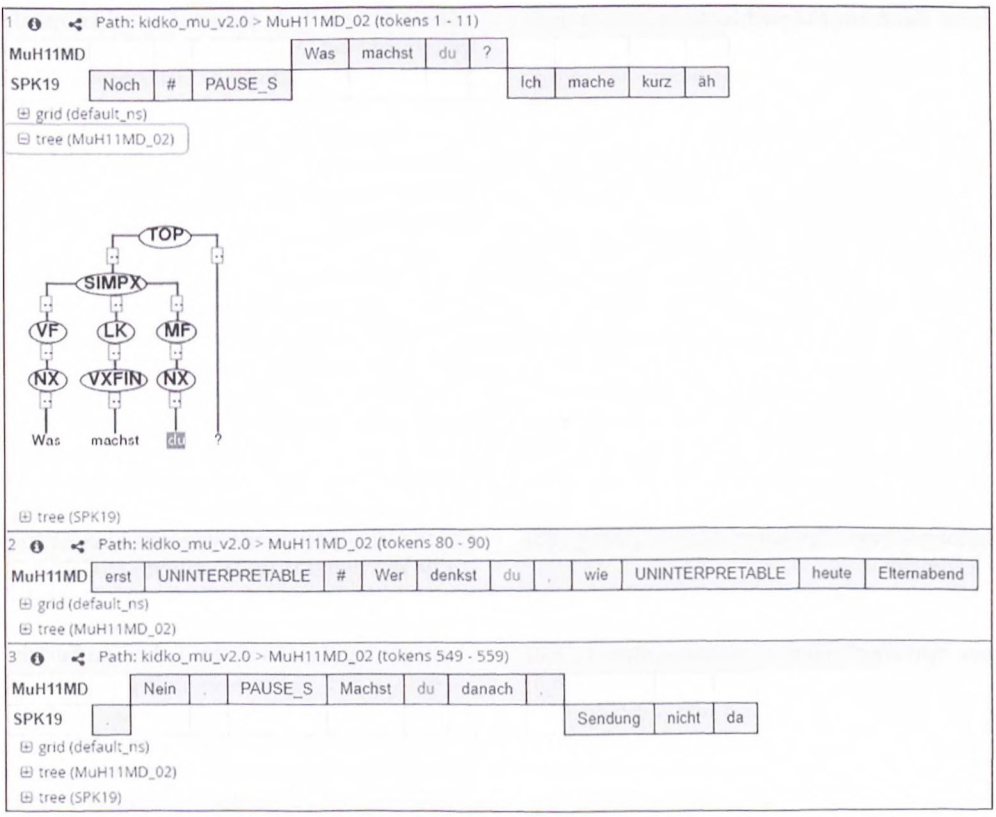

Abb. 3: Abfrage auf dem Kiezdeutsch-Korpus in ANNIS mit Annotationen auf Token-Ebene und Darstellung des Parsingbaumes nach topologischen Feldern

1 Die nicht-deutschsprachigen Bestandteile folgen in einer kommenden Version der DGD.

den Sprecher*innen selbst und ohne die Anwesenheit eines Forschenden, durchgeführt. Die Aufnahmen wurden nach einer an GAT angelehnten Systematik transkribiert. Als zusätzliche Annotationen auf Token-Ebene stehen eine orthografische Normalisierung (z. T. als kommentierte Übersetzungen aus dem Türkischen) sowie ein Part-of-Speech-Tagging zur Verfügung. Eine Besonderheit des Korpus ist, dass es zusätzlich auch nach syntaktischen Strukturen zu Chunks, also größeren syntaktischen Einheiten wie Nominalphrasen, und topologischen Feldern, also syntaktischen Positionen im Satz, annotiert wurde.

Das *Kiezdeutsch-Korpus* wird über das Hamburger Zentrum für Sprachkorpora in der Korpusplattform ANNIS[2] und als Download beim Zentrum für nachhaltiges Forschungsdatenmanagement der UHH[3] verfügbar gemacht. Aus Datenschutzgründen stehen hier allerdings nur die Transkripte, nicht die zugrunde liegenden Audiodaten zur Verfügung.

4. Variationskorpora

Variationskorpora haben ihren Ursprung in der Dialektologie, also der Erforschung der lokalen oder regionalen Variation sprachlicher Formen innerhalb eines Sprachgebiets. Für Variationskorpora werden üblicherweise keine vollständig natürlichen Daten erhoben, sondern es werden gezielte Elizitationsverfahren (s. u.) eingesetzt, um sprachliches Material zu erhalten, das über verschiedene Aufnahmen bzw. Sprecher*innen hinweg gut vergleichbar ist.

Da es bei einem Variationskorpus vor allem darum geht, den Sprachgebrauch einzelner Sprecher*innen zu dokumentieren, tritt auch die Interaktivität der Sprechsituation in den Hintergrund, bzw. es wird im Gegenteil darauf gezielt, vorwiegend monologische Sprache zu erheben. Aus demselben Grund ist das Design von Variationskorpora üblicherweise sprecherzentriert, d. h., die Organisation des Korpus erfolgt primär anhand von Eigenschaften der Sprecher*innen wie Herkunft, Geschlecht, Alter oder Bildungsgrad.

4.1 Deutsche Mundarten (Zwirner-Korpus)

Das Korpus *Deutsche Mundarten* (Zwirner/Bethge 1958) wurde ab 1955 in einem groß angelegten Projekt unter der Leitung von Eberhard Zwirner aufgebaut und ist daher auch als *Zwirner-Korpus* bekannt. Ziel war es, durch eine systematische Erhebung von Tondaten die Dialekte des Deutschen möglichst vollständig zu dokumentieren. Dazu wurden das Gebiet der damaligen Bundesrepublik sowie einige angrenzende Regionen (Vorarlberg, Elsass, Teile der Niederlande) in ein Netz sogenannter „Planquadrate" von 16km x 16km unterteilt. In jedem dieser Planquadrate wurde ein ländlicher Aufnahmeort ausgewählt und in diesem Ort wurden drei Sprecher*innen unterschiedlicher Altersstufen ausgesucht (um 20, 40, über 60 Jahre alt), mit denen verschiedene, für die Dialektologie übliche Erhebungen durchgeführt wurden. Dazu gehören:

- das „Übersetzen" hochdeutscher „Wenkersätze" (vorgegebene Sätze, die darauf ausgelegt sind, bei der Übersetzung phonetische, lexikalische oder syntaktische Variation sicht- oder hörbar zu machen. Der erste Wenkersatz lautet z. B. „Im Winter fliegen die trockenen Blätter in der Luft herum.", vgl. dazu etwa Fleischer 2017) in den jeweiligen Dialekt;
- das Aufzählen der Zahlen von 1 bis 10 und der Wochentage im jeweiligen Dialekt;
- ein narratives Interview, in dem die Sprecher*innen angehalten wurden, frei über Ereignisse in ihrem Leben zu berichten.

Insgesamt wurden für das Korpus fast 6000 Tonband-Aufnahmen im Gesamtumfang von über 1000 Stunden gemacht, die mittlerweile vollständig digitalisiert vorliegen. Für etwa die Hälfte dieser Aufnahmen liegen digitale

2 https://corpora.uni-hamburg.de/annis/kidko.
3 https://www.fdr.uni-hamburg.de/record/8247.

Transkripte vor, womit das Korpus gut 4 Millionen transkribierte Wortformen umfasst. Das Korpus *Deutsche Mundarten* ist über die Datenbank für Gesprochenes Deutsch zugänglich.

4.2 Deutsch Heute (DH)

Mit dem Korpus *Deutsch Heute* (DH, Kleiner 2015) wurde im Projekt „Variation des gesprochenen Deutsch" am IDS Mannheim ab 2005 eine Datensammlung zur Variation des Deutschen aufgebaut, die statt der basisdialektalen Variation, die im 21. Jahrhundert zunehmend an Bedeutung verliert, die großräumigere Variation regionaler „Gebrauchsstandards" in den Blick nimmt. Für die Aufnahmen wurden insgesamt 195 Orte im ganzen Gebiet, in dem Deutsch Amts- und Unterrichtsprache ist (Deutschland, Österreich, Schweiz, Südtirol, Luxemburg, Ostbelgien, Liechtenstein) ausgewählt und dort Sprachaufnahmen zum größeren Teil mit Oberstufenschüler*innen an Gymnasien, zum kleineren Teil mit Personen im Alter von 50 bis 60 Jahren durchgeführt. Die erhobenen Datentypen sind denen des Zwirner-Korpus vom Prinzip her vergleichbar. Sie umfassen:

- Elizitationsaufgaben wie verschiedene Lesetexte, Wortlisten, Bildbenennungs- und Übersetzungsaufgaben;
- spontansprachliche Daten in Form von sprachbiografischen Interviews;
- Aufnahmen sogenannter „Maptasks", bei denen zwei Sprecher*innen gemeinsam eine vorgegebene Aufgabe zur Wegbeschreibung lösen müssen.

Abbildung 4 illustriert einen Ausschnitt aus der Maptask-Aufgabe. Die abgebildeten Gegenstände sind dabei so ausgewählt, dass systematisch regional unterschiedliche Realisierungen auf verschiedenen sprachlichen Ebenen – z. B. „Fleischer" vs. „Metzger" vs. „Schlachter", „Hähnchen" vs. „Broiler" auf der lexikalischen Ebene oder „[Ch]emielaborantin" vs. „[K]emielaborantin" auf der lautlichen Ebene – elizitiert werden.

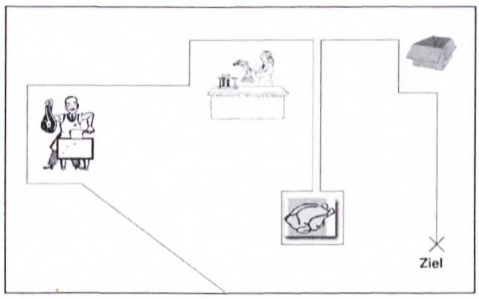

Abb. 4: Elizitationsaufgabe „Maptask" aus Deutsch Heute

Mit einer Gesamtdauer der Aufnahmen von deutlich über 1000 Stunden ist das Korpus *Deutsch Heute* vom Umfang her dem Zwirner-Korpus vergleichbar. Die Korpusdaten werden seit 2018 schrittweise über die Datenbank für Gesprochenes Deutsch verfügbar gemacht. Seit 2021 sind alle Aufnahmen und Transkripte über die DGD abrufbar.

4.3 Weitere Variationskorpora

Nach dem Vorbild des Korpus *Deutsche Mundarten* wurden in den 1960er und 1970er Jahren weitere Datenerhebungen zur Variation des Deutschen durchgeführt, um auch Dialekte außerhalb des Gebiets der damaligen Bundesrepublik zu dokumentieren, um die Datenlage für Gebiete mit besonders kleinräumiger Sprachvariation (vor allem im Südwesten Deutschlands) nachzuverdichten oder um neben der mundartlichen Variation der Basisdialekte auch Material zur Variation des Deutschen in gehobeneren Registern wie der Umgangs- oder Standardsprache zu erhalten. So entstanden die Korpora:

- *Deutsche Mundarten: DDR* mit Aufnahmen von Sprecher*innen aus dem Gebiet der damaligen Deutschen Demokratischen Republik;
- *Deutsche Mundarten: Ehemalige deutsche Ostgebiete* mit Aufnahmen von Sprecher*innen von Dialekten in ehemals zum Deutschen Reich gehörigen Regionen wie Schlesien, Pommern oder Ostpreußen;

- *Deutsche Mundarten: Schwarzwald* und *Deutsche Mundarten: Südwestdeutschland und Vorarlberg* mit besagtem Ziel einer Nachverdichtung für die betreffenden Regionen;
- *Deutsche Umgangssprachen: Pfeffer-Korpus* und *Deutsche Standardsprache: König-Korpus* als Korpora, die, wie *Deutsch Heute*, statt der basisdialektalen Variation standardnähere Varietäten des Deutschen in den Blick nehmen.

All diese Korpora sind Bestandteil der Datenbank für Gesprochenes Deutsch.

4.4 Korpora extraterritorialer Varietäten

Im Zusammenhang mit der Variation des Deutschen sind auch Varietäten interessant, die von Sprachgemeinschaften außerhalb des deutschsprachigen Kerngebiets (Deutschland, Österreich, Liechtenstein, deutschsprachige Gebiete in der Schweiz und in Belgien, Südtirol) gesprochen werden. Diese oft mit dem Terminus „Sprachinselkorpora" (vgl. Boas/Fingerhuth 2018) bezeichneten Ressourcen werden im Gegensatz zu den oben beschriebenen Variationskorpora, die meist das Resultat größer angelegter Erhebungen sind, üblicherweise durch einzelne Forscher*innen erhoben und haben daher einen deutlich geringeren Umfang. Größere, gut erschlossene Korpora zu extraterritorialen Varietäten des Deutschen sind zum Beispiel das *Texas German Dialect Corpus* mit eigener Website an der UT Austin (https://tgdp.org/, Boas 2009) oder das von Michael Clyne erhobene Korpus zum *Australiendeutsch* (Clyne 1981), das Bestandteil der Datenbank für Gesprochenes Deutsch ist. Ebenfalls in der Datenbank für Gesprochenes Deutsch finden sich kleinere Korpora zum Russlanddeutschen, zum Mennonitendeutsch in Nord- und Südamerika und zum Deutsch in Wisconsin, seit 2019 auch ein gut erschlossenes Korpus zum Deutsch in Namibia. Von Aufbau und Inhalt her sind solche Korpora in ihrer Zusammensetzung aus elizitierten Daten und narrativen oder sprachbiographischen Interviews vergleichbar mit den oben besprochenen Variationskorpora zum deutschsprachigen Kerngebiet.

5. Lernerkorpora

Lernerkorpora bestehen aus sprachlichen Produktionen von Sprecher*innen, die eine Erst-, Zweit- oder Fremdsprache erwerben. Ähnlich wie Variationskorpora sind Lernerkorpora in der Regel sprecherzentriert aufgebaut, wobei insbesondere Parameter wie Muttersprache(n) und Erwerbsdauer für das Korpusdesign und die Auswertung interessant sind. Lernerkorpora enthalten üblicherweise eher elizitiertes als vollständig natürliches Material, um Vergleiche zwischen verschiedenen Sprecher*innen zu vereinfachen.

5.1 Hamburg Maptask Corpus, Hamburg Modern Times Corpus, Berlin Maptask Corpus

Das *Hamburg Maptask Corpus* (*HaMaTaC*, Hedeland/Schmidt 2012) und das *Hamburg Modern Times Corpus* (*HaMoTiC*) enthalten Daten fortgeschrittener L2-Lerner*innen des Deutschen. Für *HaMaTaC* wurde den Sprecher*innen die gleiche Maptask-Aufgabe gestellt, die auch für das Korpus *Deutsch Heute* verwendet wurde (s. Abb. 4). Damit ergeben sich korpusübergreifende Vergleichsmöglichkeiten zwischen Muttersprachler*innen und Lerner*innen. *HaMoTiC* enthält Nacherzählungen eines Filmausschnitts von Charlie Chaplins „Modern Times". Auch hier handelt es sich um die Wiederverwendung einer Elizitationsaufgabe aus einem anderen Korpus (Perdue 1993), über die korpusübergreifende Vergleiche ermöglicht werden. Beide Korpora sind über das Hamburger Zentrum für Sprachkorpora und über die DGD zugänglich.

Analog zum *Hamburg Maptask Corpus* wurde das *Berlin Maptask-Corpus* (*BeMaTac*, Sauer/Lüdeling 2016) als ein Vergleichskorpus zwischen L1- und L2-Sprechern des Deutschen erstellt. Die verwendete Maptask-Aufgabe ist wiederum die gleiche, die auch im

Korpus *Deutsch Heute* zum Einsatz kam. *BeMaTac* wird – aufgeteilt in ein L1- und ein L2-Subkorpus und mit umfangreichen Annotationen angereichert – über die Korpusplattform ANNIS an der HU Berlin (https://korpling.german.hu-berlin.de/annis3/) verfügbar gemacht.

5.2 Korpora zum kindlichen Spracherwerb

Korpora zum kindlichen Spracherwerb sind oft longitudinal aufgebaut, d.h. es wird der Sprachgebrauch einiger weniger Kinder über einen längeren Zeitraum in regelmäßigen Abständen aufgezeichnet, sodass die Entwicklung der sprachlichen Fähigkeiten eines einzelnen Kindes über die Zeit nachvollzogen werden kann. Ein typisches Beispiel für ein solches longitudinales Spracherwerbskorpus ist das Korpus *DUFDE* (*Deutscher und Französischer Doppelter Erstspracherwerb*, Meisel 1990), das über das Hamburger Zentrum für Sprachkorpora zugänglich gemacht wird.

Im Kontrast zu Longitudinalstudien stehen Querschnittstudien, die nicht Spracherwerbsprozesse über die Zeit abbilden, dafür aber eine größere Anzahl von Sprecher*innen zu einem gegebenen Zeitpunkt dokumentieren. Ein Beispiel hierfür ist das über die Datenbank für Gesprochenes Deutsch zugängliche Korpus *Mehrsprachige Kinder im Vorschulalter* (*MEKI*, Montanari 2010), das die Sprachkompetenz von Kita-Kindern mit mehrsprachigem Familienhintergrund anhand von elizitierten Erzählungen dokumentiert.

Einige weitere Korpora zum Spracherwerb des Deutschen finden sich auch in der Datenbank des Child Language Data Exchange Systems *CHILDES* (MacWhinney 2000, https://childes.talkbank.org/).

6. Phonetische Korpora

Phonetische Korpora werden vornehmlich für instrumental-phonetische Auswertungen, teilweise auch für die Entwicklung sprachtechnologischer Anwendungen, erstellt und verwendet. Um die hierfür notwendige hohe Aufnahmequalität zu erzielen, werden die Aufnahmen für solche Korpora oft unter Laborbedingungen gemacht, womit Natürlichkeit und Spontaneität der Daten nur noch eine nachgeordnete Rolle spielen können. Die Transkription und meist auch die Alignierung von Transkript und Aufnahme erfolgen in der Regel sehr feingranular – es wird oft mit einem phonetischen Alphabet transkribiert und Zeitmarken werden mindestens auf Wortebene, oft auch auf Phonemebene, gesetzt. Dieser zusätzliche Aufwand für die Transkription hat zur Folge, dass phonetische Korpora in der Regel nur wenige Stunden Material umfassen, das dafür aber sehr detailliert erschlossen ist. Phonetische Korpora zum Deutschen werden vor allem vom Bayerischen Archiv für Sprachsignale (BAS, https://www.phonetik.uni-muenchen.de/Bas/) in München zur Verfügung gestellt. Anders als bei den oben beschriebenen Gesprächs- und Variationskorpora werden die Daten nicht nur über eine Web-Plattform, sondern auch zum Download auf den lokalen Rechner angeboten und können dort dann mit spezialisierter Software wie Praat oder Emu ausgewertet werden (→ Kapitel 27 [Transkriptionswerkzeuge] in diesem Band).

Zum Weiterlesen

Ausführlichere Einblicke in verschiedene der hier angesprochenen Korpustypen geben die Beiträge zu „Mündlichen Korpora" in Kupietz/Schmidt (2018): Kehrein/Vorberger (2018) befassen sich darin mit Variationskorpora, Boas/Fingerhuth (2018) mit Sprachinselkorpora, Draxler/Schiel (2018) mit phonetischen Korpora und Schmidt (2018) mit Gesprächskorpora. Mit Wisniewski (2021) ist eine Publikation in Vorbereitung, die sich speziell mit gesprochensprachlichen Lernerkorpora des Deutschen befasst.

Literatur

Boas, Hans Christian (2009): *The Life and Death of Texas German*. Durham: Duke University Press.

Boas, Hans Christian und Matthias Fingerhuth (2018): Deutsche Sprachinselkorpora im 21. Jahrhundert, in: Kupietz, Marc und Thomas Schmidt (Hrsg.) (2018): *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, S. 125 – 150.

Clyne, Michael (1981*): Deutsch als Muttersprache in Australien. Zur Ökologie einer Einwanderersprache*. In Zusammenarbeit mit dem Centre for Migrant Studies. Monash University. Deutsche Sprache in Europa und Übersee. Band 8. Wiesbaden: Franz Steiner Verlag.

Deppermann, Arnulf und Thomas Schmidt (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik – Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD), in: Domke, Christine und Christa Gansel (Hrsg.): *Korpora in der Linguistik – Perspektiven und Positionen zu Daten und Datenerhebung* [=Mitteilungen des Deutschen Germanistenverbandes 1/2014], S. 4–17.

Draxler, Christoph und Florian Schiel (2018): Moderne phonetische Datenbanken, in: Marc Kupietz und Thomas Schmidt (Hrsg.): *Germanistische Sprachwissenschaft um 2020: Korpuslinguistik*. Vol. 5. Berlin/Boston: De Gruyter. S. 179–208.

Fandrych, Christian, Cordula Meißner und Franziska Wallner (Hrsg.) (2017): *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora*. Tübingen: Stauffenburg.

Fleischer, Jürg (2017): *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 20 Sätzen. Dokumentation, Entdeckungen und Neubewertungen* (= Deutsche Dialektgeographie. Band 123). Olms, Hildesheim / Zürich / New York 2017.

Hedeland, Hanna und Thomas Schmidt (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German, in: Schmidt, Thomas und Kai Wörner (Hrsg.): *Multilingual Corpora and Multilingual Corpus Analysis*. (= Hamburg Studies on Multilingualism 14). Amsterdam: Benjamins, 2012. S. 25–46.

Kehrein, Roland und Lars Vorberger (2018): Dialekt- und Variationskorpora, in: Marc Kupietz und Thomas Schmidt (Hrsg.): *Germanistische Sprachwissenschaft um 2020: Korpuslinguistik*. Vol. 5. Berlin/Boston: De Gruyter. S. 125–150.

Kirk, John M. und Gisle Andersen (Hrsg.) (2016): *Compilation, transcription, markup and annotation of spoken corpora. Special Issue of the International Journal of Corpus Linguistics* [IJCL 21:3], S. 396– 418.

Kleiner, Stefan (2015): „Deutsch heute" und der Atlas zur Aussprache des deutschen Gebrauchsstandards, in: Kehrein, Roland, Alfred Lameli und Stefan Rabanus (Hrsg.): *Regionale Variation des Deutschen. Projekte und Perspektiven*. Berlin u.a. S. 489-518.

Kupietz, Marc und Thomas Schmidt (Hrsg.) (2018): *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter.

MacWhinney, Brian (2000): *The CHILDES project: Tools for analyzing talk*. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Meisel, Jürgen (Hrsg.) (1990*): Bilingual First Language Acquisition: French and German Grammatical Development. Amsterdam*: John Benjamins.

Montanari, Elke (2010): *Kindliche Mehrsprachigkeit – Determination und Genus*. Münster: Waxmann.

Ochs, Elinor (1979): Transcription as Theory, in: Schieffelin, Bambi und Elinor Ochs (ed.): *Developmental Pragmatics*. New York, NY: Academic Press.

Perdue, Clive (ed.) (1993): *Adult Language Acquisition. Vol 1: Field Methods*. Cambridge University Press.

Rehbein, Ines, Sören Schalowski und Heike Wiese (2014): The KiezDeutsch Korpus (KiDKo) Release 1.0, in: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), May 24-31, 2014*. Reykjavik, Iceland.

Sauer, Simon und Anke Lüdeling (2016): Flexible Multi-Layer Spoken Dialogue Corpora, in: International Journal of Corpus Linguistics, Volume 21, Issue 3, 2016, Special Issue: Compilation, Transcription, Markup and Annotation of Spoken Corpora, S. 419–438.

Schmidt, Thomas, Wilfried Schütte und Jenny Winterscheid (2015): *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*, [online] https://nbn-resolving.org/urn:nbn:de:bsz:mh39-46169.

Schmidt, Thomas (2017): DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim, in: *Zeitschrift für Germanistische Linguistik* 45 (3), S. 451–463.

Schmidt, Thomas (2018): Gesprächskorpora, in: Kupietz, Marc und Thomas Schmidt (Hrsg.) (2018): *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, S. 209–230.

Westpfahl, Swantje (2020): *POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Tübingen: Narr, 2020.

Wisniewski, Katrin (Hrsg.) (2021): *Gesprochene Lernerkorpora des Deutschen*. Special Issue der Zeitschrift für Germanistische Linguistik.

Zwirner, Eberhard und Wolfgang Bethge (1958): *Erläuterungen zu den Texten. Spracharchiv, Deutsches. Lautbibliothek der deutschen Mundarten*, Bd. 1. Göttingen: Vandenhoeck & Ruprecht.

Die Adressen aller Webseiten und Online-Ressourcen in diesem Beitrag wurden zuletzt auf Aktualität überprüft am 15. Juni 2021.