Online verfügbare Korpora
für Deutsch als Fremd- und Zweitsprache[1]

(Stand: 16.11.2020)

– bitte beachten Sie die Informationen zu den Korpuszugängen in den Fußnoten! –

1) Korpussammlungen

1.1 HZSK – Hamburger Zentrum für Sprachkorpora – dort Keyword ‚L2 data':

https://corpora.uni-hamburg.de/hzsk/en/repository-search?textQuery=&facetQuery=keywordOriginal%3A%22L2%20data%22 (16.11.2020)[2]

- **Commented Learner Corpus Academic Writing (KoLaS):** Authentische schriftliche Texte von Studierenden der Universität Hamburg, die Studierenden haben verschiedene Erstsprachen (L1) und verschiedene Studienfächer, alle Texte waren Gegenstand einer Schreibberatung der Schreibwerkstatt Mehrsprachigkeit, für manche Texte sind Kommentare von Tutorinnen/Tutoren sowie verschiedene Textversionen verfügbar. Sprache: German; Lizenz: HZSK-ACA (academic)

- **The Hamburg MapTask Corpus (HAMATAC):** Audioaufnahmen von Map-Task-Aufgaben bei Erwachsenen mit Deutsch als Zweitsprache. Die Kompetenzen der Sprecher/innen in Erst- und Zweitsprache variieren. Die in dieser Aufgabe benutzten Karten sind verfügbar. Sprache: German; Lizenz: HZSK-ACA (academic)

- **Hamburg Modern Times Corpus (HaMoTiC):** Audioaufnahmen von Nacherzählungen erwachsener L2 Sprecher/innen des Deutschen von Filmszenen eines Stummfilms. Die Kompetenzen der Sprecher/innen in Erst- und Zweitsprache variieren. Das Korpus umfasst 186 Minuten Aufnahmen von 29 Sprecherinnen und Sprechern mit einer jeweiligen Dauer von 2 bis 16 Minuten. Für jede/n Sprecher/in liegt eine Lernerbiographie vor. Sprache: German; Lizenz: HZSK-ACA (academic)

- **Hamburg Adult Bilingual LAnguage (HABLA):** Audioaufnahmen (semi-spontane Interviews) mit Deutsch/Italienisch und Deutsch/Französisch bilingualen Sprecherinnen/Sprechern im Alter zwischen 15 und 55 Jahren. Von jedem/jeder

[1] Siehe auch die Forschungsdatenbank CLARIN (https://www.clarin-d.net/de/), dort über „Suchen und Finden" einschlägige Stichwörter (L2, Biligualism, bilingual..).
Je nach Korpus kann es notwendig sein, sich frei verfügbare Software herunter zu laden, mit der die Korpora gelesen werden können, z.B. *Exmaralda* (https://exmaralda.org/de/) oder *CLAN* in der *CHILDES*-Datenbank (https://childes.talkbank.org/ - dort siehe „Programs").

[2] Achtung: Zugang zu den Korpora „Academic" bekommt nur, wer sich über seine/ihre Forschungsinstitution anmeldet. d.h. seine/ihre Mailadresse der Universität Potsdam verwendet. Wenn der Zugang „restructed" ist, muss er beantragt werden.

bilingualen Sprecher/in existieren 2 Aufnahmen – eine pro Sprache. Sprachen: Deutsch, Französisch, Italienisch; Lizenz: HZSK-RES (restricted)

1.2 Korpora am Max Planck Institut für Psycholinguistik in Nijmegen, „The Language Archive“[3]

L2 Acquisition: Hier sind mehrere Korpora zum Zweitspracherwerb bei Erwachsenen und Kindern zusammengestellt:[4]
https://archive.mpi.nl/islandora/object/lat%3A1839_00_0000_0000_0000_39AD_1 (16.11.2020)

- **Augsburger Korpus:** Erwerb des Deutschen als Zweitsprache bei 12 neuzugewanderten Kindern im Alter zwischen 6 und 11 Jahren mit Polnisch, Russisch und Türkisch als Erstsprache, über einen Zeitraum von zwei Jahren erhoben. Es liegen insgesamt Transkripte im Umfang von 120h vor.

- **DaZ-AF:** Fallstudie zum natürlichen Erwerb des Deutschen als Zweitsprache durch Lernende vor und nach Einsetzen der Pubertät. Die Spontansprachdaten zum Erwerb des Deutschen als Zweitsprache durch zwei russische Mädchen (8;7 und 14;2 Jahre alt) wurden ein Jahr lang je einmal wöchentlich (ca. 1h lang) erhoben.

- **ESF:** Mündliche Texte, die im Rahmen eines großen, vergleichenden Projekts der European Science Foundation (ESF) von jungen Erwachsenen aus mehreren Ländern der sog. „Gastarbeiter-Anwerbung“ in den 80er und 90er Jahren in mehreren Zielsprachen erhoben wurden, die die Probanden ungesteuert erwarben. Aus diesem Projekt ist u.a. das Konzept der „Basic Variety“ (vgl. Klein & Perdue 1997) entstanden.

- **P-MoLL: "Modalität in Lernervarietäten im** Längsschnitt". Das Projekt befasste sich mit der Untersuchung des Erwerbs der Modalität in Deutsch als Zweitsprache durch ungesteuert Deutsch erwerbende erwachsene Einwanderer mit L1 Polnisch und Italienisch. Die Längsschnittdatenerhebung umfasst etwa zweieinhalb Jahre des Erfassungsprozesses der Lernenden. Es enthält ihre mündliche Sprachproduktion aus verschiedenen Erhebungsaufgaben und freien Gesprächen mit Muttersprachlerinnen/Muttersprachlern. Daten einer Kontrollgruppe mit Deutsch L1 stehen zur Verfügung.

1.3 Falko-Lernerkorpora:

Falko ist eine Zusammenstellung von fehlerannotierten lerner/innensprachlichen Korpora des Deutschen als Fremdsprache: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/design (16.11.2020)
Suchmaske/Abfrage: https://korpling.german.hu-berlin.de/falko-suche/ (16.11.2020)

- Das **Essay-Kernkorpus** enthält argumentative Aufsätze von fortgeschrittenen Lernenden des DaF mit unterschiedlichen Erstsprachen.

[3] Korpusdarstellung in Skiba 2008.
[4] Texte aus dem DaZ-AF-Korpus, dem Augsburger Korpus und dem P-Moll-Korpus sind auch am Arbeitsbereich DaF/DaZ der UP verfügbar.

- Das **WHiG-Korpus** enthält wie das Essay-Korpus argumentative Aufsätze von fortgeschrittenen DaF-Lernenden (zu denselben Themenbereichen wie das Essay-Korpus), hier allerdings mi einem sprachlich homogenen L1-Hintergrund (Englisch).

- Das **Kobalt-DaF-Korpus** besteht aus drei Subkorpora mit schwedischer, chinesischer und weißrussischer Muttersprache sowie einem Deutsch-L1-Vergleichskorpus und folgt den Falko-Erhebungs- und Aufbereitungsrichtlinien.

- Das **KanDel-Korpus** (Nina Vyatkina, Kansas, USA) enthält im Gegensatz zu den übrigen Falko-Korpora geschriebene Daten von beginnenden US-amerikanischen Lernenden des DaF, außerdem wurden diese Daten longitudinal aufbereitet.

- Das **Zusammenfassungskorpus** enthält Textzusammenfassungen, die von fortgeschrittenen Lernenden des Deutschen erstellt wurden.

- Das **Georgetown-Longitudinalkorpus** enthält Daten, die über mehrere Semester und Lernstände an der Georgetown-Universität in Washington erhoben wurden. Dazu gibt es ein Vergleichskorpus mit Texten von Muttersprachlerinnen/Muttersprachlern für das Genre der Buchrezensionen.

- Das **BeMaTACBerlin Map Task Corpus (BeMaTaC)** ist ein frei verfügbares Korpus gesprochener Sprache. Es besteht aus einem L1-Subkorpus, welches mit deutschen Muttersprachlerinnen/Muttersprachlern aufgenommen wird, und einem identisch angelegten L2-Subkorpus mit fortgeschrittenen Lernenden von Deutsch als Fremdsprache. 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 Sprecher/innen können sich nicht gegenseitig sehen und können daher nicht non-verbal kommunizieren.

1.4 Korpora in DGD-Datenbank für gesprochenes Deutsch: http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome (16.11.2020)[5]
In der DGD-Datenbank für gesprochenes Deutsch ist eine Großzahl von Korpora zum Deutschen abgelegt. Zwei dieser Korpora enthalten lerner/innensprachliche Daten:

- **Mehrsprachige Kinder im Vorschulalter MEKI**: (von Elke Montanari) enthält gesprochensprachliche Daten von mehrsprachigen Kindern, die über einen Zeitraum von knapp einem Jahr begleitet wurden. Die Daten wurden im Rahmen einer Sprachfördermaßnahme erhoben und dokumentieren die sprachliche Entwicklung der Kinder. Näheres: http://agd.ids-mannheim.de/MEKI_extern.shtml (16.11.2020)

- **SA - Kindersprache:** Saarbrücker Korpus, untersucht den späten ungesteuerten Spracherwerb türkischer und italienischer Kinder. Das Korpus umfasst 65 Tonaufnahmen aus der Zeit von 1982 bis 1984 mit einer Gesamtdauer von 4 Stunden und 33 Minuten, die in Situationen teilnehmender Beobachtung im Saarland gemacht wurden. Es handelt sich um Aufnahmen von Kind-Erwachsenen-Interaktionen. Bei den Probandinnen/Probanden handelt es sich um zwei türkische, zwei italienische

[5] Zugang muss beantragt werden (mit der universitären Email).

und zwei deutsche Kinder im Alter von 9 bis 13 Jahren. Näheres: http://agd.ids-mannheim.de/SA--_extern.shtml (16.11.2020)

2. Einzelkorpora

2.1 MERLIN:

http://www.merlin-platform.eu/ (16.11.2020)
MERLIN bietet Zugang zu 2.286 schriftlichen Texten von Lernenden des Tschechischen, des Italienischen und vor allem des Deutschen als Fremdsprache. Die Lerner/innentexte stammen aus standardisierten Sprachtests und sind auf die GER-Niveaus bezogen.

2.2 Materialien www.daz-portal.de:

http://www.daz-portal.de/images/Berichte/bm_band_02_ricart-brede_20140730.pdf (16.11.2020)
Ricart Brede, Julia (2014): Vorschulische Sprachfördersituationen Ein aufbereiteter und kommentierter Transkriptband aus dem Projektkontext von „Sag‘ mal was – Sprachförderung für Vorschulkinder“. Materialien DaZ-Portal, Bd. 2.

2.3 DiGS „Deutsch an Genfer Schulen“:

http://www.unige.ch/lettres/alman/fr/recherche/digs/ (16.11.2020)
In dem Forschungsprojekt "DiGS" wurden Texte von Schülerinnen und Schülern mit Französischer Erstsprache und Deutsch als Fremdsprache aus unterschiedlichen Jahrgangsstufen in einem pseudo-longitudinalen Design erhoben. Die Ergebnisse des Forschungsprojektes wurden veröffentlicht:
Diehl, E.; Studer, T.; Christen, H.; Leuenberger, S.; Pevat, I.(2000): Grammatikunterricht - alles für der Katz? Untersuchungen zum Zweitprachenerwerb Deutsch. Tübingen, Niemeyer (RGL 220).

2.7 “Linguistic Fieldnotes“:

URL: http://opus.kobv.de/ubp/volltexte/2010/3683/ (16.11.2020)
schriftlich produzierte Daten, die von drei verschiedenen Sprechergruppen stammen: Jugendliche aus einem multiethnischen Berliner Wohngebiet, die untereinander Kiezdeutsch sprechen, Jugendliche aus einem monoethnischen Berliner Wohngebiet, in dem der traditionelle Berliner Dialekt vorherrscht, und türkische Jugendliche in İzmir, die Deutsch als Fremdsprache gesteuert erworben haben.
Freywald, Ulrike ; Mayr, Katharina ; Schalowski, Sören ; Wiese, Heike. 2010. Linguistic Fieldnotes II: Information structure in different variants of written German. Universität Potsdam / Schriftenreihen / Interdisciplinary studies on information structure : ISIS ; working papers of the SFB 632: Kiezdeutsch-Texte, sowie Frog-Stories schriftlich DaF Izmirer jugendlicher Deutschlerner.

2.8 GeWiss – Gesprochene Wissenschaftssprache:

https://gewiss.uni-leipzig.de (16.11.2020)[6]
GeWiss ist ein Projekt zur Erforschung der gesprochenen Wissenschaftssprache. Es verfolgt das Ziel, eine empirische Grundlage für vergleichende Untersuchungen in diesem Bereich zu schaffen. Zu diesem Zweck wurde ein Korpus erstellt, das zwei zentrale

[6] Zugang muss beantragt werden (mit der universitären Email)

Genres der gesprochenen Wissenschaftssprache erfasst: Vortrag (einschließlich Diskussion) sowie Prüfungsgespräch. Das Korpus wird fortlaufend ausgebaut und weiterentwickelt. Das GeWiss-Korpus ist ein Vergleichskorpus und enthält authentische Daten vergleichbarer Genres aus verschiedenen Sprachen. Es eröffnet Vergleichsmöglichkeiten in zwei Dimensionen: zum einen im Hinblick auf den Gebrauch des Deutschen als fremder Wissenschaftssprache in verschiedenen akademischen Kontexten, zum anderen in Bezug auf andere Wissenschaftssprachen. In der Kernversion der Ressource umfasste dies neben Daten aus Deutschland deutschsprachige Aufnahmen von L2-Sprecherinnen und -Sprechern im britischen und polnischen akademischen Kontext sowie im selben Kontext erhobene polnisch- bzw. englischsprachige L1-Daten. Im Rahmen des Kurationsprojektes kamen zwei weitere Teilkorpora hinzu: ein Korpus von in Bulgarien erhobenen deutschsprachigen Seminarreferaten gehalten von Germanistikstudierenden mit der L1 Bulgarisch sowie ein Korpus mit Konferenzvorträgen in der L1 Italienisch.

2.9 Osnabrücker Bildergeschichtenkorpus

https://docplayer.org/30656168-Osnabruecker-bildergeschichtenkorpus.html (16.11.2020)

Frei verfügbare Sammlung verschrifteter Bildergeschichten von Schülerinnen und Schülern der 2. Klasse. Die Sammlung des Korpus ist Teil des Dissertationsvorhabens ”Automatische Analyse orthographischer Fehler von Schreibanfängern“ von Tobias Thelen und ist 1999 entstanden.

Informationen zur Schrifteinführung und Vermittlung orthographischer Regularitäten sowie die Herkunft der Klassen und einzelner Schülerinnen und Schüler wurden per Fragebogen erhoben und stehen zur Verfügung. Die Texte sind hinsichtlich grammatischer und orthographischer Abweichungen sowie Fehlern hinsichtlich der Zeichensetzung annotiert.

2.10 Der kleine Prinz – Eine Weltreise in hundert Sprachen

http://der-kleine-prinz-in-100-sprachen.de/projekt (16.11.2020)

In rund 190 Aufnahmen lesen Sprecherinnen und Sprecher in Ihren Muttersprachen aus aller Welt Ausschnitte aus dem Kleinen Prinzen.