Textkorpora und Wortlisten

Hier können Verweise auf interessante Korpora abgelegt werden.

Exklusive Nutzungsrechte

Mannheimer Liste

Urheber	Korpus des Instituts für Deutsche Sprache (IDS)
Rechte	*Darf in ihrer Gesamtheit - wie vereinbart - nicht veröffentlicht oder an Dritte weitergegeben werden.* Abgeleitete Werke können nach unserer Wahl behandelt werden.
Wortformen	4.000.000
Sortierkriterium	Häufigkeitsklassen
Rechtschreibung	mittel
Zugriff	für ausgewählte Personen
Stand	9. 10. 2007

Frei

Werners Liste

Urheber	Werner Lemberg
Rechte	MIT
Wortformen	500.000
Sortierkriterium	alphabetisch
Rechtschreibung	gut
Bemerkung	manuell gepflegt
Zugriff	Git-Repositorium
Stand	26. 2. 2021

Leipziger Liste

Urheber	Liste des Wortschatzprojekts der Universität Leipzig
Rechte	GPL (?)
Wortformen	2.000.000
Sortierkriterium	Häufigkeit
Rechtschreibung	mangelhaft
Bemerkung	automatische Internetsuche (Datenbanken, Zeitungsarchive usw.)
Zugriff	
Stand	28. 3. 2008

#GoogleBooksKorpus

Google-Books-Korpus

Urheber	Google
Rechte	Creative Commons Attribution 3.0 Unported
Wortformen	3.700.000
Sortierkriterium	dateiweise alphabetisch (nicht dateiübergreifend)
Rechtschreibung	mit typischen OCR-Fehlern
Bemerkung	angekündigt am 12. 5. 2011 auf der Corpora-Mailing-Liste
Zugriff	Google Labs: googlebooks-ger-all-20090715
Stand	1. 7. 2011

Google-Books-Liste

Urheber	Stephan Hennig
Rechte	GPL
Wortformen	3.700.000
Sortierkriterium	Häufigkeitsklassen
Rechtschreibung	mit typischen OCR-Fehlern
Bemerkung	abgeleitet aus dem Google-Books-Korpus
Zugriff	Google-Books-Liste
Stand	3. 9. 2011

Korpus der deutschen Wikipedia

Urheber	Roozbeh Pournader, Wikipedia-Autoren
Rechte	CC-BY-SA
Wortformen	ca. 14.000.000
Sortierkriterium	Häufigkeit
Rechtschreibung	mangelhaft
Bemerkung	angekündigt am 3. 7. 2012 auf der http://lists.freedesktop.org/archives/harfbuzz/2012-July/002092.html
Zugriff	z.B. http://www.freedesktop.org/software/harfbuzz/testing/texts/wikipedia/, siehe Ankündigung
Stand	12. 9. 2012

Free German Dictionary

Urheber	Jan Schreiber
Rechte	Public Domain (?)
Wortformen	2.095.000
Sortierkriterium	alphabetisch
Zugriff	http://germandict.sourceforge.net/
Stand	Februar 2021

FreeDict

Urheber	Horst Eyermann u. a.
Rechte	GPL u. a.
Bemerkung	verschiedene zweisprachige Wörterbücher

Zugriff	http://freedict.org/de/
Stand	August 2011

GeoNames

Urheber	verschiedene
Rechte	Creative Commons BY 3.0
Bemerkung	Datenbank mit über 10 Millionen weltweiten geographischen Bezeichnungen, Textdateien als Datenbankdump erhältlich
Zugriff	http://www.geonames.org/
Stand	Mai 2015

Unfrei

DeReWo Wortformenliste

Urheber	IDS Mannheim
Rechte	Creative Commons BY-NC 3.0
Wortformen	100.000
Sortierkriterium	Häufigkeitsklassen
Rechtschreibung	gut; abgeleitet aus dem Deutschen Referenzkorpus
Bemerkung	siehe Mannheimer Liste
Zugriff	DeReWo
Stand	August 2011

DGT-TM (Mehrsprachiger, paralleler Korpus zum EU-Recht)

Urheber	Europäische Kommission - Generaldirektion Übersetzung
Rechte	freizügig, nicht OSI kompatibel
Wortformen	deutsch: 8.000.000
Bemerkung	mehrsprachiger Übersetzungsspeicher zum EU-Recht; enthält etwa 1 Million Sätze und ihre Übersetzungen in 24 Sprachen; angekündigt am 18. 9. 2014 auf der Corpora-Mailing-Liste
Zugriff	DGT-Translation Memory
Stand	18. 9. 2014

Korpora.org

Urheber	Universität Duisburg-Essen
Bemerkung	vier verschiedene Korpora: das Bonner Frühneuhochdeutschkorpus, Daten des Projekts Bereitstellung und Pflege von Immanuel Kants Werken in elektronischer Form, das LIMAS-Korpus, die Hypertext-Ausgabe von Gottlob Freges *Grundgesetze der Arithmetik*
Zugriff	http://www.korpora.org/
Stand	August 2011

Microsoft Web N-Gram Service

Urheber	Microsoft
Zugriff	http://web-ngram.research.microsoft.com/info/

Stand	September 2011

Kommerziell

Deutsches Wörterbuch als Text-Datei

Urheber	Reiner Keul EDV-Dienstleistungen
Rechte	kommerziell (ca. 20 Euro)
Wortformen	600.000
Sortierkriterium	alphabetisch
Zugriff	http://www.debuggen.com/
Stand	September 2012

Named Entity Recognition (NER)

Rechte	kommerziell (günstig)
Bemerkung	enthält Orts- und Personennamen aus der *Frankfurter Rundschau*
Zugriff	http://www.cnts.ua.ac.be/conll2003/ner/
Stand	August 2011

beliebteste Vornamen

Urheber	Gesellschaft für deutsche Sprache
Rechte	kommerziell (günstig)
Bemerkung	Liste von jeweils 200 Mädchen- und Jungennamen, die jährlich in Deutschland am häufigsten vergeben wurden (seit 2004)
Zugriff	http://www.gfds.de/vornamen/beliebteste-vornamen/
Stand	August 2011

Ungeklärte Nutzungsrechte

Berliner Liste

Urheber	Kernkorpus des Projekts Digitales Wörterbuch der Deutschen Sprache (DWDS)
Wortformen	2.000.000
Bemerkung	repräsentativer Wortschatz der deutschen Sprache
Stand	Juni 2009

German Political Speeches Corpus

Urheber	Adrien Barbaresi
Rechte	Fragwürdig. Angeblich gemeinfrei nach § 48 UrhG. Das UrhG bezieht sich jedoch nur auf den Wortlaut von Reden, nicht auf deren Digitalisate (Vorlagen). Für diese besteht weiterhin Urheberrechtschutz. E-Mail-Kontakt besteht.
Bemerkung	enthält Reden deutscher Bundenpräsidenten und Bundeskanzler (Kopien aus dem Web-Angebot des Bundespräsidialamtes)
Zugriff	http://purl.org/corpus/german-speeches (Weiterleitung auf http://perso.ens-lyon.fr/adrien.barbaresi/corpora/index.html)

Stand	