Technical Report Series on Corpus Building

Vol. 1

(February 2013)

Deutscher Wortschatz 2012

Uwe Quasthoff
Dirk Goldhahn
Gerhard Heyer

Abteilung Automatische Sprachverarbeitung, Institut für Informatik,
Universität Leipzig

Affiliation oft the authors:

Institut für Informatik,Universität Leipzig;
{quasthoff, dgoldhahn, heyer}@informatik.uni-leipzig.de

Technical Report Series on Corpus Building

Vol. 1: Deutscher Wortschatz 2013
Vol. 2: Danish Corpora

Inhalt

Wortschatz 2012

Text als Rohstoff

Seit den Anfängen der Schriftsprache in unserem Kulturkreis dient Text dazu, Wissen festzuhalten, zu bearbeiten und weiterzugeben. Text begegnet uns deshalb auf vielfältige Weise in Form klassischer Printprodukte: z. B. in Büchern, Zeitschriften, Zeitungen, technischen Dokumentationen, Benutzerhandbüchern und Produktbeschreibungen, Normen, Gesetzen, Kommentaren und Verträgen, Organisationsanweisungen und Korrespondenzen, um nur einige zu nennen. In digitalisierter Form stellen diese Texte und Textkollektionen als große digitale Bibliotheken, wie sie derzeit entstehen, eine bedeutende Wissensressource dar. Daneben finden sich im Internet aber zunehmend auch Texte, die nicht nur das Ergebnis einer Übertragung vom Papier ins Digitale darstellen, sondern ihre Entstehung in Form und Inhalt dem digitalen Medium selber verdanken. Hierzu zählen insbesondere Informationsseiten im Web, Firmen- und Produktpräsentationen, Foren und Blogs sowie Texte aus sozialen Netzwerken wie Facebook oder Twitter.

Mit dem Internet, den Intranets, E-Mail, Groupware, Social Networks und anderen Systemen steht mehr Information in Form von Texten in natürlicher Sprache zur Verfügung als jemals zuvor. Diese wachsenden Berge an textbasierter Information enthalten geistige Vermögenswerte, deren Nutzung zum eigenen Vorteil im Wettbewerb der Ideen immer wichtiger wird. Die Informationsnutzer wünschen einen direkten Zugang zu relevanter Information, um den Informationsinhalt schnell zu erkennen und zu erfassen sowie neue Ideen und Zusammenhänge zu entdecken. Diese Anforderungen haben insbesondere zur Entwicklung von Systemen zur Auswertung von Informationen in Textform geführt, die allgemein unter dem Stichwort Text Mining zusammengefasst werden. In der Praxis besteht ein enormer Bedarf an der Verwaltung, Strukturierung und Analyse großer Mengen von Text und anderen Informationen, die Wirtschaft und Forschung antreiben.

Ein weiterer Anwendungsbereich ist die Lexikographie. Die vorliegenden großen Textmengen liefern natürlich auch Informationen über die deutsche Sprache und ihren Gebrauch. Interessant ist der deutliche Zusammenhang zwischen der Häufigkeit eines Wortes und seiner Bekanntheit. Damit lassen sich beispielsweise Stichwörter, die in ein Wörterbuch aufgenommen werden sollen, sinnvoll auswählen. Während vorher diese Auswahl auf der Intuition des Lexikographen beruhte, stehen jetzt objektive Auswahlkriterien zur Verfügung. Auf Grund der großen Datenmenge lassen sich auch seltene sprachliche Phänomene untersuchen. Das statistische auffällige gemeinsame Auftreten von Wörtern, sog. Kookkurrenzen, zeigen meist interessante inhaltliche oder strukturelle Zusammenhänge. Der Vergleich der Texte verschiedener Jahre ermöglicht die Identifikation neuer Wörter, sog. Neologismen.

Internet-basierte Textressourcen, wie sie an der Abteilung für Automatische Sprachverarbeitung in den letzten Jahren geschaffen worden sind, bilden für Text Mining und Lexikographie eine unschätzbare Grundlage. Zum einen stellen sie einen digitalen Rohstoff Text bereit, aus dem Informationen und Wissen extrahiert werden können. Zum anderen ermöglichen sie als Referenzwortschatz den Vergleich mit anderen Texten und darauf aufbauend die Anpassung von allgemeinen Text-Mining-Verfahren durch die Berücksichtigung von sprachstatistischen und musterbasierten Besonderheiten.

Daten für den Wortschatz 2012

Im Folgenden werden die Daten beschrieben, welche seit November 2012 unter http://wortschatz.uni-leipzig.de/ abgefragt werden können. Das zugrundeliegende Korpus besteht aus reichlich 26 Millionen Sätzen mit mehr als 400 Millionen laufenden Wörtern. Mehr Zahlenangaben zum Korpus enthält **Anhang: Zusammenfassung**.

Die interne Bezeichnung der Datenbank ist deu_newscrawl_2011. Eine Übersicht über früher verwendete Daten sowie die Zugriffsmöglichkeiten darauf finden sich an Ende dieses Dokuments.

Crawlingzeit

Das Crawling der Webseiten, die die Basis für dieses Korpus darstellen, erfolgte im Januar und Februar des Jahres 2011.

Beschriebener Zeitraum

Genaue Angaben über das Veröffentlichungsdatum von Webseiten lassen sich im Normalfall nicht gewinnen. Ob ein Dokument kurz vor dem Termin des Herunterladens erstellt wurde oder schon mehrere Jahre auf dem Webserver existierte, ist nicht ohne Weiteres zu klären. Alternativ ist es jedoch möglich, das Vorkommen von Jahreszahlen zu untersuchen. Dies ermöglicht Rückschlüsse auf den Zeitraum, auf den sich das Dokument bezieht. Da es sich um Nachrichtentexte handelt und diese häufig aktuellen Bezug haben, liegen gefundene Jahreszahlen und das Jahr der Veröffentlichung oft nahe beieinander.

Anhang: Zahlen im Datumsformat (1980-2029) zeigt die Häufigkeit der Jahreszahlen zwischen 1980 und 2029. Der Statistik kann man entnehmen, dass das Jahr 2010 am häufigsten Inhalt der Betrachtungen ist, gefolgt von 2009 und 2008. Auffällig hohe Werte treten für die Zahl 2000 auf, aber auch für die Jahreszahlen 1989 und 1990, einen Zeitraum, in den der Prozess der deutschen Wiedervereinigung fällt. Auch Betrachtungen über zukünftige Entwicklungen sind Inhalt der Dokumente. Am auffälligsten ist hier das Jahr 2020.

Größte Quellen

Zur Erstellung des Korpus wurden verschiedene Domains von Nachrichtenanbietern gecrawlt.

Anhang: Größe der umfangreichsten Domains listet die Webseiten auf, deren Texte die meisten Sätze für die spätere Analyse ergaben. Teilweise wurden hier hohe sechsstellige Satzzahlen erreicht. Größte Domain ist www.n24.de.

Anhang: Größe der verschiedenen TLDs hingegen zeigt die Verteilung der Top Level Domains unter den gecrawlten Seiten. Hierbei stellen deutsche Webseiten mit der TLD „.de“ mit einem Wert von fast 65% den Großteil des Textmaterials. Es folgen Texte aus der Schweiz und Österreich.

Häufigkeitsklassen

Zu jedem Wort wird eine Häufigkeitsklasse (engl. frequency class) angegeben. Die Häufigkeitsklasse HK(w) beschreibt die Häufigkeit f(w) eines Wortes w im Vergleich zur Häufigkeit des häufigsten Wortes fmax und ist wie folgt definiert: HK(w)=[log[2](fmax/f(w))]. Dabei wird auf den nächstgelegenen ganzzahligen Wert gerundet.

Das häufigste Wort im Korpus ist *der* und somit dient dessen Häufigkeit als Referenzwert fmax. Damit gehört *der* zur Häufikeitsklasse 0. Außer *der* sind noch die Wörter *die* und *und* in dieser Häufigkeitsklasse. Eine Erhöhung der Häufigkeitsklasse um 1 entspricht näherungsweise einer Halbierung der Häufigkeit. Das Wort *der* ist somit ungefähr 32 mal so häufig wie das Wort *alle* aus der Klasse 5.

Anhang: Größe der Frequenzklassen listet den Umfang der einzelnen Frequenzklassen des Korpus auf.

Die 50 häufigsten Wörter

Anhang: Die 50 häufigsten Wörter zeigt die 50 Wörter mit den größten Vorkommenszahlen. Unter diesen Wörtern finden sich ausschließlich Funktionswörter wie Artikel, Präpositionen, Konjunktionen, Hilfsverben usw.

Längste häufige Wörter

Anhang: Längste Wörter in den Top-1000 geordnet nach Länge zeigt die 50 längsten unter den 1000 häufigsten Wörtern. Zusätzlich wird zu jedem der Wörter dessen Rang, also seine Position in der häufigkeitssortierten Wortliste angegeben. An diesen Wörtern kann man wichtige Themen der Texte erkennen.

Wortlänge

Für die Wörter wurde die Verteilung der Wortlänge in Buchstaben ermittelt.

Anhang: Wortlänge ohne Wiederholungen zeigt die Verteilung der Wortlänge, wenn jedes Wort nur einmal berücksichtigt wird. Die Grafik zeigt, dass in der gesamten Wortliste der Anteil der Wörter bestehend aus 13 Buchstaben mit reichlich 8% am größten ist.

Betrachtet man hingegen jedes Wort mit der Häufigkeit, mit der es im Text verwendet wird, ergibt sich die Wortlängenverteilung in **Anhang zu: Wortlänge mit Wiederholungen**. Hier haben kürzere Wörter eine größere Häufigkeit Die dreibuchstabigen Wörter besitzen mit rund 25% den größten Anteil. Grund dafür ist die häufige Verwendung der dreibuchstabigen Artikel und Präpositionen.

Satzlänge

Für alle Sätze wurde die Satzlänge in Wörtern bestimmt. **Anhang: Satzlänge in Wörtern** stellt die Verteilung der Satzlängen in grafischer Form dar.

Zusätzlich werden in **Anhang: Satzlänge in Buchstaben** Informationen zur Satzlänge in Buchstaben zusammengetragen.

Kookkurrenzen

Unter Kookkurrenzen versteht man ein Paar von Wörtern, die statistisch signifikant häufig gemeinsam auftreten. Dies liegt normalerweise darin begründet, dass diese Wörter sich inhaltlich oder funktional ergänzen. Wird das gemeinsame Auftreten in jeweils einem Satz betrachtet, so sprechen wir von Satzkookkurrenzen. Interessant ist auch das Auftreten als unmittelbare Nachbarn. In diesem Falle sprechen wir von Nachbarschaftskookkurrenzen.

Zur Bestimmung der Stärke einer Kookkurrenz bestimmen wir deren Signifikanz. Hierfür nutzen wir das Log-Likelihood-Ratio. Dieses Maß liefert einen umso höheren Wert, je mehr die Wahrscheinlichkeit des gemeinsamen Auftretens vom Erwartungswert unter der Annahme der statistischen Unabhängigkeit nach oben abweicht.

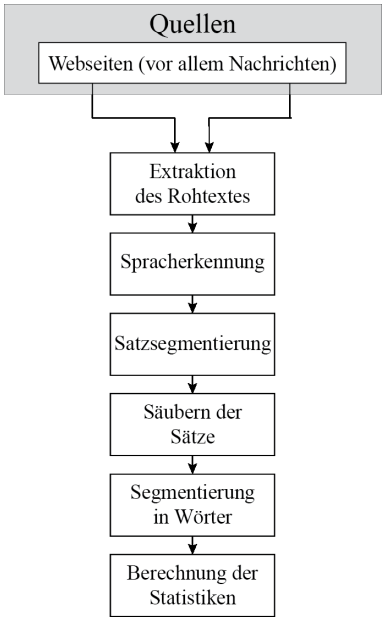
Anhang: Signifikanteste Nachbarschaftskookkurrenzen zeigt die stärksten Nachbarschaftskookkurrenzen unter den 10.000 häufigsten Wörtern. Angeführt wird die Liste vom Wortpaar *in der*.

Anhang: Signifikanteste Satzkookkurrenzen wiederum beinhaltet die stärksten Satzkookkurrenzen unter den 10.000 häufigsten Wörtern. Die stärkste Signifikanz tritt hier zwischen *gibt* und *es* auf.

Mehr statistische Auswertungen für diese und andere Korpora finden sich auf den Seiten zur Sprachstatistik unter http://cls.informatik.uni-leipzig.de/(in Englisch).

Datenaufbereitung für Wortschatz-Datenbanken

Der Prozess der Korpuserstellung umfasst mehrere Einzelschritte. Die Texte müssen beschafft, verarbeitet und analysiert werden. Wichtige Etappen dieses Prozesses werden im Folgenden näher beschrieben. Zur Übersicht soll die folgende Grafik dienen.



Crawling

Vor dem Herunterladen wird eine Liste der zu berücksichtigenden Websites benötigt. Diese Domainnamen werden dem Verzeichnis http://www.abyznewslinks.com entnommen, welche Nachrichtenquellen vieler Länder sortiert nach Sprachen auflistet. Zum Herunterladen dieser kompletten Nachrichtenwebseiten wird das Tool HTTrack (http://www.httrack.com/) genutzt. In Abhängigkeit vom Aufbau der jeweiligen Nachrichtenwebseite und den Fähigkeiten von HTTrack werden je nach Domain nur aktuelle Artikel oder zusätzlich ganze Archive geladen.

Extraktion der Rohtexte - HTML-Stripping

Der nächste Verarbeitungsschritt umfasst die Extraktion des Textes aus den heruntergeladenen HTML-Dokumenten. Dazu wird der an unserer Abteilung entwickelte HTML-Stripper Html2Text eingesetzt. Ergebnis sind die mit Quellenangaben versehenen Rohtexte.

Sprachidentifikation

Da man sich nicht immer darauf verlassen kann, dass als deutschsprachig gekennzeichnete Webseiten tatsächlich Text in deutscher Sprache enthalten, werden die Einzeldokumente noch einmal überprüft. Dafür wird der Sprachidentifizierer LangSepa unserer Abteilung genutzt. Dieses Tool vergleicht die Verteilung von Stoppwörtern sowie Buchstaben-N-Grammen vieler Sprachen mit der im Dokument vorhandenen Verteilung und bestimmt die wahrscheinlichste Sprache des Dokumentes. Dokumente, die dabei nicht der deutschen Sprache zugeordnet werden, werden entfernt.

Satzsegmentierung

Die Dokumente wurden nun in Sätze zerlegt. Hierzu wurden eine Liste mit typischen Satzendezeichen (Punkt, Ausrufe- und Fragezeichen, diverse Formen von Ausführungszeichen) verwendet. Eine zusätzliche Abkürzungsliste hilft, Punkte in Abkürzungen von Satzenden zu unterscheiden.

Säubern

Um die Qualität der Daten zu erhöhen, folgt eine Säuberung nach der Zerlegung der Texte in Sätze. Um nicht-wohlgeformte Sätze ohne eine grammatische Analyse zu erkennen, werden musterbasierte Verfahren eingesetzt. Beispielsweise wird die Menge der erlaubten Zeichen eingeschränkt, die Anzahl von Ziffern und Sonderzeichen pro Satz wird beschränkt, ebenso die Gesamtlänge eines Satzes. Außerdem wird für jeden Satz noch einmal die Sprache überprüft, nicht deutschsprachige Sätze werden aussortiert. Abschließend werden mehrfach vorkommende Sätze entfernt.

Tokenisierung

Bei der folgenden sog. Tokenisierung werden die Sätze in Wörter zerlegt. Die Zerlegung der Sätze erfolgt nicht nur an Leerzeichen, zusätzlich müssen auch Satzzeichen von den Wörtern abgetrennt werden. Dies ist die Voraussetzung für Statistiken auf Wortebene.

Wortstatistiken

Für jedes Wort wird die Häufigkeit seines Auftretens bestimmt. Darüber hinaus werden Satz- und Nachbarschaftskookkurrenzen ermittelt: Das sind diejenigen Paare von Wörtern, die statistisch auffällig häufig zusammen in einem Satz oder unmittelbar nebeneinander vorkommen. Der Grund für ein solches auffälliges gemeinsames Auftreten liegt meist in einem inhaltlichen oder grammatikalischen Zusammenhang dieser Wörter und zeigt die typische Verwendung.

Wörterbuch-Daten zur Deutschen Sprache

Wörterbuch

Die Wörterbuchdaten stammen aus verschiedenen Quellen. Daten wurden von Verlagen zur Verwendung im Wortschatz-Projekt zur Verfügung gestellt oder aus freien Quellen übernommen. Soweit möglich wurden diese Angaben mit automatischen Lernverfahren für weitere Wörter erzeugt.

Eine Übersicht über die wichtigsten Wörterbuchangaben und deren Umfang gibt die folgende Tabelle:

Angabe	Anzahl Wörter mit dieser Angabe
Grammatik	750.000 Millionen
Grundform	1.0 Millionen
Morphologie	825.000
Sachgebiet	280.000
Beschreibung	74.000
Pragmatik	30.000
semantische Relationen	130.000

Wortgruppen

Ausgewählte Wortgruppen werden wie Einzelwörter behandelt: Sie können mit der Suchmaske gesucht werden; man erhält Häufigkeitsangaben, Beispielsätze und Kookkurrenzen. Die Liste der Wortgruppen enthält folgende Arten von Einträge

Art der Wortgruppe	Anzahl	Beschreibung
Wikipedia-Titel	147.000	Titel von Wikipedia-Einträgen, die aus mehreren Wörtern bestehen. Häufig Personennamen (*Max Planck*, *Karl der Große*), aber auch geographische Namen (*Sri Lanka*) und andere Eigennamen (*Eurovision Song Contest*)
Suchformen für Phraseologismen	6.000	Entweder komplette Phraseologismen oder unveränderliche Teile davon, z.B. *Schritt für Schritt*, *Saus und Braus*, *dumm und dämlich*.
sonstige	172.000	Datumsangaben (*1. 4.*, *1. April*), nach alter Rechtschreibung zusammengeschriebene Wörter (*entgegen wirken*), Anglizismen (*Car Sharing*, *Joint Venture*) und viele andere; teilweise auch automatisch erzeugt.
Gesamt	**325.000**	

Download und Nutzungsbedingungen für Wortschatz-Datenbanken

Daten zum Download

Kleinere Versionen des Korpus stehen sowohl als reine Textdateien als auch als MySQL-Datenbanken zur Verfügung. Dazu wurden aus der Gesamtmenge aller Sätze zufällig 10.000, 30.000, 100.000, 300.000 bzw. 1 Millionen Sätze ausgewählt und entsprechend weiterverarbeitet. Abgeleitet vom internen Namen deu_nescrawl_2011 tragen diese die Namen deu_nescrawl_2011_10K, deu_nescrawl_2011_30K usw. bis deu_nescrawl_2011_1M.

Diese Daten stehen zum Download zur Verfügung unter http://corpora.informatik.uni-leipzig.de/download.html.

Nutzungsbedingungen

Sämtliche zum Download angebotenen Daten des Projekts *Deutscher Wortschatz* unterliegen der Creative Commons Lizenz cc-by Vers. 3.0.

Dies bedeutet,

- Nutzer können die Daten vervielfältigen, verbreiten und öffentlich zugänglich machen;
- Abwandlungen und Bearbeitungen der Daten anfertigen und
- die Daten kommerziell nutzen

unter der Bedingung, daß die Rechteinhaber auf folgende Weise genannt werden:

© Abt. Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig, 2012.

Ältere Versionen

Da diese Daten in unregelmäßigen Abständen aktualisiert werden, müssen die aus diesem Korpus statistisch ermittelten Daten nicht mit Daten aus älteren Versionen übereinstimmen. Die folgende Tabelle beschreibt die älteren Versionen. Alle diese Daten sind weiter verfügbar unter http:// wortschatz. uni-leipzig. de/ ws_norm/ index_wm. php

Zeitraum	Name des Korus	Anzahl Sätze	Beschreibung
2000 - 2003	deu_mixed_2000	15 Millionen	hauptsächlich Zeitungstext, erschienen 1995-2000; auch einige Handbücher und Literatur aus dem *Projekt Gutenberg*.
2003 - 2005	deu_news_1995-2003	35 Millionen	Zeitungstext, erschienen 1995-2003
2006 - 2012	deu_news_2005-2006	10 Millionen	Zeitungstext, erschienen 2005/2006
ab 2013	deu_newscrawl_2011	26 Millionen	Zeitungstext, Crawling 2011

Mitwirkende am Wortschatz 2012

Am Crawling und an der Korpuserstellung wirkten mit:

- Thomas Eckart,
- Dirk Goldhahn,
- Christoph Kuras,
- Uwe Quasthoff.

Die Software zum Crawling und zur Korpuserstellung wurde hauptsächlich entwickelt von:

- Volker Boehlke (Satzsegmentierung),
- Marco Büchler (Datenaufbereitung, Berechnung der Kookkurrenzen),
- Thomas Eckart (Prozesskette, Datenaufbereitung),
- Dirk Goldhahn (Crawling, Datenaufbereitung),
- Mao Nie (Crawling),
- Johannes Pollmächer (Sprachidentifikation),
- Fabian Schmidt (Web-Portal)
- Sven Teresniak (Sprachidentifikation).

An älteren Versionen haben mitgewirkt:

Chris Biemann, Karsten Böhm, Stefan Bordag, Chun Cui, Erla Hallsteinsdóttir, Martin Läuter, Robert Remus, Matthias Richter, Sergej Vintgolc, Christian Wolff u.a.

Literatur zum Deutschen Wortschatz

Literatur

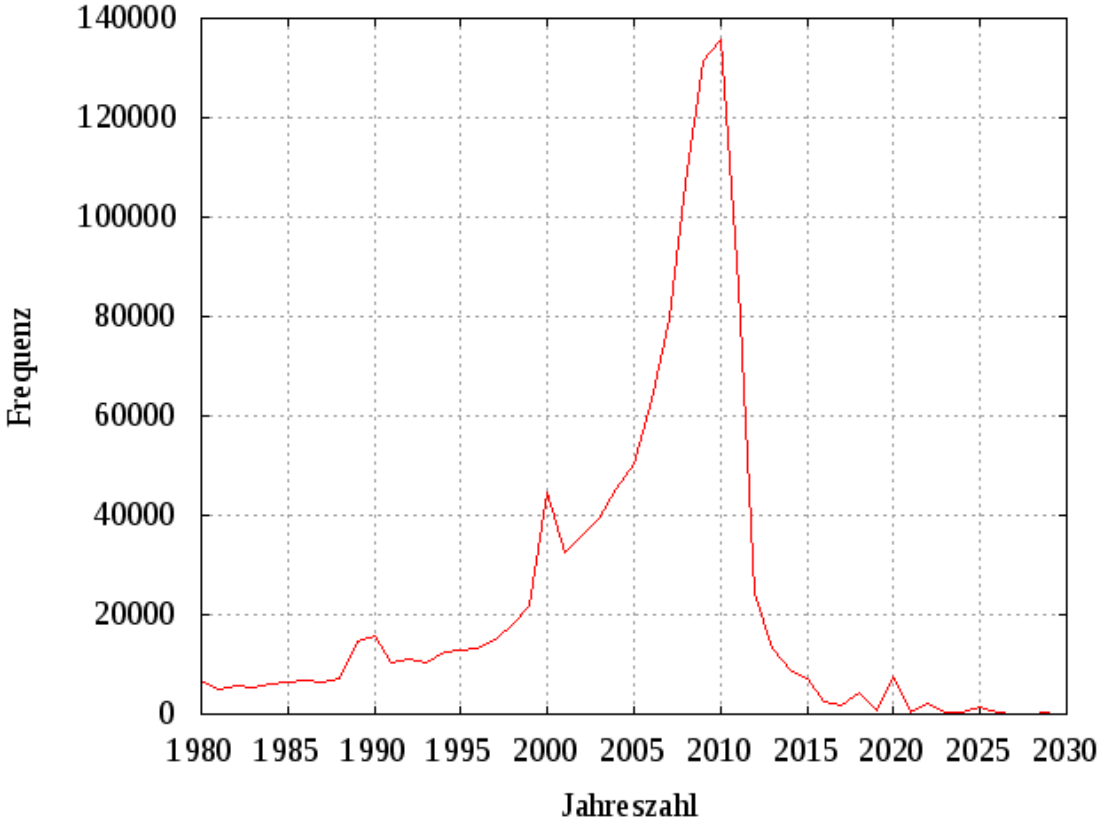
Die nachfolgende Liste enthält Literatur, in der die Korpuserstellung genauer beschrieben ist sowie Wörterbücher, welche mit Hilfe der Wortschatz-Daten erstellt wurden.

- Dornseiff, F.: *Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearb. u. mit einem vollständigen alphabetischen Zugriffsregister versehene Aufl. von Uwe Quasthoff. Mit einer lexikographisch-historischen Einführung und einer Bibliographie von Herbert Ernst Wiegand. Berlin. New York 2004.
- Goldhahn, D., Eckart, T., Quasthoff, U.: *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages*, In: Proceedings of the 8th Language Resources and Evaluation Conference (LREC) 2012
- Heyer, G., Quasthoff, U. und Wittig, Th.: *Wissensrohstoff Text; Text Mining: Konzepte, Algorithmen, Ergebnisse*. w3l-Verlag, Bochum, 2006; dritter Nachdruck 2011.
- Quasthoff, U.: *Deutsches Neologismenwörterbuch: Neue Wörter und Wortbedeutungen in der Gegenwartssprache*, Verlag Walter de Gruyter, Berlin, New York 2007
- Quasthoff, U.: *Wörterbuch der Kollokationen im Deutschen*, Verlag Walter de Gruyter, Berlin, New York 2011
- Quasthoff, U., Fiedler, S. und Hallsteinsdóttir, E. (ed.): *Frequency Dictionaries Vol. 1: Frequency Dictionary German*, Leipziger Universitätsverlag, Leipzig 2011

Anhänge

Anhang zu deu_newscrawl_2011: Zahlen im Datumsformat (1980-2029)

Häufigkeit der Zahlen von 1980 bis 2029



Anhang zu deu_newscrawl_2011: Größe der umfangreichsten Domains

Quelle	# Sätze
www.n24.de/	910582
www.pressetext.at/	743690
www.news.ch/	743331
www.20min.ch/	694800
www.az.com.na/	494969
www.oe-journal.at/	481613
www.beobachter.ch/	437028
www.bernerzeitung.ch/	435313
www.noz.de/	427833
www.gea.de/	424781
www.freitag.de/	377290
www.goettinger-tageblatt.de/	376030
www.nachrichten.at/	365354
www.general-anzeiger-bonn.de/	357452
www.haz.de/	335932
www.fr-online.de/	325083
www.net-news-global.de/	310651
www.tagesanzeiger.ch/	305133
www.nzz.ch/	297058
www.pattayablatt.com/	295899
www.handelszeitung.ch/	289742
www.epochtimes.de/	286284
www.noows.de/	283120
www.sauerlandkurier.de/	280921
www.derbund.ch/	271218

Anhang zu deu_newscrawl_2011: Größe der verschiedenen TLDs

Top-Level-Domains mit einem Anteil von mehr als 1%

TLD	# Quellen	%
.de/	1297886	64.81
.ch/	283924	14.18
.at/	216928	10.83
com/	63438	3.17
.na/	41390	2.07
.ru/	31709	1.58
.lu/	22872	1.14

Anhang zu deu_newscrawl_2011: Die 50 häufigsten Wörter

Rang	Wort	Rang	Wort
1	der	26	werden
2	die	27	am
3	und	28	aus
4	in	29	dass
5	den	30	Der
6	zu	31	sie
7	von	32	bei
8	mit	33	wird
9	das	34	sind
10	ist	35	nach
11	sich	36	er
12	im	37	Das
13	auf	38	um
14	für	39	noch
15	Die	40	wie
16	nicht	41	einer
17	ein	42	einem
18	dem	43	einen
19	des	44	vor

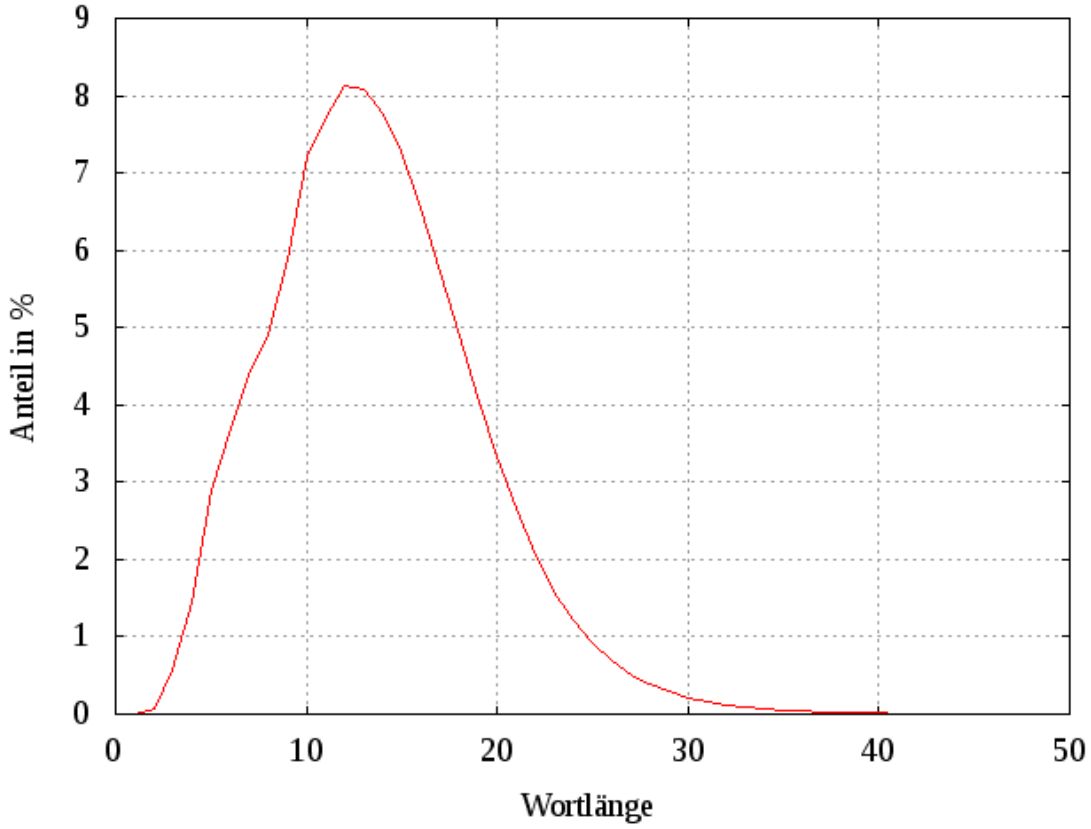
20	eine	45	haben
21	es	46	war
22	auch	47	über
23	an	48	zum
24	als	49	so
25	hat	50	aber

Anhang zu deu_newscrawl_2011: Längste Wörter in den Top-1000 geordnet nach Länge

Rang	Rang in der Wortliste	Wort	Länge
1	804	internationalen	15
2	781	beispielsweise	14
3	778	veröffentlicht	14
4	767	Zusammenarbeit	14
5	750	verschiedenen	13
6	525	Bürgermeister	13
7	687	Unterstützung	13
8	563	Informationen	13
9	966	europäischen	12
10	911	öffentlichen	12
11	904	mittlerweile	12
12	997	Jugendlichen	12
13	629	Gesellschaft	12
14	597	Entscheidung	12
15	953	erfolgreich	11
16	810	tatsächlich	11
17	807	politischen	11
18	663	Bevölkerung	11
19	460	Entwicklung	11
20	214	vergangenen	11

Anhang zu deu_newscrawl_2011: Wortlänge ohne Wiederholungen

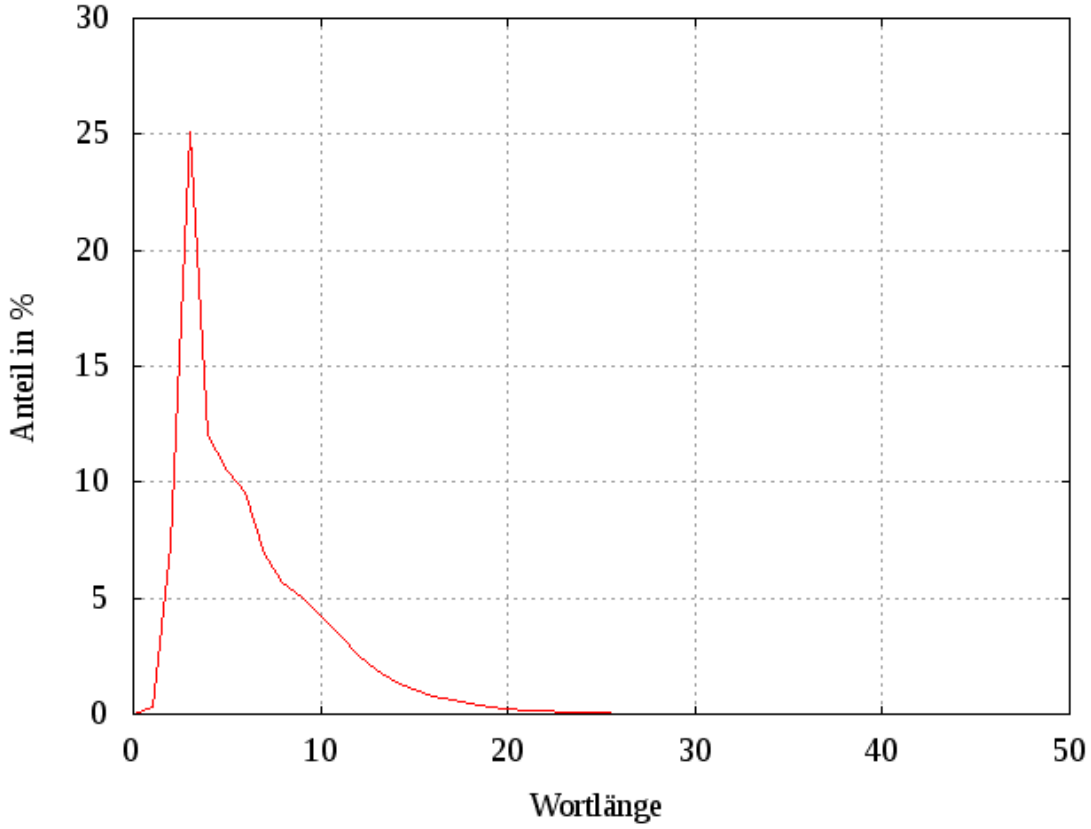
Anzahl Wörter bestimmter Länge in der Wortliste, d.h. gezählt ohne Wiederholungen



Durchschnittliche Wortlänge
13.5488

Anhang zu deu_newscrawl_2011: Wortlänge mit Wiederholungen

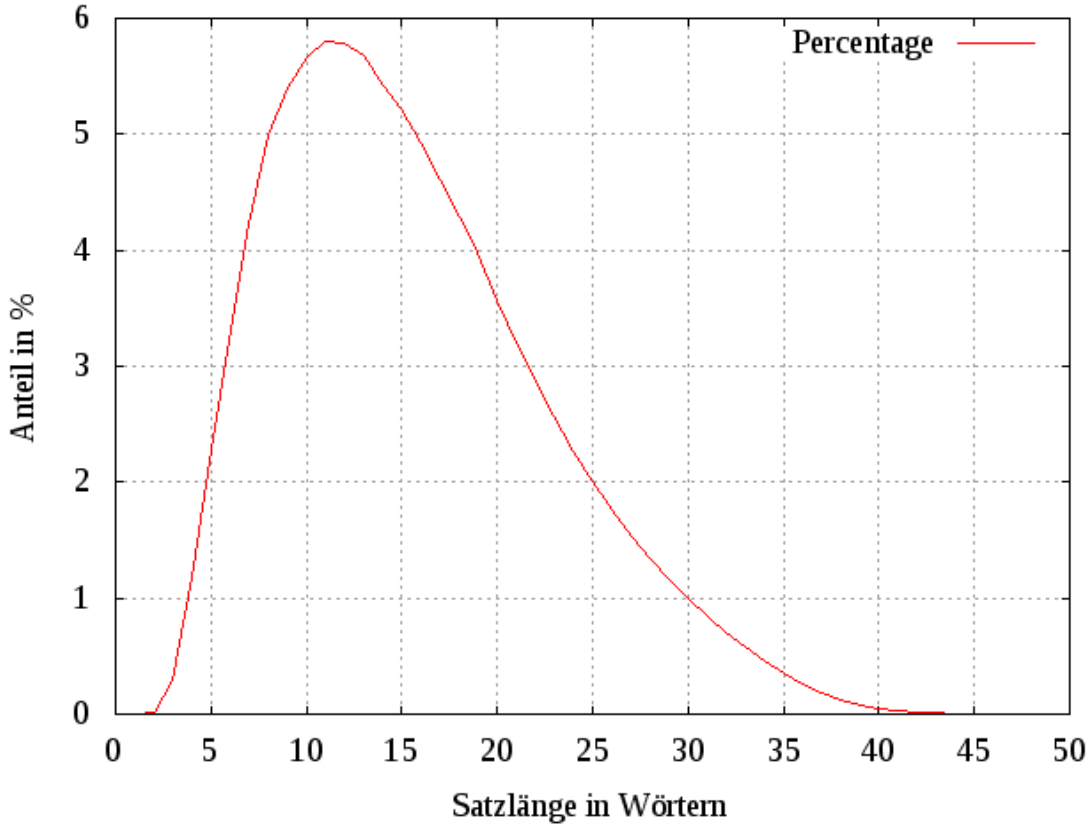
Anzahl Wörter bestimmter Länge im Text, d.h. gezählt mit Wiederholungen



Durchschnittliche Wortlänge
6.1611

Anhang zu deu_newscrawl_2011: Satzlänge in Worten

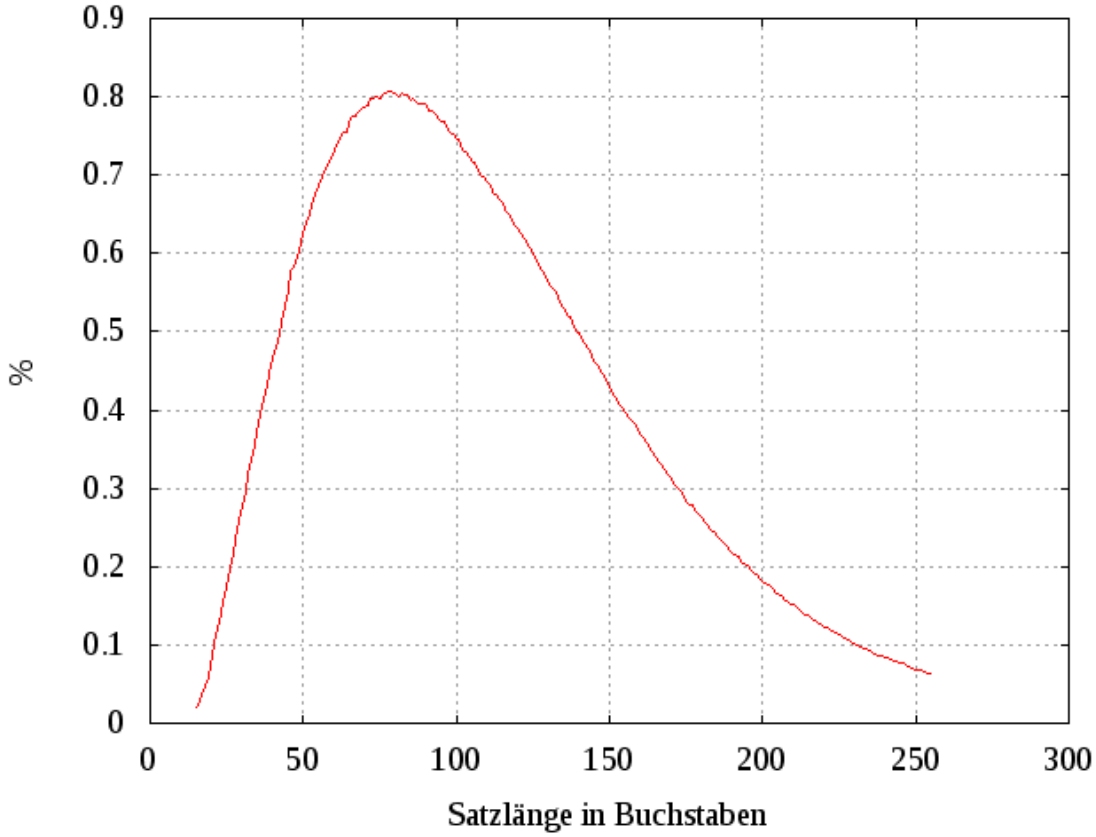
Satzlänge (gemessen in Wörtern): Verteilung und Durchschnittswert



Durchschnittliche Satzlänge
15.5977

Anhang zu deu_newscrawl_2011: Satzlänge in Buchstaben

Satzlänge (gemessen in Buchstaben): Verteilung und Durchschnittswert



Durchschnittliche Satzlänge
108.6338

Standardabweichung
51.0890

Anhang zu deu_newscrawl_2011: Signifikanteste Nachbarschaftskookkurrenzen

Die 25 stärsten Nachbarschaftskookkurrenzen innerhalb der Top-10.000 Wörter

Wort 1	Wort 2	Rank 1	Rank 2	Signifikanz
in	der	4	1	3711933.75
für	die	14	2	2113812.75
gibt	es	84	21	2000495.88
in	den	4	5	1964081.00
mit	dem	8	18	1946782.88
vor	allem	44	164	1941257.75
aus	dem	28	18	1363556.62
nicht	mehr	16	56	1294687.50
auf	dem	13	18	1286648.62
mit	einem	8	42	1135884.50
mehr	als	56	24	1095617.75
unter	anderem	72	410	1017785.12
bei	der	32	1	1004296.62
Millionen	Euro	155	79	1000780.38
für	den	14	5	915097.25
nicht	nur	16	51	872017.44
über	die	47	2	839453.94
an	der	23	1	797593.75
zur	Verfügung	61	594	783548.00
nach	dem	35	18	783282.75
auf	den	13	5	760775.12
Es	ist	66	10	702122.06
New	York	671	1011	700957.50
am	Mittwoch	27	329	691469.75
zum	Beispiel	48	336	687312.56

Anhang zu deu_newscrawl_2011: Signifikanteste Satzkookkurrenzen

Die 25 stärsten Satzkookkurrenzen innerhalb der Top-10.000 Wörter

Wort 1	Wort 2	Rank 1	Rank 2	Signifikanz
gibt	es	84	21	1156363.75
allem	vor	164	44	920568.25
in	der	4	1	738950.00
sondern	nicht	157	16	684292.56
anderem	unter	410	72	560537.38
Millionen	Euro	155	79	474396.62
York	New	1011	671	473453.94
mehr	als	56	24	454434.94
zwischen	und	137	3	434355.75
Verfügung	zur	594	61	432535.28
sondern	nur	157	51	422235.41
dem	mit	18	8	420638.59
Uhr	um	100	38	418156.97
um	zu	38	6	412866.59
seit	Jahren	114	81	406769.12
bin	Ich	263	78	401621.41
bin	ich	263	57	374353.12
habe	ich	73	57	371184.59
findet	statt	374	304	370292.25
kann	man	62	54	363175.50
mich	ich	173	57	334883.25
Sonntag	am	273	27	334713.50
Freitag	am	284	27	332081.50
Mittwoch	am	329	27	331849.53
uns	wir	120	63	328134.44