

## **Ergänzung zu Kapitel 7.3: Übersicht zu Einzelkorpora**

In dieser Datei haben wir für Sie eine Reihe von Einzelkorpora zusammengestellt, die eine gewisse Übersicht über die deutsche Korpuslandschaft geben. In den ersten beiden Auflagen des Buches war die Tabelle Teil von Kapitel 5. In der 3. Auflage haben wir uns entschieden, diese Tabelle auf die Webseite auszulagern. Die vorliegende Datei stellt einen „Schnappschuss“ aus dem Herbst 2015 dar. Die PDF-Datei enthält aktive Links, so dass Sie mit Klicken auf eine URL direkt auf die verlinkte Web-Seite weitergeleitet werden.

Manche der Korpora werden im Buch in Kapitel 7.3.9 zusätzlich im Zusammenhang mit den zentralen Archiven aufgeführt. Für die Merkmale, die bei der Beschreibung einzelner Korpora fehlen, konnten wir leider keine Werte ermitteln.

**Legende:** **Z:** Zugang, **F:** Funktion, **S:** Sprachenauswahl, **M:** Medium, **G:** Größe, **P:** Persistenz, **A:** Annotation, **SB:** Sprachbezug, **V:** Verfügbarkeit

### **Referenzkorpus Althochdeutsch (750–1050)**

**Z:** <http://www.deutschdiachrondigital.de/home/>

**F:** Sprachdokumentation

**S:** Deutsch

**M:** geschrieben

**P:** statisch

**A:** Morphosyntax, Lemma

**SB:** Referenzkorpus

**V:** online (ANNIS: <https://korpling.german.hu-berlin.de/annis3/ddd>)

### **Attritions-Korpus**

**Z:** [mschmid@essex.ac.uk](mailto:mschmid@essex.ac.uk) (Monika Schmid)

**F:** Empirische Basis zum Sprachverlust bei deutschen Emigranten im anglophonen Kanada und in den Niederlanden

**S:** Deutsch

**M:** Audiodaten und Transkripte

**G:** ca. 0,5 Millionen Token

**P:** statisch

**SB:** Deutsche Migranten in Kanada und den Niederlanden

**V:** auf Anfrage

### **Bonner Frühneuhochdeutschkorpus**

**Z:** <http://www.korpora.org/>

**F:** Materialgrundlage für die Erarbeitung der Grammatik des Frühneuhochdeutschen

**S:** Frühneuhochdeutsch

**M:** geschrieben

**G:** 1500 Texte (Gesamt), 40 Texte (Auswahlkorpus)

**P:** statisch

**A:** Morphosyntax, Morphologie

**SB:** Sprachstadium

**V:** frei

**Bonner Zeitungskorpus**

**Z:** [www.ids-mannheim.de/kl/projekte/korpora/archiv/bzk.html](http://www.ids-mannheim.de/kl/projekte/korpora/archiv/bzk.html), [korpusinguistik@ids-mannheim.de](mailto:korpusinguistik@ids-mannheim.de)  
(Institut für Deutsche Sprache)  
**F:** Teil der IDS-Korpora  
**S:** Deutsch  
**M:** geschrieben  
**G:** 3,1 Millionen Token  
**P:** statisch  
**A:** Textstruktur  
**SB:** Spezial: Tageszeitungen (Neues Deutschland (DDR), Die Welt (BRD))  
**V:** mit kostenfreier Lizenz herunterladbar (<http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html>); online über COSMAS

**Bonner Zeitungskorpus CG-annotiert**

**Z:** [http://beta.visl.sdu.dk/corpus\\_linguistics.html](http://beta.visl.sdu.dk/corpus_linguistics.html), [eckhard.bick@mail.dk](mailto:eckhard.bick@mail.dk) (Eckhard Bick)  
**S:** Deutsch  
**M:** geschrieben  
**G:** 4 Millionen Token  
**P:** statisch  
**SB:** Tageszeitung  
**V:** online mit Lizenz, Abfrage als „BZK“: <http://corp.hum.sdu.dk/cqp.de.html>

**Chat-Korpus**

**Z:** <http://www.chatkorpus.uni-dortmund.de>, [michael.beisswenger@tu-dortmund.de](mailto:michael.beisswenger@tu-dortmund.de) (Michael Beißwenger)  
**F:** Grundlage und Hilfsmittel für sprachwissenschaftliche Untersuchungen zur synchronen internetbasierten Kommunikation (IBK)  
**S:** Deutsch  
**M:** geschrieben  
**G:** 600 000 Token  
**P:** statisch  
**SB:** Logfiles aus Chatkommunikation  
**V:** frei, online

**Compact Memory – Korpus jüdischer Periodika**

**Z:** <http://www.dwds.de/>, [dwds@bbaw.de](mailto:dwds@bbaw.de) (Berlin-Brandenburgische Akademie der Wissenschaften)  
**F:** Korpus jüdischer Periodika, Teil der Textbasis für das Digitale Wörterbuch der Deutschen Sprache (DWDS)  
**S:** Deutsch  
**M:** geschrieben  
**G:** 26 Millionen Token  
**P:** statisch  
**A:** Lemma, Morphosyntax  
**SB:** Spezialkorpus  
**V:** online abfragbar unter <http://retro.dwds.de/> (Stand 11/2015; zukünftig auch über: <http://www.dwds.de/>)

**DECOW**

**Z:** <http://corporafromtheweb.org/>, [roland.schaefer@fu-berlin.de](mailto:roland.schaefer@fu-berlin.de), [felix.bildhauer@fu-berlin.de](mailto:felix.bildhauer@fu-berlin.de)

(Roland Schäfer, Felix Bildhauer)

**F:** Webkorpus zur allgemeinen Nutzung für die linguistische Erforschung der deutschen Gegenwartssprache

**S:** Deutsch

**M:** geschrieben

**G:** 20 Milliarden bzw. 12 Milliarden Token (DECOW14)

**P:** statisch

**A:** Lemma, Morphosyntax, Named Entities, Chunks

**SB:** Referenzkorpus

**V:** herunterladbar; nach Anmeldung online abfragbar über <https://webcorpora.org/> (zufällige Satzreihenfolge)

#### **Deutsches Rechtswörterbuch, Textarchiv**

**Z:** <http://drw-www.adw.uni-heidelberg.de/drw-cgi/zeige> (Heidelberger Akademie der Wissenschaften)

**F:** Bereitstellung der Textquellen für das Deutsche Rechtswörterbuch

**S:** Deutsch

**M:** geschrieben

**G:** unbekannt

**P:** statisch

**A:** keine

**SB:** Spezialkorpus

**V:** online abfragbar

#### **Deutsches Textarchiv**

**Z:** <http://www.deutschestextarchiv.de/>, [dta@bbaw.de](mailto:dta@bbaw.de) (Berlin-Brandenburgische Akademie der Wissenschaften)

**F:** Dokumentation des Sprachstadiums von 1650-1900

**S:** Deutsch

**M:** geschrieben

**G:** ca. 100 Millionen Token

**P:** statisch

**A:** Lemma, Morphosyntax, normierte Schreibung

**SB:** Referenzkorpus

**V:** online abfragbar, komplette Texte herunterladbar

#### **deWaC**

**Z:** <http://wacky.sslmit.unibo.it/doku.php?id=corpora>, [wacky@sslmit.unibo.it](mailto:wacky@sslmit.unibo.it) (Universität Bologna u.a.)

**F:** Webkorpus zur allgemeinen Nutzung für die linguistische Erforschung der deutschen Gegenwartssprache

**S:** Deutsch

**M:** geschrieben

**G:** ca. 1,7 Milliarden Token

**P:** statisch

**A:** Lemma, Morphosyntax

**SB:** Referenzkorpus

**V:** frei online nutzbar und herunterladbar

**Dolmetschen im Krankenhaus (DiK)**

**Z:** [https://www.corpora.uni-hamburg.de/sfb538/de\\_k2\\_dik.html](https://www.corpora.uni-hamburg.de/sfb538/de_k2_dik.html), corpora@uni-hamburg.de  
(Hamburger Zentrum für Sprachkorpora)

**S:** Deutsch, Portugiesisch, Türkisch

**M:** gesprochen

**G:** 92 Aufnahmen (Gesamtdauer: 23:01:18.53 Stunden), ca. 166 000 transkribierte Token

**P:** statisch

**A:** Transkription, deutsche Übersetzung

**SB:** Domäne: Arzt-Patient-Kommunikation, Laiendolmetschen

**V:** Auf Anfrage.

**DWDS-Kernkorpus 20**

**Z:** <http://www.dwds.de/>, dwds@bbaw.de (Berlin-Brandenburgische Akademie der Wissenschaften)

**F:** Teil der Textbasis für das Digitale Wörterbuch der Deutschen Sprache

**S:** Deutsch

**M:** geschrieben

**G:** 100 Millionen Token

**P:** statisch

**A:** Lemma, Morphosyntax

**SB:** Referenzkorpus

**V:** online abfragbar

**Emigrantendeutsch in Israel**

**Z:** <http://agd.ids-mannheim.de/index.shtml> (Archiv für gesprochenes Deutsch)

**F:** Sprachdokumentation

**S:** Deutsch

**M:** gesprochen

**P:** statisch

**A:** Transkripte

**SB:** Gruppensprache emigrierter deutschsprachiger Juden

**V:** nach Registrierung online nutzbar

**Europarl-Korpus**

**Z:** <http://www.statmt.org/europarl/>, pkoehn@inf.ed.ac.uk (Philipp Koehn)

**F:** Textbasis für Maschinelle Übersetzung

**S:** Version 7: Deutsch, Englisch und 19 weitere EU-Sprachen, satzaligniert mit Englisch

**M:** gesprochen (editiert)

**G:** 44,5 Millionen Token (DEU)

**P:** statisch

**SB:** Spezialkorpus: European Parliament Proceedings Parallel Corpus 1996-2011

**V:** frei

**Europarl CG-annotiert**

**Z:** [http://beta.visl.sdu.dk/corpus\\_linguistics.html](http://beta.visl.sdu.dk/corpus_linguistics.html), eckhard.bick@mail.dk  
(Eckhard Bick)

**F:** Sprachlehre, empirische Grundlage für computerlinguistische Forschung

**S:** Deutsch

**M:** gesprochen (editiert)

**G:** 29 Millionen Token

**P:** statisch

**A:** Morphosyntax, Morphologie, CG-Dependenzen

**SB:** Spezialkorpus, siehe Europarl-Korpus

**V:** frei, online abfragbar: [corp.hum.sdu.dk](http://corp.hum.sdu.dk)

#### **European Corpus Initiative CG-annotiert (ECI-DE1, Frankfurter Rundschau)**

**Z:** [http://beta.visl.sdu.dk/corpus\\_linguistics.html](http://beta.visl.sdu.dk/corpus_linguistics.html), [eckhard.bick@mail.dk](mailto:eckhard.bick@mail.dk)  
(Eckhard Bick)

**S:** Deutsch

**M:** geschrieben

**G:** 34 Millionen Token

**P:** statisch

**SB:** Zeitungsdeutsch

**V:** online mit Lizenz abfragbar: [corp.hum.sdu.dk](http://corp.hum.sdu.dk)

#### **Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache (FALKO)**

**Z:** <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko>, [anke.luedeling@rz.hu-berlin.de](mailto:anke.luedeling@rz.hu-berlin.de) (Anke Lüdeling)

**F:** Untersuchungen zur „Interlanguage“ von Fremdsprachenlernenden; Kontrollkorpus muttersprachlicher Texte

**S:** Deutsch als Fremdsprache

**M:** geschrieben

**G:** Version 2.0: 380 000 Token

**P:** statisch

**A:** Morphosyntax, Lemmata, teilw. Dependenzsyntax, Normalisierungsebenen („Zielhypothesen“), Fehlerannotation des Verbalsystems

**SB:** Lernaltersprache

**V:** frei, Online-Abfrage über ANNIS: <https://korpling.german.hu-berlin.de/falko-suche/>

#### **Referenzkorpus Frühneuhochdeutsch (1350–1650)**

**Z:** <http://www.ruhr-uni-bochum.de/wegera/ref/>

**F:** Sprachdokumentation

**S:** Deutsch

**M:** geschrieben

**G:** 7 Millionen transkribierte, 4,4 Millionen annotierte Token

**P:** statisch

**A:** Morphosyntax, Lemma

**SB:** Referenzkorpus

**V:** online über ANNIS im Aufbau

#### **Goethe-Korpus**

**Z:** <http://www.ids-mannheim.de/kl/projekte/korpora/archiv/goe.html> (Institut für Deutsche Sprache)

**F:** Teil der IDS-Korpora (Historische Korpora)

**S:** Deutsch

**M:** geschrieben

**G:** 1,4 Millionen Token

**P:** statisch

**A:** Textstruktur

**SB:** Individualsprache: 29 Texte aus Goethes Werken

V: online über COSMAS

#### **Hamburg Dependency Treebank**

Z: <https://corpora.uni-hamburg.de/drupal/de/islandora/object/treebank:hdt> menzel@informatik.uni-hamburg.de (Wolfgang Menzel)

F: computerlinguistische Forschung

S: Deutsch

M: mehrheitlich geschrieben

G: ca. 100 000 (manuell) bzw. 250 000 (teil-manuell annotierte) Sätze / mehr als 12 Millionen Token

P: statisch

A: Dependenzsyntax

SB: 75% Online-News, diverse andere Texttypen

V: frei, herunterladbar; satzweise online visualisierbar: WebLicht (frei), Tundra (nach Anmeldung)

#### **Handbuchkorpora**

Z: <http://www.ids-mannheim.de/kt/projekte/korpora/archiv/hbk.html> (Institut für Deutsche Sprache)

F: Teil der IDS-Korpora: Datengrundlage für Handbucherstellung

S: Deutsch

M: geschrieben

G: 11 Millionen Token

P: statisch

A: Textstruktur

SB: Zeitungs- und Zeitschriftenartikel: 1985-1988

V: online über COSMAS

#### **HyTex-Korpus**

Z: <http://www.hytext.tu-dortmund.de/ressourcen.html#korpora>, michael.beisswenger@tu-dortmund.de (Michael Beißwenger)

F: Forschungen zu Hypertextstrukturen

S: Deutsch

M: geschrieben

G: ca. 25 000 Normseiten

P: statisch

A: Syntax (automatisch): Chunking mit KoRoParse für 28 Dokumente; Textgrammatik

SB: Fachtexte V: herunterladbar

#### **INTERSECT Corpus**

Z: <http://arts.brighton.ac.uk/staff/raf-salkie/intersect>, r.m.salkie@bton.ac.uk (Raphael Salkie)

F: Sprachlehre

S: Deutsch, Englisch: Parallelkorpus, satzaligniert

M: geschrieben

G: 800 000 Token

P: statisch

SB: Domänen: Wirtschaft und Politik

V: auf Anfrage

**Bonner Kant-Korpus**

**Z:** [Korpora.org](http://www.korpora.org), [bernhard.schroeder@uni-due.de](mailto:bernhard.schroeder@uni-due.de) (Bernhard Schröder)

**S:** Deutsch

**M:** geschrieben

**P:** statisch

**A:** teilweise lemmatisiert

**SB:** Individualsprache: Texte des Philosophen Kant

**V:** herunterladbar und online abfragbar

**Kieler Korpus gesprochener Sprache**

**Z:** <http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html> (Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel)

**F:** Materialgrundlage für computerlinguistische Anwendungen

**S:** Deutsch

**M:** gesprochen

**G:** 4 Stunden

**P:** statisch

**SB:** Lese- und Spontansprache

**V:** kostenpflichtige Lizenz, über Webseite zu bestellen

**Korpus Magazin Lufthansa Bordbuch / DEUTSCH**

**Z:** <http://www.ids-mannheim.de/kt/projekte/korpora/archiv/mld.html>, [belica@ids-mannheim.de](mailto:belica@ids-mannheim.de) (Institut für Deutsche Sprache)

**F:** Teil der IDS-Korpora

**S:** Deutsch

**M:** geschrieben

**G:** 250 000 Token

**P:** statisch

**A:** Textstruktur

**SB:** Spezialkorpus: Tourismus, Lufthansa: 1995-1997

**V:** online über COSMAS

**LIMAS-Korpus**

**Z:** <http://www.korpora.org/Limas/>, [bernhard.schroeder@uni-due.de](mailto:bernhard.schroeder@uni-due.de) (Bernhard Schröder)

**S:** Deutsch

**M:** geschrieben

**G:** 1 Millionen Token

**P:** statisch

**A:** Morphosyntax, Lemma (maschinell)

**V:** frei, online (zusätzlich über COSMAS)

**Mannheimer Korpus 1**

**Z:** <http://www.ids-mannheim.de/kt/projekte/korpora/archiv/mk.html>, [belica@ids-mannheim.de](mailto:belica@ids-mannheim.de) (Institut für Deutsche Sprache)

**F:** Teil der IDS-Korpora

**S:** Deutsch

**M:** geschrieben

**G:** 2,2 Millionen Token

**P:** statisch

**A:** Textstruktur

**SB:** div. Textgenres: 1950-1967

**V:** mit kostenfreier Lizenz herunterladbar (<http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html>); online über COSMAS

#### **Mannheimer Korpus 2**

**Z:** <http://www1.ids-mannheim.de/kt/projekte/korpora/archiv/mk.html> (Institut für Deutsche Sprache)

**F:** Teil der IDS-Korpora

**S:** Deutsch

**M:** geschrieben

**G:** 300 000 Token

**P:** statisch

**A:** Textstruktur

**SB:** Diverse Textgenres: 1950-1967

**V:** mit kostenfreier Lizenz herunterladbar (<http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html>); online über COSMAS

#### **Mannheimer Korpus CG-annotiert**

**Z:** [http://beta.visl.sdu.dk/corpus\\_linguistics.html](http://beta.visl.sdu.dk/corpus_linguistics.html), eckhard.bick@mail.dk (Eckhard Bick)

**S:** Deutsch

**M:** geschrieben

**G:** 4 Millionen Token

**P:** statisch

**SB:** Zeitungskorpus

**V:** online-Abfrage (mit Lizenz): [corp.hum.sdu.dk](http://corp.hum.sdu.dk)

#### **Mannheimer Morgen**

**Z:** <http://www1.ids-mannheim.de/kl/projekte/korpora/archiv/mm.html> (Institut für Deutsche Sprache)

**F:** Teil der IDS-Korpora

**S:** Deutsch

**M:** geschrieben

**G:** Stand 2007: 244,20 Millionen Token (siehe Webseite (letzter Zugriff 11/2015), wächst kontinuierlich)

**P:** Monitorokorpus

**A:** Textstruktur

**SB:** Tageszeitung: 1989-2012

**V:** online über COSMAS

#### **Referenzkorpus Mittelhochdeutsch (1050–1350)**

**Z:** <http://referenzkorpus-mhd.uni-bonn.de/>

**F:** Sprachdokumentation

**S:** Deutsch

**M:** geschrieben

**P:** statisch

**A:** Morphosyntax, Lemma

**SB:** Referenzkorpus

**V:** online über ANNIS im Aufbau



**NEGRA-Korpus (Version 2)**

**Z:** <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html>

(Universität des Saarlandes)

**F:** Empirische Basis für linguistische Forschung und maschinelles Lernen

**S:** Deutsch

**M:** geschrieben

**G:** 360 000 Token

**P:** statisch

**A:** Morphosyntax, Konstituenten, grammatische Funktionen (halbautomatisch)

**SB:** Tageszeitung: Frankfurter Rundschau

**V:** Lizenz

**OPUS-Korpora**

**Z:** <http://opus.lingfil.uu.se/>, [jorg.tiedemann@lingfil.uu.se](mailto:jorg.tiedemann@lingfil.uu.se) (Jörg Tiedemann)

**F:** Datengrundlage für computerlinguistische Anwendungen

**S:** viele Sprachen, satz- oder wortaligniert

**P:** statisch

**A:** Token, Morphosyntax und partiell Syntax (maschinell)

**SB:** Spezialkorpora: politische Texte, Computermanuals, Untertitel, Belletristik, u.a.

**V:** frei, online

**Oslo Multilingual Corpus (OMC)**

**Z:** <https://www.hf.uio.no/ilos/english/services/omc/>

**S:** Deutsch und Englisch

**M:** geschrieben

**G:** ca. 700 000 Token (deutscher Teil)

**P:** statisch

**SB:** Literarische und Gebrauchstexte

**V:** auf Mitglieder der Universitäten Oslo und Bergen beschränkt

**Potsdam Commentary Corpus (PCC 2.0)**

**Z:** <http://angcl.ling.uni-potsdam.de/resources/pcc.html>,

[stede@uni-potsdam.de](mailto:stede@uni-potsdam.de) (Manfred Stede)

**F:** Untersuchungen zur Informationsstruktur von Texten

**S:** Deutsch

**M:** geschrieben

**G:** verfügbares Kernkorpus: 176 Artikel = 32 800 Token

**P:** statisch

**A:** halbautomatisch: Morphosyntax, Syntax, Morphologie, Lemma, Konnektoren, Diskursrelationen, Koreferenz

**SB:** Spezialkorpus: Zeitungskommentare aus MAZ (Märkische Allgemeine Zeitung)

**V:** frei, online (2 Texte in ANNIS-Demo: <https://korpling.german.hu-berlin.de/annis3/>)

**Deutscher Wortschatz-Korpus**

**Z:** <http://corpora.informatik.uni-leipzig.de/>, [wort@informatik.uni-leipzig.de](mailto:wort@informatik.uni-leipzig.de)

**F:** Sprachstatistische Analysen

**S:** Deutsch

**M:** geschrieben

**G:** deu\_newscrawl\_2011: ca. 425 Millionen Token

**P:** statisch

**A:** Morphosyntax, Lemma, Semantik (semantische Primitive und Relationen zwischen Wörtern)  
**SB:** Opportunistisches Korpus aus Webquellen von 2011  
**V:** frei, online

**Saarbrücken Lexical Semantics Acquisition Project (SALSA-Korpus)**

**Z:** <http://www.coli.uni-saarland.de/projects/salsa/> (Universität des Saarlandes)  
**F:** Materialbasis für computerlinguistische Anwendungen  
**S:** Deutsch  
**M:** geschrieben  
**P:** statisch  
**A:** Frame-Semantik, Koreferenz, (baut auf TIGER-Korpus auf, daher auch Syntax usw.)  
**SB:** Zeitungssprache: Frankfurter Rundschau  
**V:** Versionen 1.0 und 2.0 herunterladbar mit kostenfreier Lizenz (<http://www.coli.uni-saarland.de/projects/salsa/corpus/request/salsa-corpus-request.cgi>), Web-Demo für SALSA 1.0 (<http://fnps.coli.uni-saarland.de:8080/>).

**SMS-Korpus**

**Z:** <http://www.mediensprache.net/de/medienanalyse/corpora/>,  
info@mediensprache.net (Peter Schlobinski, Torsten Siever)  
**F:** Untersuchungen zu medienspezifischen Sprachvarietäten  
**S:** Deutsch  
**M:** geschrieben  
**G:** a) c.a 1500 SMS b) 98 SMS **P:** statisch  
**SB:** Spezialkorpus: a) SMS von Schülerinnen/Schülern und Studierenden; b) Info-SMS; Werbemittel  
**V:** a) als PDF-Datei herunterladbar und nach Anmeldung (<http://www.mediensprache.net/de/ussr/register.aspx>) online durchsuchbar als SMS-Mitteilungen (H) und SMS-Mitteilungen (OS); b) online (<http://www.mediensprache.net/de/werbesprache/handy/>)

**TIGER (Version 2)**

**Z:** <http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html>,  
clarin@ims.uni-stuttgart.de (Universität Stuttgart)  
**F:** Materialgrundlage für linguistische und computerlinguistische Forschung  
**S:** Deutsch  
**M:** geschrieben  
**G:** Version 2: 0,9 Millionen Token  
**P:** statisch  
**A:** Morphosyntax, Morphologie, Lemma, Syntax (Konstituenten und funktionale Information), Semantik (Eigennamen)  
**SB:** Zeitungssprache (Frankfurter Rundschau)  
**V:** herunterladbar mit kostenfreier Lizenz (<http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/license/index.html/>); Web-Demo für Tiger 2.1 (<http://fnps.coli.uni-saarland.de:8080/>).

**Tübinger Baubank des Deutschen / Zeitungskorpus (TüBa-D/Z)**

**Z:** [www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html](http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html),  
tuebadz-info@sfs.uni-tuebingen.de (Marie Hinrichs)  
**F:** Materialgrundlage für linguistische und computerlinguistische Forschung  
**S:** Deutsch  
**M:** geschrieben

**G:** Version 10: 1,8 Millionen Token

**P:** statisch

**A:** Morphosyntax, Morphologie, Lemma, Syntax (Konstituenten und funktionale Information), topologische Felder, Semantik (Eigennamen, GermaNet-Lesarten), Koreferenz, Diskurskonnektoren

**SB:** Spezialkorpus: die tageszeitung

**V:** mit kostenfreier Lizenz herunterladbar; online nach Anmeldung über TüNDRA (<http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra>)

#### **Tübinger Baubank des Deutschen / Spontansprache (TüBa-D/S)**

**Z:** <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-ds.html>,  
tuebadz-info@sfs.uni-tuebingen.de (Marie Hinrichs)

**F:** Datenbasis für Maschinelle Übersetzung (Verbmobil-Projekt)

**M:** gesprochen

**G:** 360 000 Token

**P:** statisch

**A:** Morphosyntax, Morphologie, Syntax (Konstituenten und funktionale Information), topologische Felder, Eigennamen (manuell)

**SB:** Spezialkorpus: Dialoge (Terminvereinbarungen)

**V:** mit kostenfreier Lizenz herunterladbar; online nach Anmeldung über TüNDRA (<http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra>)

#### **Tübinger Partiiell Gearpastes Korpus des Deutschen/Schriftsprache (TüPP-D/Z)**

**Z:** <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tuepp-dz.html>,  
tuebadz-info@sfs.uni-tuebingen.de (Marie Hinrichs)

**F:** Datenbasis für Maschinelles Lernen (u.a. Parser)

**S:** Deutsch

**M:** gesprochen

**G:** 200 Millionen Token

**P:** statisch

**A:** Morphosyntax, Chunks, topologische Felder (maschinell)

**SB:** Spezialkorpus: die tageszeitung (taz) 1986-1999

**V:** Lizenz

#### **Videokorpus zur Aphasie**

**Z:** [http://paul.igl.uni-freiburg.de/auer/?Forschung:Abgeschlossene\\_Drittmittelprojekte:Aphasie](http://paul.igl.uni-freiburg.de/auer/?Forschung:Abgeschlossene_Drittmittelprojekte:Aphasie) peter.auer@germanistik.uni-freiburg.de (Peter Auer)

**F:** Klinische Untersuchungen zur Aphasie

**S:** Deutsch

**M:** multimodal

**P:** statisch

**A:** Transkription, Digitalisierung, Alignierung von Text mit Videospur

**SB:** Spezialkorpus: Videoaufnahmen zur Familieninteraktion mit Aphasikern; Longitudinalstudie, in der 10 Familien über einen Zeitraum von einem Jahr nach Entlassung des Aphasikers aus der Klinik beobachtet wurden

**V:** nur am IGL in Freiburg (nach Anfrage)

#### **Vineta-Korpus**

**Z:** <http://archiv.ub.uni-marburg.de/sonst/2000/0003/>, stefan.rabanus@univr.it (Stefan Rabanus)

**F:** Untersuchung von intonatorischen Verfahren

**S:** Deutsch, Italienisch  
**M:** gesprochen  
**G:** 45 Minuten  
**P:** statisch  
**A:** intonatorisch (ToBI, Grundwertfrequenzen)  
**SB:** Spezialkorpus spontaner gesprochener Sprache  
**V:** online

#### **Wendekorpus Mannheim**

**Z:** <http://www.ids-mannheim.de/kt/projekte/korpora/archiv/wk.html>,  
korpusinguistik@ids-mannheim.de (Institut für Deutsche Sprache)  
**F:** Teil der IDS-Korpora  
**S:** Deutsch  
**M:** gesprochen, geschrieben  
**G:** 3,3 Millionen Token  
**P:** statisch  
**A:** Textstruktur  
**SB:** Spezialkorpus: diverse Textgenres aus DDR und BRD 1989-1990  
**V:** online über COSMAS

#### **Wendekorpus Berlin**

**Z:** <http://www.dwds.de/ressourcen/korpora/>, <http://www.dwds.de/kontakt/> (Berlin-Brandenburgische Akademie der Wissenschaften)  
**F:** Interviews aus der Wendezeit, ausgewogen zwischen Ost- und Westberliner Bürgern, Teil der Korpusbasis für das Digitale Wörterbuch der Deutschen Sprache  
**S:** Deutsch  
**M:** gesprochen  
**G:** 77 Interviews  
**P:** statisch  
**A:** Transkription, Lemma, Morphosyntax  
**SB:** Spezialkorpus  
**V:** online abfragbar unter <http://retro.dwds.de/> (Stand 11/2015; zukünftig auch über: <http://www.dwds.de/>)

#### **Wortwarte**

**Z:** <http://www.wortwarte.de>, [lothar@lemnitzer.de](mailto:lothar@lemnitzer.de) (Lothar Lemnitzer)  
**F:** Tägliche Extraktion neuer Wörter  
**S:** Deutsch  
**M:** Schriftsprache  
**G:** Pro Tag etwa 1 Millionen Token. Die Worttyp-Listen werden aufgehoben, die eigentlichen Texte nach drei Tagen gelöscht.  
**P:** Monitorkorpus  
**A:** keine  
**SB:** Tägliche Auswertung von 10 deutschen Tages- und Wochenzeitungen. Wird mit der Wortliste eines Referenzkorpus des Deutschen abgeglichen.  
**V:** online; Worttypen-Listen aller gesampelten Tage sind auf Anfrage verfügbar