

Korpora gesprochener Sprache von/für DaF-LernerInnen

Überblick über mutter- und lernersprachliche Korpora im Kontext von Deutsch als Fremdsprache

PETER PASCHKE

ABSTRACT

The aim of the present paper is to provide an overview (updated to April 2017) of the corpora of spoken German, which are relevant to the teaching of German as a foreign language and freely available on the Internet. These include, on the one hand, the L1 corpora, which are useful for designing syllabuses, developing authentic teaching materials and for direct application in the classroom, and on the other, L2 corpora for the study of second language acquisition. The following aspects are presented and discussed: scope and representativeness of the corpus; types of discourse, speakers and varieties; transcription, annotation and metadata; available query tools and download options; possible uses in the classroom; and publications. With respect to the L1 corpora, FOLK and the corpus "Gesprochene Sprache für die Auslandsgermanistik" provide a sufficient basis for research and teaching. However, it would be desirable to supplement the conversation-analytic approach with phonetically and prosodically annotated corpora. With regard to the web-based L2 corpora (GeWiss and BeMaTaC) the range of communication types (examination interview, student and expert presentation, map task dialogue) and L1 languages (English, Polish, partly Bulgarian and Italian) is quite limited. Here, it would be desirable to set up further corpora to comprehensively investigate the acquisition of German as a foreign or second language.

KEYWORDS

Language corpora, spoken language, German language, German as a foreign language, web resources

1. EINLEITUNG

Der vorliegende Beitrag setzt sich zum Ziel, einen Überblick über Korpora der gesprochenen Sprache Deutsch zu geben (Stand: Frühjahr 2017), die im Kontext der Vermittlung von Deutsch als Fremdsprache (DaF) relevant sind. Ein Interesse an Korpora allgemein ist im Bereich DaF spätestens mit dem korpuslinguistischen Schwerpunktthema der Zeitschrift „Deutsch als Fremdsprache“ (Heft 4/2007 bis 2/2009) nachweisbar. Korpora speziell der gesprochenen Sprache wurden bereits im einführenden Beitrag von Fandrych & Tschirner (2007) gefordert und waren in der Folge Gegenstand mehrerer Einzelbeiträge (Costa 2008, Schneider & Ylönen 2008, Fernández-Villanueva & Strunk 2009).

Fandrych & Tschirner (ebd., 202) unterscheiden muttersprachliche, unterrichtssprachliche und lernersprachliche Korpora. Unterrichtssprachliche Korpora erlauben es, den tatsächlichen sprachlichen Input von Fremdsprachenlernenden zu untersuchen, sind aber meines Wissens für DaF bisher nicht veröffentlicht worden. Die beiden anderen Typen können bestimmten Hauptverwendungszwecken zugeordnet werden (vgl. Fandrych & Tschirner 2007, Gut 2007a, 2007b, Imo 2013):

- Deutsch L1: Korpora mit muttersprachlichen Kommunikationsereignissen dienen der Festlegung von Lerninhalten, der Entwicklung von (authentischen) Unterrichtsmaterialien, dem induktiven Lernen im DaF-Unterricht sowie als Vergleichskorpus für die Untersuchung des Zweitsprachenerwerbs.
- Deutsch L2: Lernersprachliche Korpora dienen vor allem der Untersuchung des Zweitsprachenerwerbs, evtl. auch der Bewusstmachung im Unterricht.

Daneben kann es für vergleichende Studien (oder im Unterricht) sinnvoll sein, auch muttersprachliche mündliche Kommunikationsereignisse von L2-Lernenden zu erfassen bzw. zu analysieren (vgl. z.B. Fernández-Villanueva & Strunk 2009).

Unter „Korpus“ verstehe ich mit Lemnitzer & Zinsmeister (2006: 7) eine „Sammlung schriftlicher oder gesprochener Äußerungen“, die „typischerweise digitalisiert“ vorliegen. Neben den Daten selbst (den Texten) umfassen Korpora möglicherweise Metadaten sowie linguistische Annotationen. In jedem Fall sollten sie repräsentativ sein (ebd., 40f.), etwa für eine bestimmte Varietät oder eine Textsorte/Diskursgattung.

Korpora der gesprochenen Sprache umfassen im Prinzip alle medial mündlichen Gattungen, also spontane, aber auch elizitierte Gespräche sowie vorgelesene Texte (z.B. Wortlisten für Dialektforschung oder phonetische Analysen). Das Hauptinteresse (bei Deutsch L1) gilt aktuell allerdings den Korpora authentischer Gespräche. Deppermann & Schmidt (2014: 4, zit. n. Schmidt 2014: 198) definieren

„Gesprächskorpus“ als „[...] Sammlung von Aufzeichnungen (Audio- und/oder Videoaufnahmen) authentischer Gespräche (i.e. konzeptionell und medial mündlicher, i.d.R. spontaner, Interaktion von zwei oder mehr Teilnehmern), die nach einer wissenschaftlich begründeten und explizit dargelegten Systematik zusammengestellt und über eine Transkription, gegebenenfalls zusätzliche Annotationen und die Dokumentation von Metadaten (zu Gesprächsumständen und beteiligten Sprechern) für eine (sprach-)wissenschaftliche Analyse erschlossen wird.“ Anders als bei Textkorpora mit schriftlichen Texten erfolgt der Zugang über Transkripte, die mit den Primärdaten (Aufnahmen) nicht identisch sind, sondern in einem Abbildverhältnis dazu stehen.¹

Der vorliegende Beitrag beschränkt sich auf Korpora, die im Internet (evtl. nach kostenloser Anmeldung) frei zugänglich sind (Stand Frühjahr 2017)². Umfang, Zugänglichkeit bzw. Möglichkeiten der Abfrage unterscheiden sich jedoch erheblich. Abgeschlossene Projekte verfügen meist nicht über die finanziellen und personellen Mittel, um Webzugang und Abfragetools zu pflegen und an die technologische Entwicklung anzupassen. Da die Zahl online verfügbarer Korpora aber recht beschränkt ist, wurden auch einige ältere Projekte berücksichtigt. Folgende Fragen stehen dabei im Vordergrund:

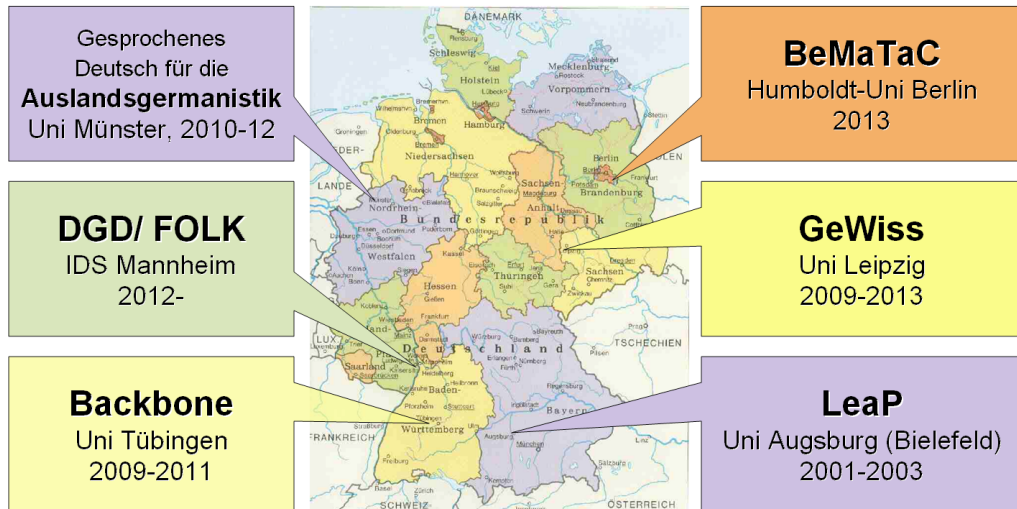
- Welchen Umfang hat das Korpus? Ist es repräsentativ?
- Welche Diskursgattungen sind vertreten? Welche Sprecher und Varietäten?
- Ist das Korpus transkribiert? Welche linguistischen Annotationen und Metadaten gibt es?
- Welche Abfragefunktionen gibt es? Ist das Korpus (sind einzelne Belege) herunterladbar?
- Ist das Korpus im Unterricht einsetzbar? Wie und auf welcher Stufe?
- Welche Veröffentlichungen über/ welche Untersuchungen am Korpus liegen vor?

Abb. 1 gibt eine Übersicht über die vorgestellten Korpora: Die L1-Korpora Gesprochenes Deutsch für die Auslandsgermanistik, DGD/FOLK und Backbone sind Gegenstand von Kapitel 2, die L2-Korpora BeMaTaC, GeWiss und LeaP folgen in Kapitel 3.

1 Zu Korpora gesprochener Sprache vgl. das Themenheft des „International Journal of Corpus Linguistics“ (Kirk & Anderson 2016) sowie Wichmann (2008).

2 Nicht berücksichtigt werden Korpora der gesprochenen Sprache, bei denen keine Audioaufnahmen, sondern nur Transkripte verfügbar sind, wie z.B. „DWDS – gesprochen“ (<http://retro.dwds.de/>), vgl. Káňa (2014: 73).

Abb. 1 – Übersicht über die vorgestellten Korpora: links Deutsch L1, rechts Deutsch L2



2. KORPORA FÜR GESPROCHENES DEUTSCH L1

2.1. Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK) in der DGD³

Die Datenbank Gesprochenes Deutsch (DGD) ist die Internet-Plattform, über die das Archiv für Gesprochenes Deutsch (AGD) des Instituts für Deutsche Sprache (IDS) Mannheim Materialien aus 24 Korpora online zur Verfügung stellt, die ab 1960 überwiegend im Kontext variationslinguistischer Untersuchungen entstanden sind (Jugenddeutsch, Emigrantendeutsch, Mundarten usw.) und sich im Allgemeinen „von authentischen gesprächsanalytischen Daten stark unterscheiden“ (Schmidt 2014: 224). Insgesamt stehen 10913 Sprechereignisse, 10962 Audio- und 102 Videoaufnahmen (mit einer Dauer von über 3100 Stunden) sowie 4571 Transkripte zur Verfügung. Im Folgenden wird aus der umfangreichen Sammlung nur das gesprächslinguistisch orientierte „Forschungs- und Lehrkorpus gesprochenes Deutsch“ (FOLK) vorgestellt.

Das seit 2008 in der Abteilung Pragmatik des IDS Mannheim aufgebaute Korpus FOLK umfasst in seiner im Web veröffentlichten Version derzeit (Stand

³ Die Darstellung berücksichtigt Erweiterungen und Verbesserungen der am 6.4.2017 veröffentlichten Version 2.8.

Abb. 2 – Übersicht über die Korpora der DGD (Screenshot, Ausschnitt)

DGD
DATENBANK FÜR
GESPROCHENES
DEUTSCH

ÜBER DIE DGD
BESCHREIBUNG
BESTAND
NUTZUNGSBEDINGUNGEN
VERSIONEN

RECHERCHE DOWNLOAD MEINE DGD HILFE AB

Über die DGD - Bestand

DGD - Bestand

In den folgenden Tabellen finden Sie Informationen über die gegenwärtig in der DGD verfügbaren Daten: Metadaten, Aufnahmen, Transkripte und Zusatzmaterialien aus 24 Korpora.

Korpora

Korpuskennung	Korpusname	Erhebungszeitraum	Beschreibung
AD	Australiendeutsch	1967	Erzählungen und Bildbeschreibungen von 82 vorwiegend älteren Frauen und Männern, deren Familien z.T. schon seit drei Generationen in Südaustralien leben
BB	Deutsche Mundarten: Kreis Böblingen	1965-1970	Erzählungen und Unterhaltungen von und mit Sprechern unterschiedlichen Alters aus dem Kreis Böblingen
BR	Biographische und Reiserzählungen	1985-1990	In der DDR aufgezeichnete Erzählungen und Interviews von und mit jungen Männern und Frauen
BW	Berliner Wendekorpus	1992-1996	Interviews mit 30 Ostberlinern und 26 Westberlinern (Frauen und Männer) im Alter von 19 bis 55 Jahren über die Themen Mauerfall 1989 sowie individuelle, soziale und gesellschaftliche Aspekte des Alltagslebens

6.4.2017) 259 Gesprächsereignisse mit 730 dokumentierten Sprechern, 295 Tonaufnahmen mit einer Gesamtdauer von 202 Stunden, 95 Videoaufnahmen mit über 77 Stunden Laufzeit, 534 Transkripte mit 1.952.159 laufenden Wörtern und ca. 200 Zusatzmaterialien (Informationen über Settings und Verläufe, Maptask-Karten, Wort- und Lemmalisten). Es ist damit das bei Weitem umfangreichste der hier vorgestellten Korpora und wird zudem stetig erweitert.

FOLK dokumentiert Gesprächsdaten aus unterschiedlichsten Bereichen des gesellschaftlichen Lebens (Arbeit, Freizeit, Bildung, öffentliches Leben, Dienstleistungen usw.) im deutschen Sprachraum. Es handelt sich um authentische, vollständige Gespräche aus „unterschiedlichsten privaten (z.B. Tischgespräche, Gespräche bei Spielinteraktionen), institutionellen (z.B. Schulunterricht, berufliche Kommunikation) und öffentlichen (z.B. Podiumsdiskussionen, Schlichtungsgespräche) Kontexten“⁴. Diese große Variationsbreite, aber auch die Berücksichtigung von Sprechermerkmalen wie regionale Herkunft, Bildungsstand und Alter bei der Erweiterung des Korpus, haben das Ziel, eine

4 http://agd.ids-mannheim.de/FOLK__extern.shtml

Abb. 3 – FOLK-Transkript mit Annotationen: literarische Umschrift nach GAT, orthographische Normalisierung, Lemmatisierung und Part of Speech-Tagging nach dem Stuttgart-Tübingen-Tagset (IDS, Datenbank für Gesprochenes Deutsch (DGD), FOLK_E_00217_SE_01_T_02_DF_01 / c523. Zur Bedeutung der Tags (VVINF, APPR etc.) vgl. <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>. Vgl. auch Schiller et al. 1999)

Annotationen für FOLK_E_00217_SE_01_T_02_DF_01 / c523																	
jetz	hatse	des	hingekriegt	mit	ihrm	e	mail	teil	die	grad	angerufen	hat	weil	des	hat	rr	
jetzt	hat sie	das	hingekriegt	mit	Ihrem	E-Mail-Teil	&	&	die	grad	angerufen	hat	weil	das	hat	m	
jetzt	haben sie	die	hingekriegt	mit	Ihrem	E-Mail-Teil	&	&	die	gerade	anrufen	haben	weil	die	haben	ic	
ADV	VAFIN	PPER	PDS	VVFIN	APPR	NN	NN	XY	XY	PRELS	ADV	VVPP	VAFIN	KOUS	PDS	VAFIN	PF

≡	0522	TZ	bis da hab ich erst (.) immer aufpassen dass das nich irgendwas +++ macht
≡	0523	PZ	guck hier ka_man jetzt sehn ob_s deckt (.) weil (.) °h hier man man wird schon wahrscheinlich nachher noch des unterschied zwischen dem weiß also wo vorher weiß war und °h (.) wo vorher gelb war sehn man bin ich echt froh jetzt hatse des hingekriegt mit ihm e mail teil (.) di[e grad an]gerufen hat °h weil des hat mich ja schon echt schwer genervt vorhin

repräsentative Auswahl aus Gesprächen im deutschsprachigen Raum anzubieten und die Bildung ausgewogener Teilkorpora zu ermöglichen (vgl. ebd.). FOLK ist „primär an den Bedürfnissen von Gesprächsforschern ausgerichtet“ und „bietet [...] als stetig wachsendes und systematisch stratifiziertes Korpus [...] zum ersten Mal einen Referenzpunkt, von dem aus sich gesprächsanalytische Einzelfallanalysen über einen Rückgriff auf eine größere Datenbasis bewerten und absichern lassen“ (Schmidt 2014, 199).

Die Aufnahmen sind nach den Konventionen für GAT-Minimaltranskripte⁵ (also ohne Fokusakzente und Grenztöne) in literarischer Umschrift mit dem (frei verfügbaren) Editor FOLKER transkribiert und in Segmenten von maximal 5 Sekunden mit den Transkripten aligniert. Die Abfrage des Korpus wird durch eine zusätzliche orthographische Normalisierung und eine Lemmatisierung erleichtert (vgl. Abb. 3).

Im Screenshot in Abb. 3 ist im unteren Teil das GAT-Minimaltranskript mit Tilgungen (jetz, ihm), Klitisierungen (hatse), Pausen (.) und Atemvorgängen (°h) zu erkennen. Im oberen Teil sind verschiedene Annotationsebenen wiedergegeben: Die erste enthält wiederum die literarische Umschrift (aber ohne Pausen, Einatmen usw.). Die zweite Ebene zeigt die orthographische Normalisierung (jetzt, ihrem, hat sie), die dritte die Lemmatisierung, d.h. die Rückführung auf die im Lexikon verwendete Grundform (z.B. hat → haben). Das in der letzten Zeile sichtbare POS-Tagging (also die Klassifizierung nach Wortarten) wurde erst 2017 allgemein freigeschaltet, da vorher die Fehlerquo-

5 Selting et al. 2009: 359ff.

te mit 20% (Westpfahl & Schmidt 2013: 140, vgl. Blombach 2017) zu hoch lag. Nach einer Anpassung des Stuttgart-Tübingen-Tagsets an Daten aus mündlicher Kommunikation (STTS 2.0, vgl. Westpfahl et al. 2017) konnte die Fehlerquote beim POS-Tagging nach Angaben der Betreiber⁶ auf ca. 5% reduziert werden. Im gezeigten Ausschnitt scheinen die Probleme eher bei der Normalisierung (Ihrem/ihrer?) bzw. Lemmatisierung (hingekriegt/hinkriegen? Ihrem/Ihr?) zu liegen; aber falsch ist auch die (automatische) Klassifizierung von „Ihrem“ als NN (normales Nomen).

FOLK erlaubt verschiedene webbasierte Abfragen: Zunächst einmal kann man sich über die Funktion „Browsing“ einen Überblick über das Korpus, über sämtliche Sprechereignisse, Sprecher, Transkripte, Audio- und Videoaufnahmen sowie die Zusatzmaterialien verschaffen. In den Listen lassen sich alle Dokumente per Mausklick aufrufen. In den Transkripten kann man beliebig scrollen und zu jeder Textstelle per Doppelklick den entsprechenden Ausschnitt der Tonaufnahme anhören.

Möglich ist zweitens eine Volltextsuche in den Metadaten zu Ereignissen (z.B. Suche nach „Prüfung“ findet alle Prüfungsgespräche) und Sprechern (z.B. Suche nach „Berlin“ findet alle Sprecher, die mit Berlin zu tun haben) sowie in den GAT-Transkripten.

Die dritte Möglichkeit ist die sog. Token-Suche: Hier kommen die Annotationsebenen ins Spiel, denn es können sowohl transkribierte als auch normalisierte Wortformen, Lemmata und Wortarten-Tags einzeln oder in Kombination gesucht werden. Mit einer Kombination von „ein“ (transkribiert) und „einen“ (normalisiert) findet man etwa akkusativische Formen des maskulinen unbestimmten Artikels, die monosyllabisch realisiert sind. Mithilfe des Wortartkriteriums lässt sich z.B. „überhaupt“ als Abtönungspartikel (PTKMA, Bsp. wenn überhaupt) von der Gradpartikel (PTKIFG⁷, Bsp. überhaupt nicht/kein) unterscheiden⁸.

Bei der Token-Suche können zudem Kontextelemente (z.B. bei der Suche nach Kollokationen) und achtzehn verschiedene Positionsbeschränkungen (z.B. Stellung direkt nach einer Pause) definiert werden. Außerdem ist es möglich, Metadaten zu Sprechern (z.B. Bildungsgrad) und Sprechereignissen (z.B. Datum und Region) als Filter einzusetzen. Die Ergebnisse einer Token-Suche werden stets als KWIC-Tabelle (keyword in context) angezeigt, in der jede Zeile mit dem vollständigen Transkript und zugehöriger alignierter Aufnahme verlinkt ist. Bei

6 Vgl. E-Mail-Rundbrief an die DGD-Nutzer vom 9.4.2017.

7 PTKMA=Modal- und Abtönungspartikeln, PTKIFG=Intensitäts-, Fokus- und Gradpartikeln.

8 Die Trennschärfe ist aber in diesem konkreten Fall nicht besonders hoch; vor allem viele Gradpartikelvorkommen werden als Abtönungspartikeln klassifiziert.

Abb. 4 – Download-Formate und -Optionen in FOLK



hoher Trefferzahl⁹ sehr nützlich (im Sinne repräsentativer Forschungsergebnisse) ist die Möglichkeit, eine zufällige Stichprobe zu ziehen. Außerdem können virtuelle Korpora definiert werden, die man zu einem späteren Zeitpunkt wieder aufrufen kann. Sämtliche KWIC-Suchergebnisse sind online speicher- oder in verschiedenen Formaten downloadbar (s. Abb. 4). Einzelne Belege können ebenfalls online gespeichert oder heruntergeladen werden (max. 1 Minute Audio + Transkript). Die Daten können im Rahmen des Zitatrechts (mit Quellenangabe und Dokument-Kennung) veröffentlicht werden. Insgesamt 16 Sprechereignisse unterschiedlicher Interaktionstypen können im vollen Umfang (Audiofile, klickbare Transkripte usw.) heruntergeladen werden.

FOLK wendet sich primär an Forschende und Lehrende aus der Gesprächsforschung, der Korpuslinguistik und verwandten Ansätzen. Die Aufnahmen, vor allem Alltagsgespräche, sind aufgrund hoher Sprechgeschwindigkeit und niedrigen Aufnahmepegels für die Darbietung in DaF-Kursen oftmals nicht geeignet. Am ehesten sind Gespräche vor Publikum oder Verkaufsgespräche, also

⁹ Übrigens wird auch eine Funktion zur automatischen Quantifizierung des Suchergebnisses angeboten.

mit Gesprächsteilnehmern, die einander nicht kennen und wenig gemeinsames Wissen teilen, einsetzbar. Die Arbeit mit den Transkripten ist dagegen durchaus möglich, etwa um die Funktion von Gesprächspartikeln zu illustrieren. In jedem Fall kann das FOLK als Referenzkorpus für die Festlegung von Lerninhalten und ihrer Progression dienen und die Entwicklung authentischer Unterrichtsmaterialien unterstützen.

Die Entwicklung, Struktur und Funktionsweise von FOLK sind umfassend dokumentiert. Eine Literaturliste findet sich auf der Webseite des Archivs für Gesprochenes Deutsch des IDS¹⁰. Besonders hingewiesen sei auf den schon zitierten Aufsatz von Thomas Schmidt, dem Leiter des Programmbereichs „Mündliche Korpora“ der Abteilung Pragmatik, der 2014 in der Online-Zeitschrift „Gesprächsforschung“ erschienen ist. Der Aufsatz stellt FOLK und DGD als Instrumente gesprächsanalytischer Forschung dar und zeigt am Beispiel der Formel „ich sag mal“, wie eine einschlägige Studie mit Hilfe von FOLK durchgeführt werden kann.

2.2. *Gesprochenes Deutsch für die Auslandsgermanistik*

Das Projekt „Gesprochenes Deutsch für die Auslandsgermanistik“ wurde in den Jahren 2010-2012 unter der Leitung von Susanne Günthner und Wolfgang Imo an der Universität Münster durchgeführt und vom DAAD finanziert. Der Untertitel „Bereitstellung und Beitrag zur Didaktisierung von Materialien gesprochener Sprache in authentischen Kommunikationssituationen“ verdeutlicht, dass sich das Projekt dezidiert an die DaF-Praxis wendet und, anders als DGD/FOLK, kein Instrument für die Forschung sein will (die Nutzung für Forschungszwecke ist sogar explizit ausgeschlossen¹¹). Ausgangspunkt des Projekts ist die Beobachtung, dass DaF-Lehrwerke sich noch immer vorrangig an den Normen der Schriftsprache orientieren und in den Dialogen erheblich vom tatsächlichen Sprachgebrauch der Muttersprachler abweichen. Um dieses Defizit auszugleichen, bietet die Datenbank eine kleine Sammlung von authentischen Kommunikationsereignissen an, die im Unterricht eingesetzt werden können. „Methodologisch ist das Projekt der Interaktionalen Linguistik und der Gesprächsanalyse verpflichtet.“¹²

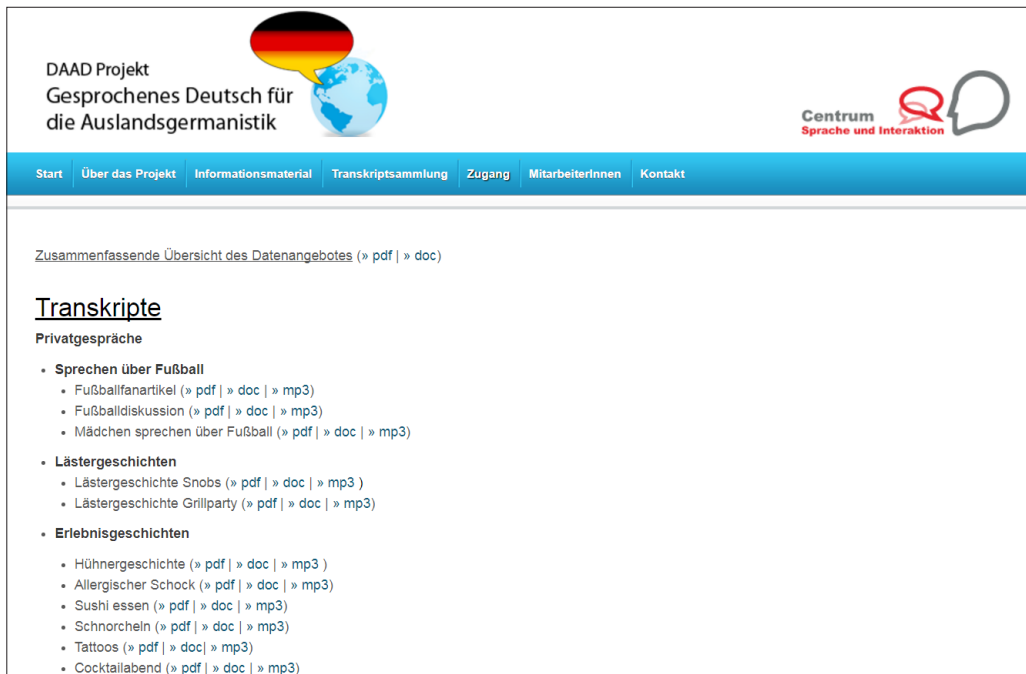
Die Datensammlung umfasst 61 Gespräche, von denen 52 als Audiodateien (mp3) angeboten werden, die übrigen (nur Gespräche an der Hochschule)

10 <http://agd.ids-mannheim.de/folk.shtml>

11 http://audiolabor.uni-muenster.de/daf/?page__id=26 (Der Hinweis erfolgt nur bei nicht registrierten Benutzern.)

12 http://audiolabor.uni-muenster.de/daf/?page__id=2

Abb. 5 – Übersicht über die Gespräche (Ausschnitt)



The screenshot shows the DAAD Projekt website interface. At the top left, the text reads 'DAAD Projekt' and 'Gesprochenes Deutsch für die Auslandsgermanistik' next to a logo of a speech bubble with a German flag and a globe. On the top right is the logo for 'Centrum Sprache und Interaktion'. Below the header is a blue navigation bar with links: 'Start', 'Über das Projekt', 'Informationsmaterial', 'Transkriptsammlung', 'Zugang', 'MitarbeiterInnen', and 'Kontakt'. The main content area has a link 'Zusammenfassende Übersicht des Datenangebotes (» pdf | » doc)'. Below this is the section 'Transkripte' with a sub-section 'Privatgespräche'. Under 'Privatgespräche', there are three main categories: 'Sprechen über Fußball', 'Lästergeschichten', and 'Erlebnisgeschichten'. Each category contains a list of specific topics with links to PDF and MP3 files.

Zusammenfassende Übersicht des Datenangebotes (» pdf | » doc)

Transkripte

Privatgespräche

- **Sprechen über Fußball**
 - Fußballfanartikel (» pdf | » doc | » mp3)
 - Fußballdiskussion (» pdf | » doc | » mp3)
 - Mädchen sprechen über Fußball (» pdf | » doc | » mp3)
- **Lästergeschichten**
 - Lästergeschichte Snobs (» pdf | » doc | » mp3)
 - Lästergeschichte Grillparty (» pdf | » doc | » mp3)
- **Erlebnisgeschichten**
 - Hühnergeschichte (» pdf | » doc | » mp3)
 - Allergischer Schock (» pdf | » doc | » mp3)
 - Sushi essen (» pdf | » doc | » mp3)
 - Schnorchelein (» pdf | » doc | » mp3)
 - Tattoos (» pdf | » doc | » mp3)
 - Cocktallabend (» pdf | » doc | » mp3)

als Videoaufnahmen, daneben stehen Transkripte (in den Formaten .doc und .pdf) sowie drei „Informationseinheiten“ zu den Themen „Sprecherwechsel“, „Parsequenzen“ und „Wenn-Sätze“ zur Verfügung.

Die Sammlung erhebt keinen Anspruch auf Repräsentativität, umfasst aber eine recht breit gefächerte Auswahl sowohl von privaten Alltagsgesprächen wie von institutionellen Gesprächen (s. Abb. 5). Die Privatgespräche gliedern sich in: Sprechen über Fußball, Lästergeschichten, Erlebnisgeschichten, Rezepte austauschen/Sprechen über Lebensmittel, Verabredungen am Telefon, Urlaubsplanung, Diskutieren, Erklären, Thema Studium. Die institutionellen und Dienstleistungsgespräche umfassen Gespräche an der Hochschule (u.a. Sprechstunde, Studienberatung, Kurzreferat, Seminardiskussion, Beratungsgespräch), Gespräche im Friseursalon, Verkaufs- und Beratungsgespräche (Bäckerei, Juwelier, Apotheke) sowie Arzt-Patienten-Gespräche. Bei den SprecherInnen handelt es sich überwiegend um junge Leute zwischen 20 und 30 Jahren; die regionale Herkunft der SprecherInnen wird nur selten vermerkt (z.B. Ruhrgebiet, Süddeutschland), dürfte aber vor allem in Norddeutschland liegen.

Das Korpus „Gesprochene Sprache für die Auslandsgermanistik“ umfasst neben den Ton- bzw. Videoaufnahmen zu jedem Gespräch ein Transkript, das

Abb. 6 – Transkript des Gesprächs „Klausurenstress“ (Ausschnitt)

001	C	aber bei mir ist es zum beispiel auch SO?
002		dass ich unter woche auch GAR nich so viel arbeiten könnte dass am wochenende-
003		dass das komplett FREI wäre glaub ich.
004		also ich bin ja bis sechs in der Uni oder so?
005		und dann komm ich WIEder?
006	F	hm-
007	C	dann kann ich noch n paar stunden was MACHen?
008		aber dann ist auch irgendwann GUT [also-]
009	N	[JA] ja.

aber nicht mit den Primärdaten aligniert ist. Die Metadaten zu Sprechern (Name, Alter, manchmal Herkunft), Situation/Thematik und Dauer des Gesprächs sind in sehr knapper Form der Transkription vorangestellt (z. B. zum Gespräch „Klausurenstress“: „Situation: Carina, Nele und Franziska, drei Studentinnen Anfang bzw. Mitte 20, sitzen gemeinsam in der Küche und essen Salat. Im Hintergrund läuft Radiomusik. Es geht um das Zeitmanagement beim Lernen für Klausuren bzw. Schreiben von Hausarbeiten.“). Weitere linguistische Annotationen (orthographische Normalisierung, Lemmatisierung, phonetische Transkription u.Ä.) stehen nicht zur Verfügung. Die Transkription erfolgt nach einem vereinfachten GAT2-System (vgl. Selting et al. 2009), das aber insofern über das sog. Minimaltranskript (in literarischer Umschrift mit Pausen, Überlappungen und para- bzw. außersprachlichen Handlungen) hinausgeht, als eine Gliederung in Intonationsphrasen erfolgt und Hauptakzente sowie finale Tonhöhenbewegungen verzeichnet werden (s. Abb. 6)¹³. Die Transkriptionskonventionen werden am Anfang jedes Transkripts kurz aufgeführt. Insgesamt erscheinen die Transkripte als dem intendierten Verwendungszweck der Materialien vollauf angemessen. Sie sind einerseits (prosodisch) informativer als z.B. die FOLK-Transkripte, andererseits aber dennoch gut lesbar und daher „optimal an die didaktischen Zwecke für den Einsatz an Schule und Hochschule angepasst“¹⁴.

Das Material der Datenbank „Gesprochenes Deutsch für die Auslandsgermanistik“ ist nicht öffentlich zugänglich, kann aber nach (kostenloser)

¹³ Vgl. http://audiolabor.uni-muenster.de/daf/?page_id=22 zu den Transkriptionskonventionen.

¹⁴ ebd., vgl. auch Spiegel 2009.

Abb. 7 – Begriffserklärungen im Transkript „Klausurenstress“ (Ausschnitt)

Begriffserklärungen:

- Z. 002: **unter der Woche:** an Werktagen (umgangssprachlich).
- Z. 013: **„sich vor den Fernseher hauen“:** sich vor den Fernseher setzen (umgangssprachlich).
- Z. 21: **Hausarbeit:** längere, schriftliche Arbeit vom Studierenden (akademisch).
- Z. 022: **„abschalten“:** sich entspannen (umgangssprachlich).

Registrierung als NutzerIn (mittels formloser Anfrage an Wolfgang Imo) genutzt werden. Sämtliche Dokumente (.pdf, .doc) und Aufnahmen lassen sich im Browser öffnen oder auf den eigenen Rechner herunterladen, es gibt aber keinerlei webbasierte Abfragewerkzeuge. Eine gewisse Orientierungshilfe bietet das Dokument „Zusammenfassende Übersicht des Datenangebotes“, in dem für jedes Gespräch Dateiname, Dauer, SprecherInnen und inhaltliche Schwerpunkte angegeben sind.

Die Datensammlung des DAAD-Projekt „Gesprochene Sprache für die Auslandsgermanistik“ wendet sich explizit an die DaF-Praxis im nichtdeutschsprachigen Ausland. Zwar handelt es sich lediglich um Materialien, noch nicht um fertige Unterrichtseinheiten, aber sie erscheinen als gut geeignet für den Unterricht mit fortgeschrittenen Lernenden. Dafür sprechen die akzeptable Tonqualität, die beschränkte Dauer der Gespräche (oftmals 1-3 Min., nur selten mehr als 5 Min.), die gut lesbaren Transkripte, die zusätzlichen „Begriffserklärungen“ (s. Abb. 7). Die bereits erwähnten „Informationseinheiten“ zu Sprecherwechsel, Paarsequenzen und Wenn-Sätzen hingegen bieten eine recht anspruchsvolle gesprächslinguistische Analyse und eignen sich in der Auslandsgermanistik wohl eher für DozentInnen oder fortgeschrittene Studierende. Es handelt sich um Textdokumente mit integrierten Audiobeispielen von 9-21 Seiten Länge.

Eine umfangreiche Literaturliste zur Vermittlung der gesprochenen Sprache im DaF-Unterricht findet sich auf der Internetseite des Projekts¹⁵. Eine knappe Darstellung der Datensammlung gibt Imo (2013: 149ff. mit weiterführenden Literaturhinweisen). Ein Vorschlag von Günthner (2015) zur Behandlung von Diskursmarkern im DaF-Unterricht stützt sich zum Teil auf das hier dargestellte Korpus. Gleiches gilt für einen Vorschlag von Moroni (2015) im selben Band zur Vermittlung der Intonation im DaF-Unterricht.

15 http://audiolabor.uni-muenster.de/daf/?page_id=5

2.3. Backbone

Auch Backbone¹⁶ ist ein webbasiertes Korpus mit dezidiert pädagogischer Ausrichtung; es wurde in den Jahren 2009-2011 vom Lifelong Learning Programme (Key Activity: Languages) gefördert und umfasst, neben dem Deutsch L1-Korpus, auf das sich die Darstellung hier beschränkt, Korpora zu Englisch, Polnisch, Spanisch, Türkisch und Englisch als Lingua franca. Koordiniert wurde das Projekt von der Abteilung Applied English Linguistics der Universität Tübingen (Prof. em. Kurt Kohn). Der Untertitel „Pedagogic Corpora for Content & Language Integrated Learning“ lässt an den bilingualen Sachfachunterricht (CLIL) denken, intendiert ist aber vor allem ein interkulturelles Lernen.

Das deutsche Korpus umfasst 25 Video-Interviews mit deutschen Muttersprachlern von je ca. 5-10 min. Länge, die auch als Audiodateien (.wav) zur Verfügung stehen. Es erhebt keinen expliziten Anspruch auf Repräsentativität, bietet aber einen plausiblen Querschnitt ungezwungener Interviews mit Fachleuten, die ihr Wissen und ihre Erfahrungen einem größeren Publikum mitteilen wollen.

Die Interviews wurden eigens für das Projekt angefertigt; es handelt sich also um elizitierte, aber spontane Gespräche, die in der Regel recht geordnet ablaufen, meist aus kurzen Fragen und längeren erzählenden oder darstellenden Antworten bestehen. Interessanterweise wurden verschiedene diatopische Varietäten bzw. Akzent-Färbungen berücksichtigt, nämlich Bairisch, Schwäbisch, Badisch, Norddeutsch, Rheinisch, Berlinisch, Alemannisch, Wienerisch. Bei der Hälfte der SprecherInnen ist die Varietät als „neutral“ eingestuft. Die SprecherInnen sind in der Mehrzahl männlich (ca. 2/3), üben die verschiedensten, meist gehobenen Berufe aus (Ladeninhaber, Berufsschullehrer, Biologin, Fahrzeugtechnikingenieur, Hochschullehrer, Bodenseefischer, Ärztin, freier Schriftsteller u.a.) und gehören offenbar überwiegend einer mittleren Altersgruppe an.

Das Korpus ist vollständig transkribiert, und zwar in orthographischer Umschrift (vgl. Abb. 9). Verzögerungs- und Rückmeldesignale (ähm, hm) erscheinen nicht. Elisionen werden nur beim unbestimmten Artikel durch literarische Umschrift ('n, 'ne) wiedergegeben, ebenso bei der 1. Person Singular der Verben (ich hab), aber in den meisten anderen Fällen nicht dokumentiert (gesprochenes „nich“ erscheint als „nicht“ usw.). Auch die Interpunktion folgt orthographischen Regeln und nicht der Intonation.

¹⁶ Zugang über <http://projects.ael.uni-tuebingen.de/backbone/moodle/> → Corpora & Search → Corpus search → (in der Titelleiste) Corpus → BB German

Abb. 8 – Backbone, deutsches Teilkorpus: Übersicht über die Interviews (Ausschnitt)





Info Korpus: BB German Menüsprache Hilfe		
Home Überblick Abschnittssuche Konkordanzsuche Konkordanzsuche Lexikalische Listen Ressourcen		
Preview	Interview	Optionen
	<p>Bayerische Spezialitäten in Berlin</p> <p>Boris ist Inhaber eines bayerischen Feinkostladens in Berlin. Er spricht über sein Sortiment, bayerische Spezialitäten und Biere. Boris spricht Hochdeutsch mit bayerischer Färbung.</p> <p>Akzent oder Varietät: Bayerisch</p>	<p>Video spielen</p> <p>Audio spielen</p> <p>Audiodatei herunterladen</p> <p>Interview zeigen</p> <p>Abschnittsübersicht zeigen</p>
	<p>Das deutsche Schulsystem</p> <p>Sylvia ist Lehrerin an einer kaufmännischen Berufsschule in Baden-Württemberg. Sie spricht über das deutsche Schulsystem, die Besonderheiten ihres Schultyps und ihre Arbeit als Lehrerin. Sylvia spricht mit schwäbischem Einschlag.</p> <p>Akzent oder Varietät: Schwäbisch</p>	<p>Video spielen</p> <p>Audio spielen</p> <p>Audiodatei herunterladen</p> <p>Interview zeigen</p> <p>Abschnittsübersicht zeigen</p>
	<p>Die Arbeit an einer Seehundstation in Schleswig-Holstein</p> <p>Eva Baumgärtner ist Biologin und arbeitet an der Seehundstation in Friedrichskoog an der Nordsee. Sie erzählt von der Informationsarbeit und Heuleraufzucht in der Station sowie über die Bedeutung der Seehundstation als touristische Attraktion in der Region. Eva kommt ursprünglich aus Baden-Württemberg und spricht Hochdeutsch.</p> <p>Akzent oder Varietät: Neutral</p>	<p>Video spielen</p> <p>Audio spielen</p> <p>Audiodatei herunterladen</p> <p>Interview zeigen</p> <p>Abschnittsübersicht zeigen</p>
	<p>Die Entwicklung von Kraftstoffanlagen</p> <p>Michael kommt von der Schwäbischen Alb und arbeitet im Raum Stuttgart in der Automobilindustrie. Er spricht über seinen Beruf, seine Ausbildung, und über die Schwaben. Er hat eine schwäbische Dialektfärbung.</p> <p>Akzent oder Varietät: Schwäbisch</p>	<p>Video spielen</p> <p>Audio spielen</p> <p>Audiodatei herunterladen</p> <p>Interview zeigen</p> <p>Abschnittsübersicht zeigen</p>

Abb. 9 – Backbone – Ausschnitt aus dem Transkript zu „Einblick in eine Arztpraxis“

I: Jetzt mal 'ne andere Sache, wir haben jetzt sozusagen das mehr aus der Patientenperspektive angeschaut, wie ist es jetzt sozusagen aus der Arztperspektive? Also, du hast ja diese Praxis noch nicht so lange. Wie sieht so 'ne Praxisgründung aus?

Ines: Ja, also ich habe diese Praxis jetzt zwei Jahre, nachdem meine Kinder groß sind, hab ich die Möglichkeit gehabt, die Praxis von einem Kollegen zu übernehmen. Das heißt hier in diesen Räumlichkeiten ist schon seit 30 Jahren eine Praxis und die hat in der Zeit zwei- oder dreimal den Arzt gewechselt, sozusagen. Die letzten fast 25, 30 Jahre war es jetzt mein Kollege, von dem ich das übernommen habe. Ja, da gibt es verschiedene Perspektiven, wir wohnen hier in einem Gebiet, wo man sich nicht einfach so niederlassen kann, sondern man gibt die Zulassung sozusagen weiter von Einem zum Andern. Und man muss diese Zulassung de facto abkaufen. De — Na, vor dem Gesetz eigentlich nicht. Das hat keinen rechtlichen Bestand. Und das ist eigentlich ein bisschen schwierig, wie in der Medizin gibt's da 'ne ganze Reihe von Grauzonen, das heißt, der Kollege hat mich angerufen, hat gesagt *ich weiß, du brauchst—du wolltest gerne meine Praxis übernehmen. Bist du da immer noch dran interessiert?* Und dann hab ich lange überlegt und bin halt zur Bank gegangen und hab überlegt, und dann hab ich ihn natürlich als Erstes mal gefragt *Was möchtest du dafür? Wie viel Scheine hast du?*, und so weiter. Und dann hab ich mir überlegt, ob ich das auch bezahlen kann und möchte, und ob sich das letztendlich auf die Dauer rechnet. Da ich jetzt schon bisschen älter bin, ist das natürlich 'ne Überlegung, ich werd so etwa noch zehn Jahre arbeiten, ob sich das in der Zeit amortisiert oder nicht. Aber im Prinzip ist es so, dass wir dann 'n Vertrag gemacht haben, wir haben noch 'n bisschen um den Preis gefeilscht und er hat mir die Praxis dann verkauft.

Die Transkription ist mit den Video- bzw. Audioaufnahmen nicht aligniert, aber die Transkripte sind in längere Abschnitte unterteilt, zu denen die passenden Audioclips abgerufen werden können. Außerdem sind die einzelnen Abschnitte im Hinblick auf Themen, Grammatik, dialektale Färbung, kommunikative Funktionen, Kompetenzstufen des Referenzrahmens sowie Dolmetschherausforderungen annotiert.

Bei der Abfrage (Menüpunkt: Abschnittssuche) können daher aus allen Interviews Abschnitte mit bestimmten Themen (Berufsleben/Arbeitswelt, Bildung, Gesundheitswesen, Kultur, Politik und Soziales, Städte und Regionen, Umwelt), bestimmten Grammatikphänomenen (z.B. Passiv oder Infinitivkonstruktionen), mit spezifischen kommunikativen Funktionen (sich vorstellen, eine Firma/Institution präsentieren, über Pläne sprechen, über die Vergangenheit sprechen, argumentieren, Meinungen äußern, Meinungen begründen), mit bestimmten Kompetenzstufen oder einer spezifischen dialektalen Färbung bzw. mit einer Kombination dieser Merkmale gefunden werden. Zusätzlich gibt es eine Konkordanz- und eine Kookkurenzsuche für das Gesamtkorpus. Sämtliche Transkripte, Audio- und Videodateien können heruntergeladen werden.

Das Korpus eignet sich zweifellos für den Einsatz im Fremdsprachenunterricht, wurde es doch mit eben diesem Ziel entwickelt. Als Lernstufe kommt etwa B2 (oder darüber) in Betracht, auch wenn einzelne Abschnitte durchaus auf der Grundstufe einsetzbar sind. Die kommunikative Gattung „Experteninterview“ eignet sich als klassische Hörverständnisübung, weil sie abfragbare Informationen zu landeskundlichen Themen bietet. Für eine Bewusstmachung der spezifischen Merkmale konzeptionell und medial gesprochener Sprache hingegen ist Backbone nicht das geeignete Material. Zu den thematischen Bereichen Berufsleben, Umwelt, Politik & Soziales, Städte & Regionen, Bildung, Tradition und Kultur, Gesundheitswesen steht (unter dem Menüpunkt „Ressourcen“) jeweils eine fertig ausgearbeitete Unterrichtsskizze bereit (als PDF-Dokument). Zudem werden (unter demselben Menüpunkt) Online-Lernmodule zu jedem Interview angeboten. Hier finden sich Hör-, Wortschatz- und Grammatikübungen, meist in Form von Lückentexten¹⁷. Eine Einführung in die Konzeption von Backbone bieten Kohn, Hoffstaeder & Widmann (2009).

2.4. Weitere Korpora kurzgefasst

Das kostenpflichtige Korpus „PhonDat2“ des Bayerischen Archivs für Sprachsignale (Institut für Phonetik, Universität München) umfasst orthographisch, segmental und prosodisch annotierte Aufnahmen vorgelesener Äußerungen

¹⁷ Die Videoclips können aufgrund veralteter Software aus den Lernmodulen heraus nicht mehr aufgerufen werden.

zum Thema Zugauskunft und ist gedacht zur Unterstützung der Entwicklung von Systemen der automatischen Spracherkennung. Es steht nur als CD-ROM zur Verfügung¹⁸.

Ebenfalls für phonetische Studien gedacht ist das kostenpflichtige „Kiel Korpus“ des (ehemaligen) „Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel“ mit vorgelesener und spontaner Sprache (vgl. Gut 2007b: 4).

Das „Korpus der gesprochenen Sprache des Ruhrgebiets“ (KGS¹⁹) (Leitung: Kerstin Kucharczik, Universität Bochum) umfasst ein „Altkorpus“ von ca. 120 Stunden Gesprächen mit Kleingärtnern aus den 1980er Jahren sowie ein „Neukorpus“ von ca. 15 Stunden, in dem seit 2013 Interviews mit Sprechern verschiedener Bevölkerungsgruppen gesammelt werden.

Im Bereich der medial schriftlichen, aber konzeptionell mündlichen Kommunikation sind das Dortmunder Chat-Korpus (TU Dortmund)²⁰ von Angelika Storrer und Michael Reißwenger sowie die Kurznachrichtendatenbank „Mobile Communication Database (MoCoDa)“²¹ von Wolfgang Imo erwähnenswert.

3. KORPORA FÜR GESPROCHENES DEUTSCH L2 (BZW. L1-L2)

3.1. *Gesprochene Sprache im akademischen Kontext (GeWiss)*

Das GeWiss-Korpus (Gesprochene Sprache im akademischen Kontext)²² ist ein kontrastives Korpus zur gesprochenen Wissenschaftssprache, das am Herder-Institut der Universität Leipzig unter Leitung von Christian Fandrych entstanden ist und in den Jahren 2009-2013 mit Mitteln des EU-Sozialfonds und der Volkswagenstiftung gefördert wurde²³. Die erheblich gestiegene Mobilität von Studierenden und Lehrenden macht vergleichende Untersuchungen der Wissenschaftssprache notwendig, für die das GeWiss-Korpus eine empirische

18 <https://www.phonetik.uni-muenchen.de/Bas/BasPD2deu.html>, vgl. Gut 2007b: 5.

19 <http://www.ruhr-uni-bochum.de/kgstr/>

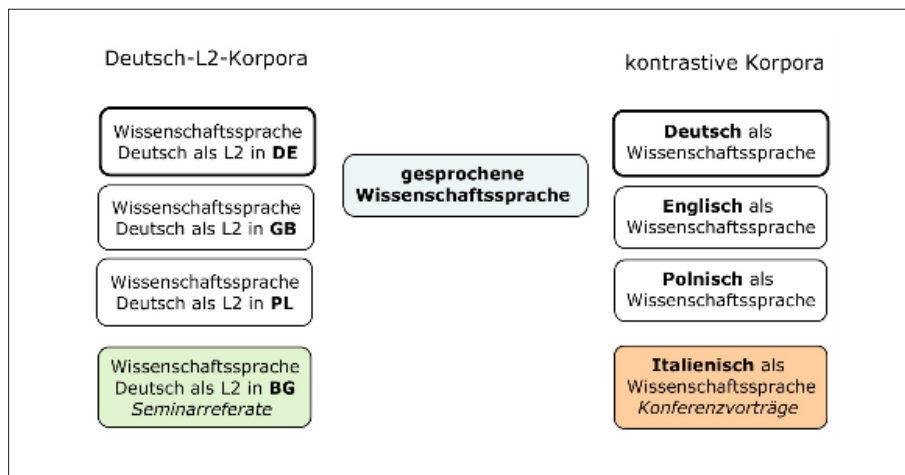
20 <http://www.chatkorpus.tu-dortmund.de/>, vgl. Imo & Moraldo 2015: 1.

21 <http://mocoda.spracheinteraktion.de/>, vgl. ebd.

22 <https://gewiss.uni-leipzig.de/index.php?id=home>; laut Mitteilung des DGD-Teams ist das GeWiss-Korpus ab Version 2.9 (27.11.2017) in DGD integriert. Das GeWiss-Portal bleibe aber bis auf weiteres aktiv. Dort seien die gleichen Daten (teilweise mit zusätzlichen Annotationen) und die nicht-deutschsprachigen Bestandteile des Korpus abrufbar.

23 Vgl. https://gewiss.uni-leipzig.de/index.php?id=about__gewiss zu den Folgeprojekten.

Abb. 10 – GeWiss-Teilkorpora



Grundlage bildet. Das Hauptaugenmerk liegt auf sprachlichen Routinen der mündlichen Wissenschaftskommunikation.

Das (Kern-)Korpus umfasst 371 Audio-Aufnahmen mit 462 Sprechern, einer Gesamtdauer von 126 Stunden und mehr als 1 Million Token. Es gliedert sich (vgl. Abb. 10) in verschiedene Deutsch-L2-Teilkorpora (erhoben in Deutschland, Großbritannien und Polen), denen entsprechende L1-Korpora (Deutsch, Englisch und Polnisch) gegenüberstehen. Hinzu kommen assoziierte Teilkorpora für Deutsch L2 in Bulgarien und Italienisch L1 in Italien.

Das Teilkorpus Deutsch L1 umfasst 30 Stunden, die Teilkorpora für Deutsch L2 insgesamt 58 Stunden. Nach eigenen Angaben wird das Korpus „fortlaufend ausgebaut“. Ein Teil (Prüfungsgespräche Deutsch L1) ist in FOLK (s. Abschnitt 2.1) integriert worden. Das GeWiss-Korpus ist nach kostenloser Registrierung frei zugänglich und kann für Forschung, Lehre und Studium genutzt werden.

Das Korpus umfasst drei Diskursarten: das Prüfungsgespräch, den Experten-vortrag und den studentischen Vortrag. Die Vorträge sind zum Teil vollständig oder teilweise abgelesen oder auswendig gelernt; es handelt sich also nicht in jedem Fall um spontane mündliche Sprache. Die Aufnahmen wurden im Fach Deutsch als Fremdsprache bzw. Germanistik sowie in den Philologien der Partnerländer (Anglistik und Polonistik) vorgenommen. Man darf davon ausgehen, dass für die erhobenen Diskursarten sowie die genannten Ausgangs- und Zielsprachen-Kombinationen Repräsentativität gegeben ist.

Das Korpus ist vollständig nach den Regeln des GAT2-Minimaltranskripts transkribiert (also mit Tilgungen und Klitisierungen, Verzögerungs- und Rezeptionssignalen, Pausen, Atmen, nonverbalen Ereignissen, aber ohne proso-

dische Merkmale, vgl. Selting et al. 2009). Orthographische Normalisierung und Lemmatisierung sind nicht verfügbar. Es gibt jedoch verschiedene Arten von Annotationen: Annotiert wurden zunächst einmal Sprachwechselphänomene, die vor allem in den Deutsch-L2-Daten zu beobachten sind (z.B. als Kommunikationsstrategie oder beim Wechsel auf eine Meta-Ebene). In einer eigenen Kommentarspur werden diese Äußerungsteile sinngemäß übersetzt. Weiterhin wurden (in den Deutsch-L1-Expertenvorträgen) sog. Diskurskommentierungen²⁴ annotiert, also z.B. Passagen, in denen der/die Vortragende auf die Makrostruktur des eigenen Vortrags Bezug nimmt, explizit eigene Sprachhandlungen verbalisiert (performative Elemente) oder sich auf den situativen Rahmen (z.B. die Vorstellung durch den Moderator, die spätere Diskussionsphase usw.) bezieht. Eine dritte Annotationsebene (nur Experten- und studentische Vorträge in Deutsch L1) betrifft Verweise auf bzw. Zitate von wissenschaftlichen Publikationen, Studien und Konzepten²⁵. Das Korpus umfasst auch Metadaten (vgl. Abb. 11) zu den kommunikativen Ereignissen (SprecherInnen und ihre Beziehung zueinander, eingesetzte Medien, Grad der Mündlichkeit usw.) und zu den jeweiligen SprecherInnen (u.a. Informationen zum Bildungsweg und zu sprachlichen Kompetenzen).

Das GeWiss-Korpus bietet eine Vielzahl an webbasierten Abfragemöglichkeiten. Bei der Volltext-Suche kann man, ausgehend von einer Übersicht der einzelnen Teilkorpora (d.h. einer spezifischen Kombination von Zielsprache, sprachlichem Kontext und Diskursart) und der jeweils zugehörigen Kommunikationsereignisse und Sprecher die einzelnen, mit den Audiodateien alignierten Partitur-Transkripte aufrufen und durch Mausklick gezielt bestimmte Segmente anhören (vgl. Abb. 12).

Die Aufnahmen können in voller Länge als MP3-Dateien, die Transkripte als PDF-Dokumente heruntergeladen werden. Die Konkordanzsuche (Abb. 13) gestattet außer einer Suche in der Verbalspur der Transkripte (auch mit regulären Ausdrücken, d.h. Platzhaltern) die Suche auf den oben genannten Annotationsebenen (Sprachwechsel, Diskurskommentierungen, Verweise und Zitate).

Die Suche kann auf mehr oder minder große Teilkorpora eingeschränkt werden. Bei der erweiterten Suche lassen sich zudem die Metadaten der Kommunikationsereignisse und SprecherInnen als Filter verwenden. Bei Suchen in der Verbalspur mag die literarische Umschrift, d.h. die fehlende orthographische Normalisierung, Probleme bereiten; hilfreich ist hier die Liste der verwendeten Reduzierungen und Klitisierungen im Anhang des GeWiss-

24 Vgl. die Dokumentation

https://gewiss.uni-leipzig.de/fileadmin/documents/Annotationsdokumentation__GeWiss.pdf

25 Vgl. die Dokumentation

<https://gewiss.uni-leipzig.de/fileadmin/documents/VZDokumentation.pdf>

Abb. 13 – GeWiss-Konkordanzsuche

Konkordanzen

Weltere Hinweise zu den Diskurskommentierungen und Verweisen/Zitaten

DEU_L2 überhaupt Verbalspur

Suche - Kontext (10) + Kontext (10)
Erweiterte Suche Export

1 - 50 von 220 Gesamttreffern

Kommunikat	Sprecher	Linker Kontext	Treffer	Rechter Kontext
SV_BG_032	IK_0057	(0.7) erpressungsbriefen ^{0h} und (.) n atürlich (.) ahm ü	überhaupt	(mit) ah anonyme (.) ah briefe (0.4) ah das
SV_BG_032	NG_0056	(.) sowie (0.4) strafvollzüge und äh (0.3) (verhörungen) (xxx)	überhaupt	(1.3) ah (.) ((räuspert sich)) die forensik (1.3) ((schmatzt))
SV_BG_033	IT_0054	vorschlagen wenn am ende zeit bleibt [[stotternd] wenn] es	überhaupt	nötig ist (0.3) klären wir noch die anderen begriffe
SV_BG_033	II_0055	und (.) der ist so verschieden dass () äh	überhaupt	nicht in diesen merkmalen passt ^{0h} e lgentlich passt aber
SV_BG_034	DG_0058	(2.8) hm (1.1) gibt es stockende spracheproduktion / y oder	überhaupt	keine verbale äußierungsfähigkeit (0.8) äh mangel an sprachliche äußierung

Handbuchs (Gräfe et al. 2015). Schließlich bietet GeWiss mit den „Webservices“ eine Abfragemöglichkeit mittels standardisierter Eingaben in der Adresszeile des Webbrowsers, was eine automatisierte Abfrage und Weiterverarbeitung in anderen Anwendungen gestattet.

GeWiss, insbesondere das Teilkorpus der deutschen L1-Sprecher, kann im Unterricht zur Vermittlung der gesprochenen Wissenschaftssprache Deutsch in Expertenvortrag, studentischem Vortrag und Prüfungsgespräch eingesetzt werden, z.B. in DaF-Kursen für ausländische Studierende an universitären Sprachenzentren in Deutschland oder in der Auslandsgermanistik. Besonders geeignet sind die Vorträge, weil in ihnen wissenschaftssprachliche Routinen annotiert sind. Außerdem ist das Korpus natürlich für den Zweck einsetzbar, für den es speziell entwickelt wurde, also für vergleichende Untersuchungen zur Wissenschaftssprache und zum Erwerb der deutschen Wissenschaftssprache durch nicht muttersprachliche SprecherInnen.

Eine gute Einführung in die Konzeption des GeWiss-Korpus und seinen Einsatz in empirischen Studien zur Wissenschaftssprachkomparatistik bietet der von Fandrych, Meißner & Slavcheva (2014) herausgegebene Sammelband „Gesprochene Wissenschaftssprache. Korpusmethodische Fragen und empirische Analysen“.²⁶ Auf der Website des Korpus²⁷ findet sich zudem eine umfangreiche Liste von Fachartikeln sowohl über das Korpus selbst und methodische Aspekte als auch zu einzelnen Phänomenen der mündlichen Wissenschaftssprache.

26 Ein neuerer Sammelband (Fandrych, Meißner, Sadowski & Wallner 2017) befasst sich hingegen mit der Weiterentwicklung von Annotationsverfahren (z.B. Wortarten- und pragmatische Annotation) und digitaler Auswertung.

27 <https://gewiss.uni-leipzig.de/index.php?id=papers>

3.2. Berlin Map Task Corpus (BeMaTaC)

Das Berlin Map Task Corpus (BeMaTaC²⁸), ein tief annotiertes multimodales Map-Task-Korpus gesprochener Lerner- und Muttersprache, ist unter der Leitung von Anke Lüdeling und Simon Sauer an der Humboldt-Universität Berlin entstanden; die aktuelle Version ist von 2013. Im Bereich Korpuslinguistik des Instituts für deutsche Sprache und Linguistik der HU ist eine Vielzahl von weiteren Korpora erstellt worden, darunter das Fehlerannotierte Lernerkorpus des Deutschen als Fremdsprache (Falko²⁹) mit schriftlichen Produktionen von DaF-Lernenden. BeMaTaC umfasst in der Version von 2013 ein L1-Subkorpus mit 12 Dialogen (66 Minuten, 8900 normalisierte Token) sowie ein L2-Subkorpus mit 5 Dialogen (77 Minuten, 9228 normalisierte Token). Es war zunächst als L1-Ergänzung seines Vorläufers HaMaTaC³⁰ gedacht, wurde dann aber doch auf L2 ausgedehnt. Die Map Tasks wurden vom Projekt „Variation des gesprochenen Deutsch“ des IDS Mannheim übernommen. BeMaTaC ist webbasiert und ohne Anmeldung frei zugänglich.

Die Diskursgattung des BeMaTaC ist sehr speziell: „BeMaTaC verwendet ein Map-Task-Design, hierbei instruiert ein/e Sprecher/in (sog. Instructor) eine/n andere/n Sprecher/in (sog. Instructee), eine Route auf einer Karte mit Landmarken zu reproduzieren. Die SprecherInnen können sich nicht gegenseitig sehen und können daher nicht non-verbal kommunizieren. Die Dialoge werden mit zwei separat platzierten Mikrofonen aufgezeichnet, zusätzlich wird ein Video aufgezeichnet³¹, welches die zeichnende Hand des Instructees zeigt.“³² Der Vorteil dieses experimentellen Designs ist, dass „[...] eine spontansprachliche Gesprächssituation erzeugt [wird], welche jedoch thematisch klar abgegrenzt ist, was Vergleichbarkeit und Generalisierungen erleichtert.“ (Sauer et al. 2013: 82). Allerdings ist eine solche Aufgabe in der realen sprachlichen Kommunikation kaum anzutreffen, zumal die verwendete Karte nicht mit normalen Stadtplänen oder Straßenkarten zu vergleichen ist. Das Korpus ist daher nicht als repräsentativ für eine bestimmte Diskursgattung anzusehen. Repräsentativ

28 <http://u.hu-berlin.de/bematac>

29 <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/standardseite>

30 Das Hamburg MapTask Corpus (HaMaTaC, <https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus%3Ahamatac>), entstanden 2009/10 an der Universität Hamburg, ist ein Korpus mit Aufnahmen von 24 fortgeschrittenen Deutsch-L2-LernerInnen, die nach HIAT transkribiert sind. Es gibt kein Web-Interface, aber man kann alle Audio-Dateien und die Transkripte in verschiedenen Formaten herunterladen.

31 Videoaufnahmen sind nur für das L1-Teilkorpus verfügbar.

32 Vgl. <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/bematac>

sind aber möglicherweise Phänomene spontansprachlicher Kommunikation, vor allem im L1-Teilkorpus (das L2-Teilkorpus umfasst nur SprecherInnen mit Englisch als L1).

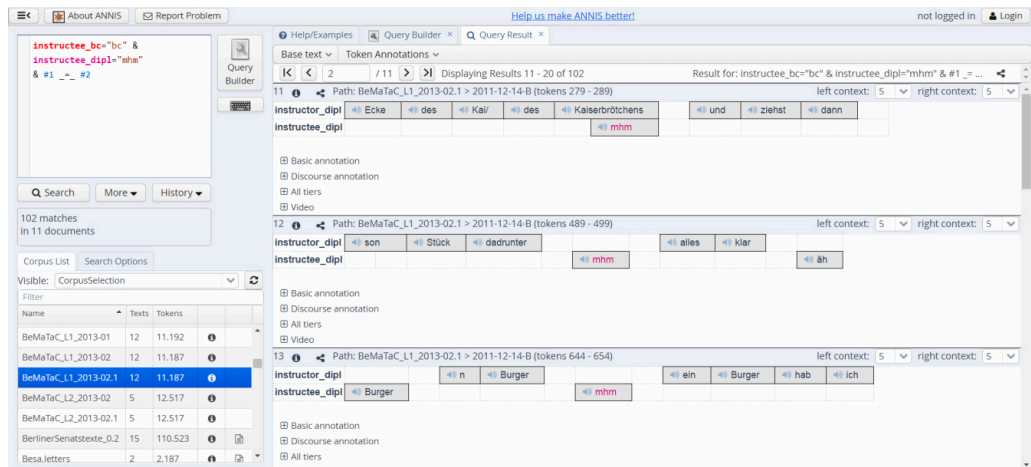
BeMaTaC zeichnet sich durch eine umfangreiche Transkription und Annotation in separaten Layern aus. Die erste Ebene enthält die literarische Umschrift, in der gesprochensprachliche Phänomene (Verzögerungssignale, Wortabbrüche, Elisionen, Verschleifungen) und idiosynkratische Aussprache berücksichtigt werden. Die zweite Ebene bildet die orthographische Normalisierung. Außerdem gehören zur „Basisannotation“ Layer mit automatischer Lemmatisierung und Wortarten-Bestimmung (nach dem Stuttgart-Tübingen-Tagset), die Gliederung in (syntaktisch motivierte) Äußerungen („utt“), nicht linguistische hörbare Ereignisse (wie z.B. Luftholen) sowie ungefüllte Pausen. Weitere Ebenen sind diskurslinguistisch motiviert und betreffen Backchanneling (Hörersignale), Verzögerungssignale („disfluencies“) sowie Reparaturen.³³ Eine phonetische-phonologische Transkription bzw. Annotation ist/war längerfristig geplant, aber liegt bisher nicht vor.

Das Korpus bzw. die einzelnen Kommunikationsereignisse können webbasiert visualisiert und abgefragt werden (s. Abb. 14). Die Suche kann einen oder mehrere Layer gleichzeitig betreffen. Z.B. kann man auf der Backchanneling-Ebene alle Vorkommnisse (verschriftet als „bc“) oder zusätzlich auf der Ebene der literarischen Transkription bestimmte Hörersignale (z.B. „okay“, „mhm“, „ja“) suchen. Bei der Erstellung der Abfrage helfen ein „Query Builder“ und eine Hilfsfunktion. In jedem Suchergebnis können alle Transkriptions- und Annotationsebenen aufgedeckt werden; durch Klicken auf ein Token öffnet sich ein Fenster, in dem ein kurzer Audio- oder Videomitschnitt abgespielt wird. In diesem Mitschnitt kann man beliebig vor- oder zurückspulen und, falls gewünscht, auch die gesamte Aufnahme anhören, wobei die Transkription aber nicht synchron visualisiert wird. Die Audio- (mp3/wave), Video- (QuickTime, WebM) und Annotationsdateien (EXMARaLDA-Partituren) lassen sich auch herunterladen.

Wie schon erwähnt, ist die Diskursgattung Map-Task-Dialog kommunikativ kaum von Interesse für den DaF-Unterricht. Andererseits können die diskurslinguistischen Phänomene, die in BeMaTaC annotiert wurden und die für verschiedenste Formen spontaner gesprochensprachlicher Kommunikation kennzeichnend sind, durchaus auch für DaF relevant sein. Bei den Hörersignalen ist z.B. auffällig, dass L1- und L2-SprecherInnen ganz unterschiedliche Signale benutzen: Unter 290 L2-Signalen gibt es nur 6 Vorkommen von „mhm“ (es dominiert „okay“), während bei den L1-Sprechern jedes dritte von 307

³³ Für weitere detaillierte Informationen vgl. <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/bematac/documentation/discourse>

Abb. 14 – Das Recherche-Interface ANNIS des Korpus BeMaTaC



Hörersignalen als „mhm“ realisiert wird. Die Audio- bzw. Videodateien ermöglichen es den Lernenden, viele dieser authentischen Hörersignale in adäquaten Kontexten und mit der korrekten muttersprachlichen Intonation zu hören. Da die Map-Task-Dialoge sprachlich einfach und inhaltlich vorhersehbar sind, lassen sie sich auch im Grundstufenbereich einsetzen.

Es gibt keine zusammenfassende Darstellung von BeMaTaC, aber einschlägige Poster, die auf der Homepage des Projekts³⁴ verlinkt sind. Sauer & Lüdeling (2016) diskutieren allgemein Anforderungen an flexibel einsetzbare bzw. erweiterbare Korpora gesprochener Sprache, die z.T. anhand von BeMaTaC veranschaulicht werden. Außerdem findet sich auf der Homepage eine Liste der anhand von BeMaTaC durchgeführten Untersuchungen, insbesondere zu Reparaturen, Verzögerungsphänomenen und Hörersignalen (Poster, Abstracts, Präsentationen, Abschlussarbeiten, vgl. auch Belz u.a (2017)).

3.3. *LeaP - Learning the Prosody of a foreign language*

Das LeaP-Korpus ist, dank einer Finanzierung durch das nordrhein-westfälische Wissenschaftsministerium in den Jahren 2001-2003, unter der Leitung von Ulrike Gut an der Universität Bielefeld entstanden. Eine kurze Projektbeschreibung findet sich auf der Website der Universität Augsburg³⁵, wo Ulrike

34 <http://u.hu-berlin.de/bematac>, u.a. Giesel et al. 2013, Sauer & Rasskazova 2014.

35 https://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-linguistics/zum_loeschen/Forschung/leap/index.html

Gut mittlerweile lehrt. Das Korpus wurde mit dem Ziel zusammengestellt, die Prosodie im Fremdsprachenerwerb (Deutsch und Englisch) zu untersuchen und unterscheidet sich damit grundsätzlich von den anderen hier vorgestellten Korpora. Es gibt keine webbasierte Abfrage, auch das ursprüngliche Abfrage-Tool³⁶ ist obsolet, die Audio- und Annotationsdateien sind stehen aber noch zum Herunterladen³⁷ zur Verfügung. Das L2-Korpus besteht aus 359 annotierten Aufnahmen von 131 Sprechern mit 32 verschiedenen Ausgangssprachen sowie aus 18 Aufnahmen von Muttersprachlern, mit einer Gesamtlauzeit von mehr als 12 Stunden. Die Zielsprachen Deutsch und Englisch sind gleichgewichtig vertreten (vgl. *LeaP Corpus Manual*, S. 3). Das LeaP-Korpus beansprucht Repräsentativität im Hinblick auf die Prosodie der Lernaltersprache (vgl. Gut 2009).

Aus der prosodischen Zielsetzung erklären sich auch die vertretenen Sprechereignisse: Vorlesen einer Geschichte, Nacherzählung derselben Geschichte, freies Interview und Vorlesen einer Nonsense-Wort-Liste. Das Interesse gilt nicht bestimmten Diskursgattungen, sondern unterschiedlichen Graden von Spontaneität (vorgelesen, vorbereitet, frei gesprochen). Einige L2-SprecherInnen sind weit fortgeschrittene LernerInnen, die von Muttersprachlern fast nicht zu unterscheiden sind („superlearner“), denn es sollte geprüft werden, ob ein perfekter Erwerb der Zielsprachlichen Prosodie möglich ist. Außerdem wurden ProbandInnen vor und nach einem Prosodiekurs getestet, um den Einfluss von gesteuertem Lernen und von bestimmten Unterrichtsmethoden zu untersuchen. Die Rolle des ungesteuerten Erwerbs wurde durch die Erhebung von Daten vor und nach einem Auslandsaufenthalt berücksichtigt. Weitere Versuchspersonen wurden einbezogen, um eine ausgeglichene Verteilung der Ausgangssprachen zu erzielen (vgl. *LeaP Corpus Manual*, S. 3). Als Kontrollgruppe dienten einige deutsche und englische MuttersprachlerInnen.

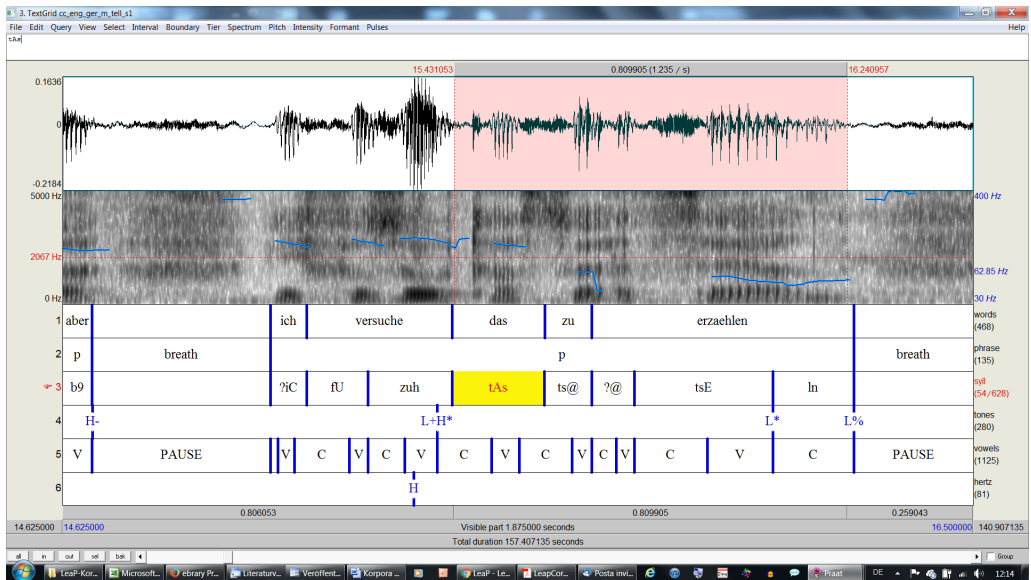
Das LeaP-Korpus ist (mit Ausnahme der vorgelesenen Wortlisten) auf mehreren Ebenen transkribiert und annotiert, die mit Hilfe von Praat (Boersma & Weenink 2001) auf acht ‚tiers‘ visualisiert werden können (vgl. Abb. 15): orthographische Transkription (words tier, etikettiert als ‚words‘); segmentale phonetische Transkription mit SAMPA (syllable tier, etikettiert als ‚syll‘); Vokale und Konsonanten (segments tier, etikettiert als ‚vowels‘); Intonationsphrasen, Pausen, nicht-sprachliche Ereignisse u.a. (phrase tier, etikettiert als ‚phrase‘); TOBI-Töne (tone tier, etikettiert als ‚tones‘); Melodieverlauf (pitch tier, etikettiert als ‚hertz‘); automatische generierte Wortartenklassifizierung (POS tier); automatisch erzeugte Lemmata (lemma tier). Allerdings sind nicht alle Aufnahmen auf allen acht Ebenen annotiert; insbesondere fehlt bei den meisten

36 TASX-Browser von Jan-Torsten Milde, vgl.

<http://www.annotation.exmaralda.org/index.php/TASX>

37 https://sourceforge.net/projects/leapcorpus/?source=typ__redirect

Abb. 15 – LeaP-Annotationsebenen in Praat (Proband „cc“ mit L1=Eng, L2=Deu, Nacherzählung, „superlearner“, hier mit Annotationen auf den Ebenen ‚tones‘ und ‚hertz‘)



L2-Aufnahmen die intonatorische Annotation (tone tier, pitch tier), außerdem wurden im deutschen Teilkorpus Wortarten und Lemmata nicht annotiert³⁸. Metadaten wurden erfasst zu den Aufnahmen (Datum, Ort, Interviewer, Interview-Sprache) und zu den L2-SprecherInnen (Alter, Geschlecht, Sprachkenntnisse, Sprachkontakt und -unterricht; Motivation und Einstellungen zu Aussprache, Musik und Theater).

Wie schon erwähnt, sind sämtliche LeaP-Aufnahmen und -Annotationen zwar herunterladbar, aber es gibt für das Korpus keine Abfragemöglichkeiten (mehr). Es lässt sich daher (in der vorliegenden Form) nur schwer für eigene prosodische Korpusstudien nutzen, könnte aber als Vorbild oder Anregung für andere Korpora zum L2-Ausspracheerwerb dienen. Zu den prinzipiellen Möglichkeiten der Nutzung von lernersprachlichen Korpora in Forschung und Lehre zur L2-Prosodie vgl. Gut (2007a).

Eine Übersicht über das Korpus bieten das „LeaP Corpus Manual“, Gut (2007a, 151-153) und Gut (2009: 63-75). Eine ausführliche Darstellung der mit Hilfe des LeaP-Korpus erzielten Forschungsergebnisse zur lernersprachlichen Prosodie des Deutschen und Englischen bietet Gut (2009). Je eigene Kapitel dieser

³⁸ Weitere Informationen zur Annotation im LeaP Corpus Manual, S. 5f. Eine Übersicht über die verfügbaren Annotationen findet sich im Anhang, S. 12ff.

Monographie sind den Themen Flüssigkeit, Silbengliederung und Konsonan-
tenclusterreduktion, Sprechrhythmus und Vokalreduktion, Intonation sowie
fremdem Akzent gewidmet.³⁹

3.4. Weitere Korpora kurzgefasst

In der einschlägigen Literatur werden weitere Korpora für Deutsch als
gesprochene Zweit-/Fremdsprache erwähnt, die nicht im Internet zugänglich
sind. Dazu zählt das an der Universität Barcelona entstandene Korpus Varkom
(vgl. Fernández-Villanueva & Strunk 2009). Es umfasst neben mündlichen
Produktionen in den Zielsprachen Katalanisch und Spanisch Aufnahmen von
18 SprecherInnen in der Mutter-, Fremd- und Zweitsprache Deutsch. Es sind
verschiedene Diskurstypen bzw. thematische Entfaltungen (Interview zur
Person, Erzählung, Beschreibung, Argumentation, Erörterung) berücksichtigt
worden. „Als Produkt liegt eine CD in Betaversion vor mit Transkriptionen
von den Videoaufnahmen mit den Informanten; die Texte können gelesen und
durchsucht werden, Videoaufnahmen sind ebenfalls integriert.“⁴⁰

Costa (2008) berichtet von einem Projekt zum Aufbau der Datensammlung
„Deutsch als Fremdsprache in berufsbezogenen Kontexten“, das „eine Daten-
grundlage für die Beschreibung der Interaktion zwischen Muttersprachlern
und Nichtmuttersprachlern in berufsbezogenen Situationen und unter Berück-
sichtigung des Sprachenpaars Italienisch–Deutsch bieten soll“ (ebd., 134). Bisher
wurde vor allem die Gattung der Stadtführungen dokumentiert.

4. ZUSAMMENFASSUNG

Unter den vorgestellten L1-Korpora kommt dem FOLK-Korpus eine besondere
Stellung zu. Dank eines repräsentativen Querschnitts von privaten, institutionel-
len und öffentlichen Gesprächen, eines Umfangs von über 200 Stunden und seiner
institutionellen Verankerung am IDS Mannheim sowie der damit verbundenen
Bestandsgarantie und Erweiterungsmöglichkeit kann es als Referenzkorpus für
das gesprochene Deutsch der Gegenwart gelten. FOLK verfügt über ein hoch
entwickeltes Online-Abfragesystem, mit dem sich komplexe Fragestellungen,
einschließlich quantitativer Aspekte bearbeiten lassen. Das Korpus eignet sich für

³⁹ Zu weiterer Literatur vgl. https://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-linguistics/zum_loeschen/Forschung/leap/index.html

⁴⁰ Vgl. http://www.ub.edu/lada/?page_id=14. Information bei Prof. Dr. Fernández (fernandezvillanueva@ub.edu).

Forschung und Lehre in Gesprächsforschung und Korpuslinguistik, aber auch zur Festlegung von Lerninhalten für die mündliche Interaktion in der Fremdsprache Deutsch. Ein direkter Einsatz im DaF-Unterricht ist nicht intendiert, aber sofern für Korpusrecherchen der Lernenden selbst genug Zeit zur Verfügung steht, kann FOLK auch zum induktiven Lernen und zur Sprachbewusstheit beitragen (vgl. Gut 2007a: 146, Ylönen 2012).

Speziell für den Einsatz in Deutsch als Fremdsprache wurde hingegen das Korpus „Gesprochene Sprache für die Auslandsgermanistik“ erstellt, das mit einer Gesamtlaufzeit von ca. 2 Stunden einen vergleichsweise geringen Umfang hat, aber dafür mit einer Auswahl von kurzen privaten und institutionellen Gesprächen direkt im Unterricht einsetzbar ist. Die Transkripte sind gut lesbar und bieten wichtige prosodische Informationen (Satzakzente, Grenztöne). Die Materialien können nicht online abgefragt werden und sind nicht aligniert, lassen sich aber in vollem Umfang herunterladen und somit im Unterricht nutzen, um immer noch bestehende Defizite von DaF-Lehrmaterialien auszugleichen.

Für den Einsatz im Fremdsprachenunterricht konzipiert ist auch das Backbone-Korpus, das durch seine Interviews mit „Experten“ aus verschiedenen gesellschaftlichen Bereichen aber weniger auf bestimmte Gesprächstypen oder spezifische Merkmale gesprochener Sprache abhebt, als auf interkulturelle Information und auf die Schulung des Hör-Seh-Verständnisses (auch mit unterschiedlichen diatopischen Varietäten). Wortschatz- und Grammatikübungen gehören ebenso zu diesem Ansatz. Ein differenziertes Abfragesystem ermöglicht die Suche sowohl nach inhaltlichen wie nach lexikalischen und grammatischen Aspekten. Mit einem Umfang von ca. 3 Stunden für die Zielsprache Deutsch bietet Backbone eine Fülle an didaktisiertem Material etwa ab der Kompetenzstufe B2.

Alle hier vorgestellten Deutsch-L2-Korpora enthalten zwecks Vergleichsmöglichkeiten jeweils auch ein L1-Teilkorpus. Von Umfang (58 Stunden L2, 30 Stunden L1) und Abfragemöglichkeiten her dominiert das kontrastive angelegte GeWiss-Korpus zur gesprochenen Sprache im akademischen Kontext. Der speziellen Zielsetzung entsprechen die Beschränkung auf Prüfungsgespräche, Experten- und studentische Vorträge und die Annotation von Sprachwechselphänomenen, Diskurskommentierungen und Verweisen/Zitaten. Das Korpus eignet sich für vergleichende Untersuchungen zur Wissenschaftssprache und zum Erwerb der deutschen Wissenschaftssprache durch nicht muttersprachliche SprecherInnen, kann aber auch im Unterricht zur Vermittlung der gesprochenen Wissenschaftssprache Deutsch eingesetzt werden, z.B. in DaF-Kursen für ausländische Studierende an universitären Sprachenzentren in Deutschland oder in der Auslandsgermanistik.

Das BeMaTaC-Korpus (Berlin Map Task Corpus) ist ein von Umfang (ca. je 1 Stunde L1 und L2) und Diskurstyp (Map Task) sehr beschränktes und spezifisches

Korpus, das aber mit einer umfangreichen Annotation und einer sehr guten Online-Abfrage aufwartet. Interessant sind die für spontane, interaktionale Sprache insgesamt charakteristischen und in einem komplexen Annotations-system registrierten Phänomene des Backchanneling (Hörersignale), der Verzögerungssignale („disfluencies“) sowie der Reparaturen. Diese können anhand der authentischen Beispiele von BeMaTaC auch im DaF-Unterricht veranschaulicht werden.

Das LeaP-Korpus (Learning the Prosody of a foreign language) ist das einzige der hier vorgestellten Korpora, das explizit zur Untersuchung von phonetisch-prosodischen Merkmalen von Lerner Sprache konzipiert und realisiert wurde. Bei der Datenerhebung ging es demzufolge weniger um bestimmte kommunikative Gattungen als darum, gesprochene Sprache mit unterschiedlichen Graden von Spontaneität zu erfassen (vom vorgelesenen Text bis zum freien Interview). Das Korpus ist auf mehreren Layern durchgehend segmental und prosodisch annotiert (die L2-Produktionen sind jedoch nur zu einem geringen Teil intonatorisch annotiert). Vom Umfang her (12 Stunden für alle Teilkorpora zusammen: Deutsch/ Englisch, L1/L2) nimmt es eine mittlere Stellung ein. Zwar gibt es für LeaP keine funktionierenden Abfragetools mehr, aber die aus dem Web herunterladbaren Audio- und Praat-Dateien können visualisiert und als Anregung für andere phonetisch und prosodisch orientierte Lerner Sprachkorpora genutzt werden.

Bei den L1-Korpora stehen mit FOLK und dem Korpus „Gesprochene Sprache für die Auslandsgermanistik“ zwei Datensammlungen zur Verfügung, die eine ausreichende Grundlage für Forschung und DaF-Unterricht bieten. Zu wünschen bliebe allenfalls, dass die gesprächsanalytische Orientierung durch phonetisch-prosodisch annotierte Korpora ergänzt würde. Hinsichtlich der L2-Korpora ist anzumerken, dass bei den webbasiert abfragbaren Korpora (GeWiss und BeMaTaC) das Angebot an Kommunikationsgattungen (Prüfungsgespräch, studentischer und Expertenvortrag, Map-Task-Dialog) und Ausgangssprachen (Englisch, Polnisch, z.T. Bulgarisch und Italienisch) recht beschränkt ist. Hier wäre der Aufbau weiterer Korpora wünschenswert, anhand derer der Erwerb des Deutschen als Fremd- oder Zweitsprache umfassend untersucht werden könnte.

LITERATUR

HINWEIS

Sämtliche im Beitrag angegebenen Webadressen sind zuletzt am 7.11.2017 abgefragt worden.

WEBADRESSEN DER VORGESTELLTEN KORPORA

FOLK – Forschungs- und Lehrkorpus gesprochenes Deutsch:

http://agd.ids-mannheim.de/FOLK_extern.shtml

Gesprochenes Deutsch für die Auslandsgermanistik:

<http://audiolabor.uni-muenster.de/daf/>

Backbone: <http://projects.ael.uni-tuebingen.de/backbone/moodle/>

Dann in der Menüleiste: Corpora & Search, Corpus search, Corpus: BB German

GeWiss – Gesprochene Sprache im akademischen Kontext

<https://gewiss.uni-leipzig.de/index.php?id=home>

BeMaTaC – Berlin Map Task Corpus:

<http://u.hu-berlin.de/bematac>

LeaP – Learning the Prosody of a foreign language:

https://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-linguistics/zum_loeschen/Forschung/leap/index.html

Download: https://sourceforge.net/projects/leapcorpus/?source=typ__redirect

LITERATUR

Belz, M., S. Sauer., A. Lüdeling und C. Mooshammer. Fluently Disfluent. Pauses and Repairs of Advanced Learners and Native Speakers of German. In: *International Journal of Learner Corpus Research* 3 (2017), 118-148.

Blombach, A. *Anleitung zur Benutzung von Korpora zu geschriebenem und gesprochenem Deutsch*. Erlangen: Universität Erlangen-Nürnberg, 2017.

<http://sprachwissenschaft.fau.de/personen/daten/blombach/korpora.pdf>

Boersma, P. und D. Weenink. *Praat: doing phonetics by computer (software tool)*. Amsterdam, 2001.

Costa, M. Datensammlungen zum gesprochenen Deutsch als Lehr- und Lernmittel. In: *Deutsch als Fremdsprache* 45, 2008, 133-139.

Deppermann, A. und T. Schmidt. Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). In: *Mitteilungen des deutschen Germanistenverbandes* 1, 2014, 4-17.

Fandrych, C. und E. Tschirner. Korpuslinguistik und DaF. Ein Perspektivenwechsel. In: *Deutsch als Fremdsprache* 4, 2007, 195-204.

- Fandrych, C., C. Meißner, S. Sadowski und F. Wallner (Hgg.). *Gesprochene Wissenschaftssprache – digital: Verfahren zur Annotation und Analyse mündlicher Korpora*. Tübingen: Stauffenburg, 2017.
- Fandrych, C., C. Meißner und A. Slavcheva (Hgg.). *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag, 2014.
- Fernández-Villanueva, M. und O. Strunk. Das Korpus Varkom – Variation und Kommunikation in der gesprochenen Sprache. In: *Deutsch als Fremdsprache* 46, 2009, 67-73.
- Giesel, L., M. Klapi, D. Krüger, I. Nunberger, O. Rasskazova und S. Sauer. Berlin Map Task Corpus: A deeply annotated multimodal map-task corpus of spoken learner and native German. Poster presented at the 35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Potsdam, 2013.
http://korpling.german.hu-berlin.de/bematac/publications/Giesel-et-al__2013__DGFS-CL-2013.pdf
- Gräfe, K. D. Lange, M. Sieradz, C. Meißner, A. Slavcheva und D. Stoppel. *GeWiss Gesprochene Wissenschaftssprache. Handbuch zum Korpus*. Leipzig 2015.
<https://gewiss.uni-leipzig.de/fileadmin/documents/Handbuch.pdf>
- Günthner, S. Diskursmarker in der Interaktion – zum Einbezug alltagssprachlicher Phänomene in den DaF-Unterricht. In: *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*, hrsg. von W. Imo und S. Moraldo. Tübingen: Stauffenburg, 2015, 135-164.
- Gut, U. Learner corpora in second language prosody research and teaching. In: *Non-native prosody: phonetic description and teaching practice*, hrsg. von J. Trouvain und U. Gut. Berlin usw.: Mouton de Gruyter, 2007a, 145-167.
- Sprachkorpora im Phonetikunterricht. In: *Zeitschrift für interkulturellen Fremdsprachenunterricht* 12, 2007b, 1-21.
- *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt/M. usw.: Peter Lang, 2009.
- Imo, W. *Sprache in Interaktion: Analysemethoden und Untersuchungsfelder*. Berlin: De Gruyter, 2013.
- Imo, W. und S. Moraldo (Hgg.). *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*. Tübingen: Stauffenburg, 2015.
- Káňa, T. *Sprachkorpora in Unterricht und Forschung DaF/DaZ*. Brno: Masarykova univerzita, 2014. https://is.muni.cz/repo/1201835/Sprachkorpora__DaF.pdf
- Kirk, J. M. und G. Andersen (Hgg.). *Compilation, transcription, markup and annotation of spoken corpora*. Special issue of *International Journal of Corpus Linguistics* 21(3), 2016.
- Kohn, K., P. Hoffstaedter und J. Widmann. BACKBONE – Pedagogic Corpora for Content & Language Integrated Learning. In: *Proceedings of the EUROCALL Conference, Valencia-Gandia, 9-12 Sept 2009*. Macmillan ELT, 2011.
- LeaP Corpus Manual* (o.J.) http://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-linguistics/workshop/pdfs/LeapCorpus__Manual.pdf
- Lemmitzer, L. und H. Zinsmeister. *Korpuslinguistik: Eine Einführung*. Tübingen: Narr, 2006.
- Moroni, M. Intonation im Gespräch. Zur Vermittlung der Intonation im DaF-Unterricht. In: *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*, hrsg. v. W. Imo und S. Moraldo. Tübingen: Stauffenburg, 2015, 67-82.

- Sauer, S. und A. Lüdeling. Flexible Multi-Layer Spoken Dialogue Corpora. In: *International Journal of Corpus Linguistics* 21(3), 2016, 419-438. [Preprint: http://korpling.german.hu-berlin.de/bematac/publications/Sauer-Luedeling_2016_IJCL.pdf]
- Sauer, S. und O. Rasskazova. BeMaTaC: Eine digitale multimodale Ressource für Sprach- und Dialogforschung. Poster presented at the workshop „Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen“. Berlin, 2014. http://korpling.german.hu-berlin.de/bematac/publications/Sauer-Rasskazova_2014_3WS-DHB.pdf
- Sauer, S., L. Giesel, M. Klapi, D. Krüger, I. Nunberger und O. Rasskazova. Gesprochene Muttersprache vs. Lernaltersprache. Aufbau und Auswertung eines Korpus. In: *Forschendes Lernen an der Humboldt-Universität zu Berlin. Die Q-Tutorien im Wintersemester 2012/2013. Eine Bilanz*. Berlin, 2013, 81-86. http://korpling.german.hu-berlin.de/bematac/publications/Giesel-et-al_2013_FLHU-QT-2012-2013.pdf
- Schiller, A., S. Teufel, C. Stöckert und C. Thielen. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Stuttgart, Tübingen, 1999. www.sfs.uni-tuebingen.de/resources/stts-1999.pdf
- Schmidt, T. Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: *Gesprächsforschung* 15, 2014, 196-233.
- Schneider, B. und S. Ylönen. Plädoyer für ein Korpus zur gesprochenen deutschen Wissenschaftssprache. In: *Deutsch als Fremdsprache* 45, 2008, 139-150.
- Selting, M. et al. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung* 10, 2009, 353-402.
- Spiegel, C. Transkripte als Arbeitsinstrument: Von der Arbeitsgrundlage zur Anschauungshilfe. In: *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung*, hrsg. v. K. Birkner und A. Stukenbrock. Verlag für Gesprächsforschung: Mannheim, 2009, 7-15.
- Westpfahl, S., T. Schmidt, J. Jonietz und A. Borlinghaus. *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS)*. Online-Publikation. Mannheim: Institut für Deutsche Sprache, 2017. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf
- Westpfahl, S. und T. Schmidt. POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: *Journal for Language Technology and Computational Linguistics* 28, 2013, 139-153.
- Wichmann, A. Speech corpora and spoken corpora. In: *Corpus linguistics: an international handbook*, hrsg. v. A. Lüdeling und M. Kytö. Berlin usw.: De Gruyter, 2008, 187-207.
- Ylönen, S. Qualitative und quantitative Methoden datengeleiteten Lernens. In: *German as a foreign language*, 2012 (2-3), 75-113. <http://www.gfl-journal.de/2-2012/Ylonen.pdf>