Wolfgang Klein

Von Reichtum und Armut des deutschen Wortschatzes*

„Und Gewinn und Verlust wäget ein sinniges Haupt“
Hölderlin

Lexik, Grammatik, Ausdrucksreichtum

Die Tauglichkeit einer Sprache bemisst sich letztlich daran, was an Gedanken und Gefühlen, Verboten und Wünschen man mit ihr auszudrücken vermag. Das hängt zum einen davon ab, über welche Ausdrucksmittel die Sprache verfügt, und zum andern davon, wie die Sprecher mit diesen Mitteln umgehen, wenn sie bestimmte Inhalte auszudrücken versuchen. Wenn in der öffentlichen Diskussion vom „Verfall des Deutschen“ die Rede ist, so ist oft nicht die Sprache selbst gemeint, sondern ein bestimmter Gebrauch, der von ihr gemacht wird. Im Folgenden geht es nur um die Möglichkeiten, die das Deutsche seinen Nutzern bereitstellt. Dieses Potential hat in allen menschlichen Sprachen zwei Quellen: Lexik und Grammatik. Mit Lexik ist der Bestand an elementaren Ausdrücken – an Wörtern – gemeint. Mit Grammatik meint man all jene Regeln, nach denen sich aus einfachen Ausdrücken komplexere bilden lassen. Hier unterscheidet man gewöhnlich zwischen wortinternen (morphologischen) Regeln, mit denen man beispielsweise die verschiedenen Flexionsformen bildet (*trankst, trank, getrunken* zu dem Verb *trinken; Vaters, Väter, Vätern* zu dem Nomen *Vater*; *feister, feiste, feisten* zu dem Adjektiv *feist*), und wortübergreifenden (syntaktischen) Regeln, die mehrere Wörter zu einem Satzteil oder einem Satz verbinden. Eine gewisse Zwischenstellung zwischen Lexik und Grammatik nimmt die Wortbildung ein. Sie umfasst all jene Regeln, nach denen man komplexe Wörter aus einfachen bilden kann. Dies sind in erster Linie Ableitungen wie *abstellen, bestellen, einstellen, verstellen, zustellen* von *stellen*, *fällig* von *Fall*, *Schwimmer* von *schwim-*

* Ich danke den beteiligten Akademien, dem Beauftragten der Bundesregierung für Kultur und Medien, der Fritz Thyssen Stiftung und dem Max-Planck-Institut für Psycholinguistik, die dieses Vorhaben getragen haben, Barbara Seelig und ihren studentischen Mitarbeitern, die das Berichtskorpus redigiert haben, sowie Alex Geyken, Julian Heister, Bernhard Ulreich und Kay-Michael Würzner, von denen die meisten Berechnungen stammen. Alex Geyken, Wolf-Hagen Krauth und Lothar Lemnitzer danke ich für vielfältige Hilfe und Diskussionen innerhalb und außerhalb dieses Vorhabens.

men oder Komposita wie *hellgelb, radfahren, Gehhilfe* oder *Rotlauf*; beides ist oft miteinander verbunden, wie in *Vorfälligkeitsentschädigung* oder *Wissenschaftsfreiheitsgesetz*. Die Wortbildungsregeln gehören zur Grammatik, das Ergebnis ist aber wiederum ein Wort und daher ein Teil der Lexik. Solche Wortbildungsregeln finden sich in allen Sprachen, aber im Deutschen wird besonders reicher Gebrauch davon gemacht. Gleichfalls eine Zwischenstellung zwischen Lexik und Grammatik nehmen Ausdrücke ein, die syntaktisch zusammengesetzt sind, aber ihrer Bedeutung nach einem einzelnen Wort entsprechen, wie etwa *zur Welt bringen*. Der Form nach gehören sie in den syntaktischen Teil der Grammatik, der Bedeutung nach ins Wörterbuch: *zur Welt bringen* entspricht ja *gebären* (eine wichtige Klasse solcher Mehrwortwörter wird in dem Beitrag von Storrer in diesem Band behandelt).

Die Linguisten sind nicht in allzu vielen Dingen einer Meinung, aber darüber, dass jede menschliche Sprache – im Gegensatz zu manchen anderen Zeichensystemen – über eine Lexik und eine Grammatik verfügt, gibt es kaum Dissens, auch wenn die Grenzen beider nicht immer völlig klar sind. Beide sind zwar unabdinglich; das Ausdrucksvermögen einer Sprache hängt jedoch vor allem von einer reichen Lexik ab. Wenn jemand 100 000 Wörter des Japanischen perfekt beherrscht, aber keine einzige der für das Japanische spezifischen grammatischen Regeln, dann würde ihm zwar viel entgehen, wenn er in Japan weilt; aber er könnte sich im Alltagsleben gut durchschlagen und eine Zeitung, einen Roman, eine Gebrauchsanleitung einigermaßen verstehen. Wenn jemand hingegen alle morphologischen und syntaktischen Regeln des Japanischen in- und auswendig kennt, aber nur ein Dutzend Wörter, dann würde sich der Nutzen in Grenzen halten. Beide Szenarien sind nicht sehr realistisch, aber sie deuten doch das relative Gewicht von Wörtern und Regeln für die Verständigung an.

Dem entspricht die Art und Weise, in der sich Lexik und Grammatik im Laufe der Zeit verändern. Zwar kommen durchaus Wörter außer Gebrauch (wer weiß noch, was *dalest* bedeutet?[1]). Aber alle Kultursprachen der Welt leben von einer gewaltigen Anreicherung ihres Wortbestandes über die Jahrhunderte, ja, es ist dies vielleicht überhaupt das Definitionsmerkmal einer Kultursprache. Die Grammatik der Kultursprachen wird hingegen im Laufe der Jahre eher einfacher.[2] Im

1 Dem entsprechenden Eintrag im Grimmschen Wörterbuch zufolge ist dieses Wort vom 14. bis zum 16. Jahrhundert belegt. Allerdings war sich Wilhelm Grimm, der diesen Eintrag geschrieben hat, auch nicht ganz klar darüber, was es denn wirklich bedeutet hat. Jedenfalls war es wohl entbehrlich, denn niemand scheint es zu vermissen.

2 Davon gibt es einige Ausnahmen, z. B. die vor 500 Jahren einsetzende Aspektunterscheidung im Englischen (*he walked – he was walking*) oder das synthetische Futur in den romanischen Sprachen (*je viendrai* aus *[ego] venire habeo*); das ist jedoch sehr untypisch.

Germanischen oder Lateinischen, also den Sprachen, die dem Deutschen und Englischen, dem Französischen, Spanischen und Italienischen zugrunde liegen, wird beispielsweise regelhaft ein Unterschied zwischen Nominativ und Akkusativ markiert: *femina – feminam, feminae – feminas* haben wir in der Schule lernen müssen. In den modernen Sprachen, die daraus entstanden sind, ist die Unterscheidung zwischen Nominativ und Akkusativ weitestgehend weggefallen. Das Deutsche hat sie immerhin noch für den Nominativ Singular maskulin bewahrt (*der Löffel – den Löffel*), da allerdings nur selten beim Nomen selbst (*der Bär – den Bären*); niemand kann sagen, ob *die Frau, das Haus, die Männer* Nominativ oder Akkusativ sind. Die wenigsten werden diesen „Verfall" einer grammatischen Markierung als Verlust empfinden, der sie ernsthaft in ihren Ausdruckswünschen beschränkt; sie hat offenbar kaum einen funktionalen Wert für das, was man auf Deutsch, Englisch oder Französisch sagen kann. Die *Kritik der reinen Vernunft* oder *Das Kapital* würden nicht viel an Verständlichkeit gewinnen, wenn ihre gelehrten Verfasser bei den Nomina immer fein zwischen Nominativ und Akkusativ unterschieden hätten. Diese Werke leben aber davon, dass es Wörter wie *Verstandesbegriff* und *transzendental*, *Tauschwert* und *ausbeuten* gibt, und es wäre schwer möglich, sie in eine Sprache zu übersetzen, die zwar eine wundervoll regelreiche Grammatik hat, aber kein Gegenstück zu diesen und zahllosen anderen Wörtern, die wir zur Lexik des Deutschen rechnen. Oder um ein anderes Beispiel zu geben: So liebenswert es ist, wenn das Deutsche nach altem Brauch zwischen *der Löffel, die Gabel, das Messer* oder zwischen *der silberne Löffel, ein silberner Löffel, (mit) silbernem Löffel* unterscheidet – niemand kann allen Ernstes behaupten, dass diese Vielgestaltigkeit der ererbten Flexionsformen nennenswert zum Ausdrucksreichtum unserer Sprache beiträgt. Fülle und Differenziertheit des Wortschatzes sind es, in denen sich die geistige und kulturelle Entwicklung einer Gemeinschaft spiegelt und die umgekehrt diese Entwicklung tragen. Jacob Grimm, sichtlich hin- und hergerissen zwischen der Faszination durch eine staunenswert komplexe Formenwelt, wie sie aus den älteren Sprachen vertraut ist, und der schwer abweislichen Erkenntnis, dass ihre Entwicklung, von sporadischen Ausnahmen abgesehen, genau die umgekehrte Richtung nimmt, hat es in seiner *Geschichte der deutschen Sprache* (1848: 5f.) so formuliert: „Aus der geschichte der sprachen geht zuvorderst bedeutsame bestätigung hervor jenes mythischen gegensatzes: in allen findet absteigen von leiblicher vollkommenheit statt, aufsteigen zu geistiger ausbildung." Das heutige Deutsch mag leiblich nicht mehr so vollkommen sein wie das Althochdeutsche, das Westgermanische oder gar das Urgermanische, aber wir können sehr viel mehr damit sagen. Wenn wir verstehen wollen, was das Deutsche leistet oder zumindest leisten könnte, dann müssen wir seinen Wortschatz und dessen Entwicklung betrachten.

Was weiß man eigentlich über den Wortschatzreichtum?

Wie umfangreich ist der deutsche Wortschatz heute, und wie sieht die Gewinn- und Verlustrechnung für das 20. Jahrhundert aus? Das sind zwei einfache Fragen, und man möchte annehmen, dass die Germanistik als hier zuständige Wissenschaft klare Antworten darauf hat, Antworten, die über den subjektiven Eindruck hinausführen und sich auf wohlgesicherte Fakten stützen. Das ist jedoch nicht so, und das wiederum ist kein Zufall: Es fällt aus einer Reihe von Gründen sehr schwer, solche Antworten zu geben. Wenn es denn überhaupt versucht wird, so orientiert man sich an den umfangreichsten Wörterbüchern, die es derzeit gibt (siehe etwa Haß-Zumkehr 2001: 381ff.). In der *Wikipedia* heißt es dazu (Stand 9. 8. 2013, hier ohne Anmerkungen zitiert; siehe auch Best 2006: 13ff., auf den sich der *Wikipedia*-Eintrag offenbar stützt):

> Der Wortschatz der deutschen Standardsprache umfasst ca. 75 000 Wörter, die Gesamtgröße des deutschen Wortschatzes wird je nach Quelle und Zählweise auf 300 000 bis 500 000 Wörter bzw. Lexeme geschätzt. So gibt Duden *Deutsches Universalwörterbuch* an, der Wortschatz der Alltagssprache werde auf etwa 500 000, der zentrale Wortschatz auf rund 70 000 Wörter geschätzt. Das *Deutsche Wörterbuch* von Jacob und Wilhelm Grimm (1852–1960) wird auf ca. 350 000 Stichwörter geschätzt; Wahrig (2008) gibt im abgedruckten Vorwort zur Neuausgabe 2006 an, dieses einbändige Wörterbuch enthalte über 260 000 Stichwörter. Solche Angaben geben Aufschluss darüber, als wie groß der deutsche Wortschatz mindestens geschätzt werden muss. Diese Wörterbücher enthalten jedoch nur geringe Anteile der vielen Fachwortschätze und sind auch insofern unvollständig, da Ableitungen und Komposita nur teilweise aufgenommen werden und die neuesten Neubildungen naturgemäß fehlen. Ein entscheidendes Kriterium für die Aufnahme von Wörtern ist ihre Verwendungshäufigkeit und Gebräuchlichkeit; ausgeschlossen werden solche Wörter, die aus einfachen zusammengesetzt sind und sich bei Kenntnis ihrer Bestandteile von selbst verstehen lassen. Damit ist klar, dass der Wortschatz insgesamt noch wesentlich größer sein muss; die Angabe von 500 000 Wörtern ist kaum übertrieben. Nimmt man Fachwortschatz hinzu, ist mit mehreren Millionen Wörtern zu rechnen. Allein die Fachsprache der Chemie enthält nach Winter (1986) rund 20 Millionen Benennungen. Vor diesem Hintergrund erscheint Lewandowskis Bemerkung: „Der Gesamtwortbestand des Deutschen wird auf 5 bis 10 Millionen Wörter geschätzt“ als noch zu tief gegriffen.

Demnach scheint gesichert zu sein, dass der Umfang des deutschen Wortschatzes zwischen 70 000 und mehr als zehn Millionen Wörtern liegt. Dass dieses Fazit etwas diffus ist, kann man aber nicht der *Wikipedia* anlasten; vielmehr illustriert das Zitat recht gut den Stand unseres Wissens und zugleich die Schwierigkeiten, ihn zu verbessern. Wie viele Stichwörter das genannte Grimmsche Wörterbuch in Wirklichkeit hat, weiß niemand. Die auf der digitalen Version des *Deutschen Wörterbuchs* beruhende Zählung von Schares (2006) kommt auf 319 295; die kor-

rigierte Fassung, die als Teil des *Digitalen Wörterbuchs der deutschen Sprache* (www.dwds.de) zugänglich ist, umfasst 336 925; einige Stichwörter sind jedoch nicht als solche markiert, und so wird die Gesamtzahl in der Tat um die 350 000 betragen. Damit ist es das bei weitem umfassendste deutsche Wörterbuch. Es spiegelt jedoch im Wesentlichen nur den deutschen Wortschatz bis etwa 1900 wider, und das auch nur in den später geschaffenen Teilen; die noch von den Brüdern Grimm selbst bearbeiteten Buchstaben A bis F waren bereits 1863 abgeschlossen.[3] Über den Umfang des gegenwärtigen deutschen Wortschatzes kann man daraus wenig ableiten, erst recht nicht über die Veränderung seit 1900. Das zehnbändige *Große Wörterbuch der deutschen Sprache* (Duden, zuletzt 1999), das umfassendste deutsche Wörterbuch aus neuerer Zeit, zählt nach eigenen Angaben etwa 200 000 Stichwörter; etwa 25 000 davon sind jedoch reine Querverweise („Corpus, siehe Korpus“) oder kleinere Varianten, sodass man eher mit etwa 175 000 rechnen muss. Zum Vergleich: Das berühmte *Oxford English Dictionary*, wie das Grimmsche ein historisches Wörterbuch, aber auf sehr aktuellem Stand, weist derzeit etwa 620 000 Stichwörter auf (siehe www.oed.com, 1. 8. 2013); der *Grand Robert* (zuletzt gedruckt 2001) beschreibt für das Französische nach eigenen Angaben 100 000 Stichwörter mit insgesamt 350 000 Bedeutungen. In all diesen Fällen gibt es aber einen erheblichen Unterschied zwischen dem, was Sprecher oder Schreiber tatsächlich alles als Wort verwenden, und dem, was in ein gedrucktes Wörterbuch als Stichwort aufgenommen wird: Ein Wörterbuch zeichnet immer nur ein vom jeweiligen Zweck bestimmtes, stets jedoch stark verengtes Bild vom tatsächlichen Wortschatz und damit vom lexikalischen Ausdrucksreichtum einer Sprache.

So viel zum Gesamtumfang. Was nun die Frage nach Gewinn und Verlust angeht, so kann jeder eine Reihe von Wörtern nennen, die man um 1900 noch nicht verwendet hat und die heute gang und gäbe sind, zumeist weil man die Sache vor hundert Jahren noch nicht oder nicht unter diesem Namen kannte: *fernsehen, rumgurken, aufmischen, Sex, Fernseher, Versorgungsausgleich, Auszeit, abgezockt*. Weitaus schwieriger ist es, klare Beispiele für Verluste anzugeben. Zwar muten uns viele Wörter – *Droschke, Leibstuhl, füglich, behufs, weiland, abzwecken* – ungebräuchlich an. Aber man versteht sie immer noch,[4] und ob sie

3 Die 1962 begonnene Neubearbeitung der Buchstaben A bis F, die kurz vor dem Abschluss steht, wird etwa 20 000 Stichwörter mit Bedeutungsbeschreibungen umfassen; eine Reihe weiterer wird nur durch Belege illustriert.

4 Natürlich versteht nicht jeder, was *behufs* oder *weiland* bedeuten; aber das gilt auch für viele neue Wörter. Der Einzelne versteht immer nur einen kleinen Teil des gesamten Wortschatzes, wobei dieses Verstehen ganz unterschiedlich weit geht. In vielen Fällen hat man nur eine ungefähre Vorstellung (*Spund* hat etwas mit Fässern zu tun, *sintern* irgendwie mit legieren oder verschmelzen), in manchen anderen eine ganz falsche.

tatsächlich nicht mehr aktiv verwendet werden, ist schwer zu entscheiden. Dass man ein Wort selber nicht gebraucht oder schon eine Weile nicht mehr gehört oder gelesen hat, heißt ja nicht, dass es nicht mehr da ist; ich selber habe auch schon seit langem keinen Maikäfer mehr gesehen. Wie kann man über das intuitive Empfinden des Einzelnen hinaus zu einigermaßen gesicherten Fakten über den derzeitigen Umfang des deutschen Wortschatzes und seine Veränderung im 20. Jahrhundert kommen?

Das ist schwer, schwerer, als die Zahl der Bäume im Amazonasbecken und die Veränderung dieser Zahl über die letzten hundert Jahre anzugeben. Die Gründe sind im Prinzip ähnlich. Erstens ist nicht leicht zu sagen, was denn die Einheiten sind, die gezählt werden sollen. Zweitens ist nicht ganz klar, wo die genauen Grenzen liegen, innerhalb derer gezählt werden soll. Drittens ist es in der Praxis sehr aufwendig, die erforderlichen Daten zu beschaffen und hinlänglich zu untersuchen. Wörter zu zählen ist zwar nicht so gefährlich, wie das Amazonasbecken zu bereisen, aber es ist aus anderen Gründen mühselig und oft wenig ertragreich.

Beim Wortschatz sind die Einheiten „Wörter" – aber was ist eigentlich ein Wort? Anders als einen Baum kann man ein Wort nicht sehen: Wörter sind abstrakte Einheiten, Verbindungen von wahrnehmbaren Formen – der Lautgestalt oder Schriftgestalt des jeweiligen Wortes – mit Bedeutungen. Diese Formen, also etwa die Lautfolge [nɔx] oder die Buchstabenfolge noch, sind aber nicht die Wörter selbst. Sie sind nur das, was man davon hören oder lesen kann. Etwas anders gesagt: *noch* ist kein Wort – es ist die Art, wie das Wort *noch* nach gängiger Orthographie geschrieben wird. Leider gibt es, anders als man zu glauben geneigt ist, in aller Regel keine 1:1-Beziehung zwischen Form und Bedeutung. Um zu einer vernünftigen Aussage zu kommen, muss man daher etwas genauer betrachten, was eigentlich ein „Wort" ist oder als solches gelten soll; darauf komme ich gleich zurück.

Beim Amazonasbecken sind die Grenzen, innerhalb derer man zählen soll, zwar nicht völlig, aber doch einigermaßen klar, und man kann mit einer Luftaufnahme feststellen, um wie viel kleiner ungefähr die bewaldete Fläche in einer bestimmten Zeit geworden ist. Beim deutschen Wortschatz ist dies nicht so, selbst dann nicht, wenn man sich auf eine bestimmte Zeit, etwa die Gegenwart, beschränkt: Was ist „der" deutsche Wortschatz? In dem oben zitierten *Wikipedia*-Auszug ist von der „deutschen Standardsprache" die Rede, deren Umfang mit 75 000 Wörtern beziffert wird, von der deutschen „Alltagssprache", die von der Dudenredaktion auf etwa 500 000 Wörter geschätzt wird, wobei „der zentrale Teil" etwa 70 000 Wörter ausmache. Nimmt man den „Fachwortschatz" hinzu, so kommt man hinwieder auf einige Millionen Wörter. Um hier etwas Sinnvolles

sagen zu können, muss man zunächst einmal klar festlegen, was man zum deutschen Wortschatz zählen will.[5]

Das dritte Problem ist eher praktischer Natur: Wie bekommt man Zugang zu den Einheiten, die man zählen möchte? Wo findet man den Wortschatz einer Sprache? Für Sprachen, die keine Schrifttradition haben, gibt es dafür nur eine Antwort: in den Köpfen derer, die die Sprache beherrschen, denn wo soll er sonst sein? Aber schon bei nur gesprochenen Sprachen und erst recht in bedeutenden Kultursprachen mit langer Schrifttradition geht das, was die Sprache ausmacht, weit über das Wissen hinaus, das ein Einzelner davon hat. Der einzige Zugang zu einem Wortschatz besteht daher darin, den Gebrauch zu untersuchen, den die Einzelnen von ihrem sprachlichen Wissen gemacht haben. Daran haben auch die modernen bildgebenden Verfahren nichts geändert: Sie erlauben einen Blick ins Gehirn, aber dort sehen wir Nervenzellen oder Veränderungen des Sauerstoffgehalts, nicht jedoch Wörter oder Regeln. Grundlage für das Studium der Lexik sind daher große Textkorpora, die den Gebrauch dokumentieren. Anders als das Amazonasbecken sind diese Korpora jedoch nicht vorgegeben. Sie müssen nach verschiedenen Kriterien zusammengestellt und durchsuchbar gemacht werden. Streng genommen können wir daher niemals sagen, welchen Umfang „der Wortschatz einer Sprache“ tatsächlich hat. Wir können lediglich sagen, wie viele Wörter in Korpora einer bestimmten Zusammensetzung verwendet werden und wie sich dies im Laufe der Zeit ändert. Nur in diesem Sinn sind die beiden eingangs dieses Abschnitts gestellten, so einfach wirkenden Fragen überhaupt beantwortbar, und in diesem Sinne müssen auch die im Folgenden gegebenen Angaben verstanden werden. Aber selbst dann ist jede Antwort mit einer Reihe von Problemen und Unsicherheiten behaftet, die man sich vor Augen halten muss, wenn man die Tragweite des im Folgenden Gesagten recht einschätzen will. Dazu müssen wir etwas ausführlicher auf den hier zugrundegelegten Begriff von Wort und auf die hier zugrundegelegten Korpora eingehen.

5 Anders ist es, wenn es von Anfang an ein klar definiertes Korpus von Texten gibt, etwa alles, was ein bestimmter Autor geschrieben hat. Dann kann man den Wortschatz dieses Autors im Prinzip klar angeben. Der Wortschatz Goethes beispielsweise umfasst rund 91 000 Wörter (siehe www.bbaw.de/bbaw/Forschung/Forschungsprojekte/gwb). Dieser für einen einzelnen Autor ganz ungewöhnliche Reichtum erklärt sich aus der großen inhaltlichen Vielfalt seiner Schriften. Der Wortschatz Georg Trakls in all seinen Dichtungen umfasst nur rund 3 800 Wörter (Klein & Zimmermann 1971).

Wörter

Aus der Warte des Linguisten ist ein Wort ein komplexes Bündel von zumindest drei Arten von abstrakten Eigenschaften. Bei dem einfachen deutschen Wort *Uhr* sieht dies etwa so aus:

1. Formeigenschaften: Hier unterscheidet man gewöhnlich zwischen der phonologischen Form, die man lautschriftlich als [u:ɪ] beschreiben kann, und der graphematischen Form, hier also der Buchstabenfolge *Uhr*. Letztere gibt es natürlich nur, wenn die Sprache überhaupt verschriftet ist.

2. Grammatische, d. h. syntaktische und morphologische Eigenschaften: hier etwa „Nomen, femininum, wird gemäß Flexionsklasse xyz dekliniert".

3. Semantische Eigenschaften: hier etwa „Gerät zum Messen der Zeit".

Das entspricht nur teilweise der Alltagsvorstellung davon, was ein Wort ist. Diese Vorstellung ist sehr stark von der geschriebenen Sprache bestimmt; demnach ist ein Wort so etwas wie eine Folge von Buchstaben zwischen zwei Leerzeichen oder einem Leerzeichen und einem Satzzeichen. Das ist im Grunde seltsam, denn die geschriebene Sprache ist, so wichtig sie sein mag, gegenüber der gesprochenen in vierfacher Hinsicht sekundär: a) die meisten Sprachen in der Geschichte der Menschheit wurden und werden nur gesprochen, dennoch haben sie natürlich einen Wortschatz; b) fast alle Menschen lernen zu sprechen, nicht alle aber lernen zu schreiben; c) jene, die es tun, lernen es normalerweise erst, nachdem sie eine gesprochene Sprache beherrschen: Wir benutzen Wörter längst, bevor wir sie schreiben; d) auch jene, die schreiben und lesen können, machen oft nur einen sehr eingeschränkten Gebrauch davon, während es nur wenige gibt, die nicht regelmäßig sprechen und Gesprochenes verstehen. Primär sind daher Wortschatz und Wort der gesprochenen Sprache.[6]

Im Folgenden werden wir uns dennoch am geschriebenen Deutschen orientieren und das gesprochene vernachlässigen. Dies ist eine klare und bedauerliche Beschränkung, die aber aus zwei Gründen unvermeidlich ist. Zum einen bildet sich ein reicher Wortschatz erst mit der Schriftsprache und der von ihr getragenen Kultur aus.[7] Zum andern ist es kaum möglich, stichhaltige Aussagen

6 Wie stark die Fixierung durch die geschriebene Sprache hier ist, sieht man sehr schön daran, dass fast alle Studenten in sprachwissenschaftlichen Einführungskursen fest davon überzeugt sind, dass es beim Sprechen zwischen „den Wörtern" kleine Pausen gibt.

7 Derzeit gibt es allenfalls für die Hälfte der rund 7 000 Sprachen, die noch auf der Welt gesprochen werden, auch eine Schrift. Das heißt aber nicht, dass es entsprechend viele Schriftkulturen gibt; eine lange Tradition der Schriftnutzung mit all ihren Folgen für die kulturelle und soziale Entwicklung einerseits, für den Ausbau des Wortschatzes anderseits ist bei vergleichsweise we-

über den Wortschatz der älteren gesprochenen Sprache zu machen: Wir haben wenig verlässliche Zeugnisse, wie man um 1900 gesprochen hat, von den Zeiten davor ganz zu schweigen. Selbst für die Gegenwart gibt es nur wenige brauchbare Datensammlungen zum gesprochenen Deutsch (die dann verschriftlicht sein müssen; die reichste dieser Sammlungen findet sich am Institut für Deutsche Sprache, Mannheim, unter http://dsav-oeff.ids-mannheim.de/DSAv/KORPORAI.HTM). Beide Gründe haben übrigens auch dazu geführt, dass die gesprochene Sprache in den herkömmlichen Wörterbüchern nur in engen Grenzen berücksichtigt wurde (siehe jedoch Ruoff 1990). Ein weiterer Grund dafür ist die oft unbewusste Vorstellung, dass, was nicht in die Schriftsprache Eingang gefunden hat, auch nicht so recht verdient, in einem Wörterbuch beschrieben zu werden.[8] Das mag aus einer bestimmten Werthaltung heraus verständlich sein („odi profanum vulgus"), aber ein Wissenschaftler sollte sich mit solchen Wertungen zurückhalten und redlich zu sagen versuchen, wie es eigentlich ist. Wo er das nicht vermag, sollte er diese Begrenzung zugeben. Also: Hier wird die gesprochene Sprache und ihre Entwicklung von 1900 bis in die Gegenwart leider nicht erfasst, genauer gesagt: Sie wird nur insoweit erfasst, als sie mit der geschriebenen Sprache übereinstimmt.

Die oben gegebene Definition von Wort als Verbindung dreier Eigenschaftsbündel bezieht sich nun auf das Wort als lexikalische Einheit – also als Element der Lexik einer Sprache. Davon muss man sehr scharf das Vorkommen eines Wortes in einem Text unterscheiden; für ein solches Vorkommen sagt man auch oft „Wort". Der vorige Satz ist 22 „Wörter" lang (= Vorkommen von Wörtern), das Wort *ein* kommt darin dreimal vor, die Wörter *man, Vorkommen* und *Wort* zweimal, alle anderen einmal (= Wort als lexikalische Einheit). Damit nicht genug: In einer flektierenden Sprache wie dem Deutschen muss man nun noch einmal zwischen der lexikalischen Einheit und ihren verschiedenen Flexionsformen unterscheiden: die lexikalische Einheit *geh-* kann in den Flexionsformen *gehe, gehst, geht, ging, gegangen* und anderen vorkommen, die lexikalische Einheit *Vater* in den Flexionsformen *Vater, Vaters, Väter*. Im Folgenden werden wir daher, wenn es zu Missverständnissen kommen könnte, zwischen drei Wortbegriffen unterscheiden:

---

nigen Sprachen ausgebildet (die beste Information über die Sprachen der Welt findet sich unter www.ethnologue.com; dort wird für etwa 3 000 Sprachen eine Schrift verzeichnet, für nur rund 70 aber eine lange Schriftkultur).

8 Eines der wenigen ist Küppers verdienstliches *Wörterbuch der deutschen Umgangssprache* in sechs Bänden (1955–1970), das zwar in vielem überholt ist, aber leider bis heute keinen besseren Nachfolger gefunden hat.

Lexem: Das ist das Wort als lexikalische Einheit; in einem Wörterbuch ist ein solches Lexem, falls flektierbar, bei Verben gewöhnlich als Infinitiv (*sagen*) verzeichnet, bei Nomina als Nominativ Singular (*Lamm*) und bei Adjektiven in der prädikativen Form (*schrill*).[9]

Wortform: Damit sind die verschiedenen flektierten Formen gemeint, unter denen ein Lexem auftreten kann – bei nichtflektierbaren Wörtern eben nur eine.

Textwort: Das bezieht sich auf die mehr oder minder häufigen Vorkommen einer Wortform (und damit auch des Lexems, das es repräsentiert) in einem fortlaufenden Text.

Wenn man von einem Korpus spricht, das eine Milliarde Wörter umfasst, so meint man damit zunächst einmal Textwörter im obigen Sinn, manche davon sind sehr häufig (z. B. *die* oder *und*), andere kommen sehr selten, oft sogar nur einmal im Korpus vor. Im Englischen spricht man gewöhnlich von „tokens" – das sind die Textwörter – und „types" – das sind die verschiedenen Wortformen. Das Verhältnis zwischen beiden, die sogenannte type token ratio (TTR), ist eine wichtige Kenngröße in der quantitativen Sprachforschung: Je höher die type token ratio eines Textes, desto differenzierter ist seine Lexik, weil bei gleicher Länge mehr verschiedene Wörter vorkommen.

Bei dieser Redeweise trennt man allerdings oft nicht zwischen Wortformen und Lexemen. Das kann man sich im Englischen noch einigermaßen leisten, weil es wenig flektiert und man daher ein Lexem oft mit einer einzigen Wortform gleichsetzen kann. Im Deutschen mit seiner reicheren Flexion würde dies zu starken Verzerrungen führen, denn wenn man den Umfang der Lexik bestimmen will, kommt es ja auf die Lexeme an, nicht auf die verschiedenen Flexionsformen, die ein Lexem haben kann. Für eine sinnvolle Untersuchung müssen daher alle Texte „lemmatisiert" werden; dabei wird jede vorkommende Wortform auf die zugrundeliegende lexikalische Einheit zurückgeführt – also *Vater, Vaters, Väter* werden allesamt als Ausdruck eines einzigen Lexems angesehen. Bei kleinen

9 Leider hat sich in der Sprachwissenschaft hier keine einheitliche Terminologie eingependelt; der Ausdruck „Lexem" wird manchmal auf lexikalische Einheiten beschränkt, die einen gewissen deskriptiven Gehalt haben, wie etwa *Uhr, lachen, mulmig* oder *gestern* (Inhaltswörter) im Gegensatz zu lexikalischen Einheiten, die lediglich eine grammatische Funktion haben, wie *die, dass* oder *es* (Funktionswörter). Das tun wir hier nicht. Umgekehrt wird für das, was hier Lexem genannt wird, gelegentlich auch der Ausdruck „Lemma" verwendet. Normalerweise ist dies jene Flexionsform, unter der das Lexem in einem Wörterbuch angeführt ist – es ist sozusagen der Name eines Lexems; deshalb entsprechen 1000 Lemmata auch 1000 Lexemen. In der Psycholinguistik hat Lemma wiederum eine ganz andere Bedeutung. Wir bleiben deshalb hier bei dem Terminus „Lexem" für „lexikalische Einheit".

Korpora kann man das von Hand machen. Bei großen Korpora muss es automatisch geschehen; niemand kann ein Korpus, das eine Milliarde Textwörter umfasst, von Hand lemmatisieren (eine Milliarde Sekunden entspricht knapp 32 Jahren). Nun kann ein und dieselbe Wortform zu verschiedenen Lexemen passen (die Wortform *sein* kann der Infinitiv des Hilfsverbs *sein* sein, aber auch zu dem Possessivpronomen *sein* gehören, wie in *sein Hut*; die Wortform *heute* ist meistens ein Zeitadverb, aber es kann auch das Präteritum des Verbs *heuen* sein). Dies automatisch aufzulösen gelingt nicht immer, und deshalb hat auch die beste Lemmatisierung eine gewisse Fehlerquote, die man bei allen Aussagen über die Anzahl der Lexeme im Deutschen in Rechnung stellen muss. Das ist schlecht, aber es geht nicht anders, und es wird auch in Zukunft nicht anders gehen, denn die derzeitigen Lemmatisierungsverfahren lassen sich nur noch in Grenzen verbessern. Daher bleibt bloß die Wahl zwischen schweigen oder mit einer gewissen Fehlerquote leben.

Weitaus problematischer als diese Fehlerquelle ist ein Umstand, der sich aus der Natur eines Lexems selbst ergibt. Ein Wort verbindet immer eine bestimmte Form – in der Schriftsprache eine Zeichenfolge – mit Bedeutungen. Man wäre geneigt zu denken, dass diese Zuordnung normalerweise eindeutig ist und dass bekanntermaßen mehrdeutige Wörter wie *Strauß* oder *Schloss* die Ausnahme sind. Tückischerweise ist das nicht so; der Normalfall ist vielmehr, dass ein Wort mehrere Bedeutungen hat. Um sich davon zu überzeugen, braucht man nur einen Blick in ein etwas umfangreicheres Wörterbuch zu werfen. Das Folgende ist in etwas vereinfachter Form der Eintrag „Absatz“ aus dem *Digitalen Wörterbuch der deutschen Sprache* (www.dwds.de):

Absatz
mask., -es, Absätze

I 1 Unterbrechung eines fortlaufenden Textes
a *einen Absatz machen* (einen neuen Abschnitt beginnen)
b Abschnitt
in diesem Absatz behandelt der Verfasser …

2 Unterbrechung einer Fläche; Vorsprung
der Absatz des Berges, der Mauer
Unterbrechung der Stufen, Podest
der (obere) Absatz der Treppe

3 Unterbrechung einer Tätigkeit
in Absätzen reden (stockend reden)

II Erhöhung der Schuhsohle unter der Hacke
der Absatz des Schuhs

bildlich
Die Welt hatte eiserne Absätze – A. Zweig Grischa 146

III Verkauf, Vertrieb (fast nie im Plural)
Waren, Erzeugnisse haben reißenden, schnellen, großen, guten, sicheren, langsamen, geringen, schlechten Absatz

IV Ablagerung, Niederschlag (fast nie im Plural)
der Absatz von Kesselstein, Gestein, Land

Offenbar sind hier vier ganz verschiedene Bedeutungen mit ein und derselben Form, nämlich *Absatz*, verbunden. Handelt es sich hier um eine lexikalische Einheit oder um vier? Anders gesagt: Soll man die gemeinsame Form als einziges Kriterium werten, oder soll man die Bedeutungen, mit denen die Form verbunden sein kann, mit in Rechnung stellen? Wenn man den Ausdrucksreichtum der deutschen Sprache erfassen will, dann müsste man eigentlich von vier Lexemen reden, die alle *Absatz* geschrieben werden. Nimmt man dies ernst, dann würde eine umfassende und dennoch gut abgesicherte Untersuchung schier unmöglich, weil man dann nicht mehr von den beobachtbaren Formen, so wie sie sich im Text darstellen, ausgehen kann, sondern die einzelnen Wortvorkommen auf ihre Bedeutung analysieren müsste.

Noch heikler wird es, wenn eine Buchstabenfolge nicht nur mit verschiedenen Bedeutungen verbunden ist, sondern auch unterschiedliche grammatische Eigenschaften hat. Sind *der Verdienst* und *das Verdienst* ein oder zwei Wörter? Der Plural von *Wort* kann – bei klarer Differenzierung in der Bedeutung – *Wörter* oder *Worte* lauten: Ist *Wort* als ein Wort zu rechnen oder als zwei? Die Buchstabenfolge *als* kann als Konjunktion verwendet werden (*als wir in die Stadt reinkamen*), als Vergleichswort beim Komparativ (*dümmer als die Polizei erlaubt*) oder als noch etwas anderes, für das die Grammatiker keinen guten Namen haben (*als Liebhaber ist er eine Katastrophe*): eines, zwei oder drei Wörter im Sinne von lexikalischen Einheiten? Eigentlich sollte man von dreien reden, denn jede dieser drei Verwendungsweisen bereichert das Ausdrucksvermögen unserer Sprache. Wir könnten, wenn man es etwas näher beschaut, im Deutschen leicht ohne die so auffällige Unterscheidung zwischen *trägt* und *trug, abstellt* und *abstellte, hat* und *hatte* auskommen, also die Unterscheidung von Präsens und Präteritum. Der damit markierte Zeitbezug ist zwar in vielen Fällen wichtig, aber er könnte ohne weiteres durch Adverbien wie *jetzt* (d. h. in der Gegenwart) und *früher* (d. h. in der Vergangenheit) angezeigt werden. Man hätte sogar die Möglichkeit, den Zeitbezug einfach offenzulassen. Das wird uns durch die in diesem Punkt etwas zwänglerische Beschaffenheit der deutschen Grammatik verwehrt: Wir müssen immer sagen, ob etwas in der Gegenwart oder der Vergangenheit ist, ob wir wollen oder

nicht. Hingegen wäre man in seinem Mitteilungsdrang deutlich beschränkt, wenn man die drei Bedeutungen, die mit der Buchstabenfolge *als* einhergehen, nicht zur Verfügung hätte. Man möge nur einmal versuchen, die drei oben mithilfe von *als* ausgedrückten Inhalte ohne *als* zu formulieren.

Wie steht es schließlich, wenn „ein Wort" – jetzt im Sinne einer lexikalischen Einheit – aus rein grammatischen Gründen auf verschiedene Stellen im Satz verteilt wird, beispielsweise das Verb *absetzen*: *Sie setzten den Vorsitzenden ab – sie haben den Vorsitzenden abgesetzt – weil sie den Vorsitzenden absetzten*. Soll man dann *setzten* und *ab* als eigene Wörter rechnen? Betrachtet man nur die Form, so handelt es sich um zwei klar getrennte Einheiten – zwei Wörter also. Aber welche Bedeutung haben dann diese beiden Wörter, und wie ergibt sich die Bedeutung des Verbs *absetzen* aus der Bedeutung dieser beiden Einzelwörter? Offensichtlich kann ein Lexem aus zwei im Satz vollkommen getrennten Formbestandteilen bestehen. So einleuchtend das im Prinzip ist, so schwierig ist es, wenn man bei der Analyse großer Textkorpora die darin vorkommenden Lexeme bestimmen will: Man muss eine komplizierte syntaktische Analyse des einzelnen Satzes vornehmen, um solche getrennten Teile wieder zusammenzuführen oder zusammen zu führen. Wie bei der Lemmatisierung kann man das, sobald die Korpora eine gewisse Größe erreichen, nur noch automatisch machen, und wie bei der Lemmatisierung ergibt sich daraus eine Anzahl von Fehlern, deren Zahl von der Komplexität des Satzes abhängt. Auch hier hat man nur die Möglichkeit, entweder stille zu sein oder sich mit einer gewissen Fehlerquote zu arrangieren.

Dies sind nur drei von vielen Problemen, vor die man gestellt ist, wenn man allgemeine Aussagen über den Reichtum des deutschen Wortschatzes und seine Veränderungen machen will. Im Prinzip kann man die meisten davon durch scharfsinnige theoretische Überlegungen und sorgfältige Analyse des Einzelfalls lösen. Das ist aber in der Praxis unrealistisch: Wenn ein Korpus eine Milliarde Textwörter umfasst, bräuchte man bei Achtstundentag und Siebentagewoche rund 96 Jahre, sofern man es schafft, ein Textwort pro Sekunde zu analysieren. Das dauert zu lang. Bescheidet man sich hingegen mit einem kleinen Korpus von vielleicht 100 000 Textwörtern – das entspricht einem kürzeren Roman –, dann kann man kein umfassendes Bild von der deutschen Lexik zu gewinnen. Wir werden uns daher im Folgenden an den geschriebenen Wortformen orientieren, also keine Differenzierungen nach Bedeutungen und verschiedenen grammatischen Funktionen vornehmen: *Absatz, als, noch, auf* (der Leser möge sich überlegen, welche Bedeutungen mit *noch* oder *auf* einhergehen) werden als ein einziges Lexem gerechnet. Phraseologismen wie *zur Welt bringen*, *den Teufel an die Wand malen, im Großen und Ganzen, Schlag ins Kontor* gelten hingegen als aus mehreren Lexemen zusammengesetzt, obwohl sie ihrer Bedeutung nach als Einheit gelten können. Als ein Lexem hinwieder betrachten wir Verben wie *absetzen*, die

bald zusammen, bald getrennt geschrieben werden (*absetzten, abgesetzt, setzten ... ab*), weil es dafür automatisierte Verfahren der syntaktischen Analyse gibt, auch wenn diese eine gewisse Fehlerquote aufweisen. Entsprechende automatische Analysen der Bedeutung gibt es hingegen bislang nicht.[10] Das sind klare Einschränkungen, mit denen man jedoch beim gegenwärtigen Stand der Wissenschaft leben muss, und es ist fraglich, ob sich das so bald ändern wird.

Daraus ergibt sich aber eine wichtige Folgerung für alle später genannten Zahlen. Diese Zahlen geben an, wie viel Wörter nach Ausweis der untersuchten Korpora zur Verfügung stehen. Wenn man aber den Reichtum des deutschen Wortschatzes nicht daran bemisst, welche Wörter zur Verfügung stehen, sondern daran, welche Bedeutungen man mit diesen Wörtern ausdrücken kann – und das ist ja letztlich entscheidend –, so muss man von weitaus höheren Zahlen ausgehen. Um wie viel höher, hängt davon ab, wie viele Bedeutungen man pro Wort ansetzt; das ließe sich aber nur aufgrund einer Einzelanalyse eines jeden Wortes sagen.[11]

Korpora

Fassbar wird der Wortschatz einer Sprache nur in seinem Gebrauch, so wie er in Texten dokumentiert ist. Welche und wie viele solcher Texte braucht man, um ein angemessenes Bild vom deutschen Wortschatz zu bekommen? Darauf gibt es keine eindeutige Antwort, weil es, wie oben im Abschnitt „Was weiß man eigent-

10 Die verschiedenen Bedeutungen von *Absatz* lassen sich bis zu einem gewissen Grad trennen, wenn man schaut, mit welchen anderen Wörtern sie bevorzugt vorkommen – beispielsweise mit *Schuh*, *Sohle*, *abbrechen*, mit *Zeile, einrücken, Paragraph* oder mit *Ware, Umsatz, einbrechen*. Diese Präferenzen kann man automatisch berechnen und beispielsweise in Form von „Wortwolken" darstellen (Didakowski & Geyken 2013). Im *Digitalen Wörterbuch der deutschen Sprache* (www.dwds.de) werden für alle Wörter aus dem Kernkorpus (siehe folgenden Abschnitt) solche Wortwolken angegeben; das führt aber immer noch nicht zu einer automatischen Trennung der Bedeutungen.

11 Eine hier nicht weiter behandelte Frage ist, wie viele „Wortstämme" es im Deutschen gibt. Damit sind gleichsam die Wortkerne gemeint, aus denen sich durch verschiedene Mittel der Wortbildung ganze „Wortfamilien" herleiten lassen. Das Wort *Verabredung* hat beispielsweise den Kern *red-*, der dann durch zwei Präfixe und ein Suffix ausgebaut wird. Eine genaue Definition ist nicht einfach zu geben und hier auch nicht weiter bedeutsam. Es gibt aus neuerer Zeit zwei Versuche, die deutschen Wortstämme zu erfassen, Augst (2009; erstmals 1998) und Splett (2009). Beide gehen etwas unterschiedlich vor, kommen aber jeweils auf 8 000 bis 9 000 solcher Stämme. Sie bilden gleichsam die „Keime" aller deutschen Wörter.

lich über den Wortschatzreichtum?“ erörtert, „den“ deutschen Wortschatz nicht gibt. Immerhin kann man sinnvolle Überlegungen darüber anstellen, welcher Ausschnitt dessen, was an lexikalischen Einheiten insgesamt gebraucht wird, erfasst werden soll.

Hier kommen drei Faktoren ins Spiel. Zum einen schwankt der Wortgebrauch sehr stark mit Thematik, Zeit, Ort und Stilebene, um nur die vier wichtigsten Dimensionen der Variation zu nennen. Zwar gibt es Wörter, die in allen Texten vorkommen, beispielsweise *als, die, es, in, so, und, weil* und dergleichen mehr, also Funktionswörter, die im Gegensatz zu Inhaltswörtern nur wenig inhaltliche Bedeutung tragen. Die meisten lexikalischen Einheiten gehören zu den Inhaltswörtern, die häufig – aber nicht zwangsläufig – thematisch gebunden sind; nicht in allen Texten ist gleichermaßen von *Steilpass, Strafmaß, fiedeln* oder *mulmig* die Rede. Je nachdem, welche Texte man daher zugrunde legt, erhält man sehr unterschiedliche Ausschnitte aus der gesamten Fülle von Lexemen, die einem Sprecher des Deutschen zu Gebote stehen. Man muss also versuchen, ein einigermaßen balanciertes Korpus oder auch, je nach Zweck, mehrere solcher Korpora aufzubauen, die in digitaler Form vorliegen müssen, um handhabbar zu sein.[12] Für die folgenden Untersuchungen werden drei solcher Korpora genutzt:[13]

1. Die Ausgangsbasis ist das Kernkorpus des *Digitalen Wörterbuchs der deutschen Sprache*, das derzeit an der Berlin-Brandenburgischen Akademie der Wissenschaften entsteht (www.dwds.de). Es besteht aus einer sorgfältigen, über das ganze 20. Jahrhundert gleichmäßig gestreuten Auswahl repräsentativer Texte, die sich zu etwa gleichen Teilen auf vier Textsorten verteilen:
   – Belletristik, d. h. Romane und Erzählungen;
   – Zeitungstexte;
   – Gebrauchstexte, z. B. Ratgeber, Kochbücher, Rechtstexte;

12 Die reichsten Textkorpora des Deutschen überhaupt sind natürlich die großen Bibliotheken und der deutschsprachige Teil des Internets. Beide sind aber für den vorliegenden Zweck nur von begrenztem Nutzen. Die Texte der Bibliotheken müssten, damit man sie in großem Maßstab auswerten kann, nicht nur digitalisiert, sondern auch so aufbereitet werden, dass man sie sinnvoll untersuchen kann. Das ist eine alles andere als triviale Aufgabe; ein Exempel für die dabei auftretenden Probleme findet sich in Anm. 17. Das Internet ist zwar schier unendlich reich, aber die Art der Texte ist einseitig, und die Texte müssten gleichfalls zuerst aufbereitet werden (zwei umfangreiche Sammlungen von Internet-Texten finden sich unter http://wortschatz.uni-leipzig.de sowie unter http://hpsg.fu-berlin.de/cow). Damit soll keineswegs bestritten werden, dass solche gigantischen Textmengen auch von großem wissenschaftlichen Nutzen sein können. Man muss sich nur der genannten Probleme bewusst sein.

13 Keines dieser Korpora enthält in einem Dialekt verfasste Texte. Dies schließt aber nicht aus, dass sich vereinzelt kleinere Dialekteinsprengsel oder, dies häufiger, einzelne dialektspezifische Wörter finden.

– wissenschaftliche Texte aus verschiedenen Gebieten.[14]

Für die Zeit ab etwa 1925 enthält es auch etwa 5% (transkribierte) Texte der gesprochenen Sprache, die aber aus den oben genannten Gründen hier nicht berücksichtigt sind.[15]

2. Für den Vergleich wurde ein Berichtskorpus zusammengestellt, das in drei den etwas unterschiedlichen Bedürfnissen angepassten Varianten vorliegt (zu den Unterschieden siehe Seelig im Anhang dieses Bandes). Für diesen Beitrag wurde Berichtskorpus A verwendet; es besteht aus drei Zeitscheiben, die jeweils Texte im Umfang von 10 Millionen Textwörtern aus einem Jahrzehnt umfassen (1905–1914, 1948–1957, 1995–2004). Für die letzte, noch halb ins 21. Jahrhundert hineinreichende Zeitscheibe wurden auch Texte aus den sehr umfangreichen Korpora des Instituts für Deutsche Sprache verwendet (www1.ids-mannheim.de/kl).

3. Für manche Zwecke war es auch wichtig, ein wesentlich umfassenderes Korpus heranzuziehen. Dieses Grüne Korpus besteht aus dem schon genannten Kernkorpus des DWDS, das durch Zeitungstexte neuerer Zeit (*Die Zeit*, *Süddeutsche Zeitung*, *Berliner Zeitung*, (Berliner) *Tagesspiegel*) auf einen Umfang von insgesamt einer Milliarde Textwörter erweitert wurde. Es ist daher nicht mehr so gut nach Textsorten balanciert, liefert aber ein breiteres Bild von den tatsächlich genutzten Wörtern.

Der zweite wichtige Faktor, den man bei der Korpuszusammenstellung berücksichtigen muss, ist das Anwachsen der Lexemzahl mit zunehmender Anzahl von Textwörtern. Wie umfangreich muss ein Textkorpus des Deutschen sein, wenn man alle oder doch annähernd alle Lexeme des Deutschen erwischen will? Darauf gibt es bei einer lebenden Sprache keine wirklich schlüssige Antwort, weil man immer wieder neue lexikalische Einheiten hinzufügen kann. Es ist aber so, dass sich die Chancen, neue, d. h. noch nicht vorgekommene Wörter in einem Korpus zu finden, zunehmend verringern, je umfangreicher das Korpus ist. Man kann

14 Bei letzteren handelt es sich durchweg um Texte, die nur ein geringes Maß an ausgesprochen fachspezifischen Termini (*Trimethylxanthin*) aufweisen. Dies ergibt sich aus dem Vorgehen bei der Auswahl: Es wurden einige hervorragende Wissenschaftler gefragt, was jeweils die prägenden deutschsprachigen Texte ihres Faches für einen bestimmten Zeitraum waren. Ein etwas trauriges Ergebnis für manche Fächer war dabei, am Rande bemerkt, dass es für manche Fächer in den letzten zwei oder drei Jahrzehnten nur noch wenig wirklich wichtige deutschsprachige Texte gibt.

15 Das DWDS-Kernkorpus wird durch eine Reihe sehr umfangreicher Zeitungskorpora (u. a. *Die Zeit*, *Süddeutsche Zeitung*, *Berliner Zeitung*, (Berliner) *Tagesspiegel*, *Bild*, *Welt*) ergänzt, die zum Teil nur intern genutzt werden dürfen, zum Teil in das Grüne Korpus eingegangen sind. Ein vergleichbares Korpus der *Frankfurter Allgemeinen Zeitung* stand uns leider nicht zur Verfügung.

sich das Problem vor Augen führen, wenn man anfängt, die *Gesammelten Werke* von Karl May, Band 1 bis Band 93, in dieser Reihenfolge zu lesen. Am Anfang sind die meisten Wörter neu, nach einigen Seiten ist schon viel vorgekommen, und am Ende des *Schut* hat man die meisten schon einmal gelesen. Dann gibt es eine thematische Verschiebung mit Szenenwechsel (*Winnetou I*), es gibt wieder mehr neue Wörter, usw., bis man schließlich ermattet bei Band 93 (*Briefwechsel mit Sascha Schneider*) angekommen ist. Dort, so würde man annehmen, sind alle Wörter schon einmal vorgekommen. Das ist aber falsch: Zwar flacht sich die Kurve der hinzukommenden Wörter ab, je umfangreicher ein Korpus wird, aber die Zunahme ist immer noch beträchtlich. Das gilt selbst für ein Korpus, das bereits eine Milliarde Wörter umfasst – allerdings nur unter der Voraussetzung, dass man immer weiter Texte unterschiedlicher Autoren und unterschiedlicher Thematik aufnimmt; wenn man tausendmal dasselbe Werk aufnähme, dann hätte man auch ein Korpus mit vielen Textwörtern, aber keine Zunahme an lexikalischen Einheiten.

Man kann die tatsächliche Zunahme recht gut nach dem sogenannten Heaps'schen Gesetz berechnen, das je nach Sprache ein wenig unterschiedliche Werte liefert (siehe etwa Tudjman 2005). Dieses Gesetz ist wichtig, wenn man von begrenzten Korpora auf die tatsächlichen Verhältnisse extrapolieren will. Wie oben erwähnt, stützt sich unsere Analyse der Wortschatzzunahme im 20. Jahrhundert auf drei Zeitscheiben mit je zehn Millionen Wörtern. Das ist viel, aber nicht genug. Für einen Teil der folgenden Untersuchungen wurden daher diese drei Zeitscheiben als Stichproben genommen und nach dem Heaps'schen Gesetz hochgerechnet („gehebelt").[16] Die berechneten Werte der 3. Zeitscheibe (1995–2004) wurden anhand des oben genannten, eine Milliarde Textwörter langen Grünen Korpus überprüft. Vergleichbar umfassende Korpora für die früheren Zeitscheiben gibt es nicht. Aber wenn sich der berechnete Wert für die 3. Zeitscheibe durch diese Überprüfung bestätigt, dann gibt es keinen gewichtigen Grund für die Annahme, dass die Berechnung nicht auch für die 1. und 2. Zeitscheibe gültig ist.

16 Das Heaps'sche Gesetz, das nach dem Informatiker Harold Stanley Heaps benannt ist (Heaps 1978), aber in ähnlicher Form bereits zuvor von dem Sprachstatistiker Gustav Herdan gefunden wurde (Herdan 1960), lautet: $V(n) = K n^{\beta}$. V ist dabei die Zahl der Wortformen, n die Zahl der Textwörter (d. h., n gibt die Textlänge in Wörtern an). K und β sind zwei freie Parameter, die sich von Sprache zu Sprache unterscheiden und experimentell bestimmt werden müssen. Man beachte, dass sich V auf die Wortformen (z. B. *geht, ging, gehst* usw.) bezieht, nicht auf die zugrundeliegende lexikalische Einheit (also das Verb *gehen*); dies muss auf Basis der Lemmatisierung entsprechend korrigiert werden. Allgemein zur quantitativen Lexikologie siehe Köhler (2005) und Köhler et al. (2008).

Der dritte Faktor schließlich ist der Umstand, dass ein Korpus, ganz gleich wie es zusammengestellt ist, immer auch zahlreiche Zeichenfolgen enthält, die man nicht so ohne weiteres zu den „deutschen Wörtern" schlagen würde. Dies sind:

a) falsch geschriebene Wörter, die sich bereits im Original finden oder durch Fehler bei der Digitalisierung entstanden sind;[17]
b) Namen: *Leica, Müritz, Spiesen, Uwe*;
c) Abkürzungen: *usw., z. B.*;
d) Akronyme: *BBAW*, *IDS*;
e) Ziffern und ähnliches: *319 295*, *8. 5. 1998*, *4711*;[18]
f) Wörter aus anderen Sprachen: *Giaur, Pizza*, *vulgo*.

Bei jeder dieser Kategorien lässt sich lange und mit guten Gründen darüber diskutieren, ob man die betreffenden Einheiten zur deutschen Lexik zählt oder nicht. Ist *Leica* ein Name oder eine Bezeichnung für eine Gruppe von Geräten (*Er besaß sieben Leicas*). Am schwierigsten ist das Problem bei Wörtern aus anderen Sprachen: Ab wann würde man ein aus einer anderen Sprache übernommenes Wort zur deutschen Lexik zählen? Hier gibt es einen gleitenden Übergang, der von bloßen Zitaten auf der einen Seite (*er sagte „schto skasaesch?"*) zu nach Bedeutung, Form und grammatischen Eigenschaften wohlintegrierten Wörtern,

17 Ein Beispiel aus *Google Books*, in dem der Leser unschwer einen Auszug aus Hegels *Wissenschaft der Logik* erkennt: „bamit fo fefyr erweiterten 2lnafyft8 auf bei @cometrie überljaupf, gekört ljat. T-aô problem ljat bei ifym bei ftorm ber Slufgabc, gerabe Sinien fenfredit auf beliebige Jdrte einer Gur>>e ju jieben, alo woburd) @ubtangente u.f.f. beftimmt tt>irb; man begreift bie Sefriedigung, bie er bafelbft über feine Cmtbecfung, bie einen Qòv genftanb >>on allgemeinen wiffenfdjaftlidjcn Sntcresse ber bamaltget." Die hohe Fehlerquote rührt daher, dass der Frakturtext des Originals mithilfe einer OCR-Software automatisch und ohne nachträgliche Korrektur in durchsuchbaren Volltext umgewandelt wurde. Dies funktioniert bei guten Antiqua-Vorlagen oft recht gut, bei älteren Frakturtexten führt es jedoch zu wenig befriedigenden Ergebnissen. Das dritte Wort *fefyr* – es kommt in *Google Books* insgesamt 94 100 Mal vor (9. 8. 2013) – entspricht dem in Fraktur geschriebenen Wort *sehr*. Aus diesem Grund ist *Google Books* – in vieler Hinsicht eine wundervolle Datenquelle, in der keineswegs alle Texte derart fehlerbehaftet sind – für wissenschaftliche Zwecke wie die vorliegenden nur von begrenztem Nutzen.

18 Ein ganz anderer Problemfall sind Zahlwörter, die sich ja nach Belieben zusammensetzen lassen. Wörter wie *drei* oder *hundert* sind sicher Bestandteile des deutschen Wortschatzes und eine wesentliche Bereicherung seiner Ausdrucksmöglichkeiten; in Kulturen ohne schriftsprachliche Tradition ist die Anzahl der Zahlwörter oft sehr beschränkt. Aber auf der anderen Seite wäre es sinnlos, alle bildbaren Zahlen zum deutschen Wortschatz zu rechnen, denn es gibt ihrer unendlich viele. In der Praxis löst sich das Problem dadurch, dass sich sinnvolle quantitative Aussagen über den deutschen Wortschatz immer nur relativ zu einem bestimmten Korpus machen lassen. In diesem Sinne wurden Zahlen, sofern sie nicht als Ziffern geschrieben sind, hier einbezogen.

etwa *Balkon* oder *toppen*, auf der anderen Seite reichen (siehe hierzu den Beitrag von Eisenberg in diesem Band, in dem das Problem am Beispiel der Anglizismen grundsätzlich diskutiert wird).

Bei unseren Analysen wurden Zeichenfolgen, die unter die Kategorien a bis e fallen, ausgeschlossen; da dies mit automatischen Verfahren gemacht wurde, gibt es wiederum eine gewisse Fehlerquote (bei Wörtern wie *Kohl* ist oft gar nicht zu entscheiden, ob es sich um einen Eigennamen oder die Bezeichnung eines Gemüses handelt). Schwieriger ist es mit der unter lexikalischen Aspekten besonders heiklen Kategorie f; hier wurden bei der Erstellung des Berichtskorpus manuell sehr viele Einzelentscheidungen durch die Bearbeiter getroffen; es sollte jedoch klar sein, dass man in vielen Fällen mit ebenso guten Gründen auch anders hätte entscheiden können.

Damit sind die Hauptprobleme einer Korpuserstellung und die hier gewählte Weise, mit ihnen umzugehen, genannt. Es ist ja eine lästige Unart der Wissenschaftler, unentwegt von den Problemen zu reden, statt zu den Ergebnissen zu kommen. Hier ist es aber nötig, weil sonst ein falsches Bild davon entsteht, wie gesichert die Befunde wirklich sind, und es ist zweifellos eine schlimmere Unart, eine Sicherheit vorzutäuschen, die nicht besteht und beim gegenwärtigen Stand unseres Wissens nicht bestehen kann. Es ist kein Zufall, dass man über den derzeitigen wie auch den früheren Umfang des deutschen Wortschatzes so wenig weiß. Man hat da nur die Wahl, diese Unsicherheiten offenzulegen – in der Hoffnung, dass damit ein Anstoß gegeben wird, sie Schritt für Schritt zu beseitigen –, oder aber das gegenwärtige Ignoramus zu einem Ignorabimus zu erklären.

Mit all diesen Kautelen (deutsches Wort?) kommen wir nun zu einigen Aussagen über die Lexik des Deutschen und ihre Entwicklung im 20. Jahrhundert. Irgendwann muss man aufhören, vom Springen zu reden, und springen.

Gesamtumfang des deutschen Wortschatzes 1905–2004

Ausgangspunkt sind die drei genannten Zeitscheiben im Umfang von je 10 Millionen Textwörtern. Rechnet man diese Zeitscheiben nun nach dem Heaps'schen Gesetz auf einen Gesamttext von einer Milliarde Wörter hoch, so kommt man auf die folgende Anzahl verschiedener Zeichenfolgen, die man als Wortformen betrachten könnte:

Zeitspanne	1905–1914	1948–1957	1995–2004
Wortformen	5 307 001	7 207 296	7 612 131

Aus diesen Rohwerten lässt sich immerhin bereits ablesen, dass es in den letzten hundert Jahren einen erheblichen Aufwuchs gibt, der in der zweiten Hälfte des Jahrhunderts jedoch offenbar geringer ausfällt. Will man die Zahl der Lexeme wissen, so müssen diese Werte bereinigt werden um: a) verschiedene Flexionsformen ein und desselben Wortes, b) all jene Zeichenfolgen, die nach dem im vorigen Abschnitt Gesagten nicht zu den deutschen Lexemen gerechnet sind, also Wörter aus anderen Sprachen, Ziffern, Akronyme, Abkürzungen, Namen und Fehler, die in den Originaltexten stehen (Druckfehler) oder bei der Korpuserstellung entstanden sind. Der genaue Anteil dieser auszuschließenden Formen schwankt in gewissen Grenzen. Man kann ihn nach unseren Erfahrungen im DWDS-Kernkorpus auf 25% bis 30% veranschlagen. Um auf der sicheren Seite zu sein, ziehen wir daher von den Rohwerten 30% ab. Dies ergibt folgendes Bild (von nun an sind die Zahlen auf Tausend gerundet, weil jede genauere Angabe eine nicht vorhandene Präzision vortäuschen würde):

Zeitspanne	1905–1914	1948–1957	1995–2004
Lemmata	3 715 000	5 045 000	5 328 000

Dies führt zu einem ersten wichtigen Ergebnis:

In einem Textkorpus der deutschen Gegenwartssprache, das eine Milliarde Textwörter lang ist, kommen etwa 5,3 Millionen lexikalischer Einheiten – also Wörter, so wie sie im Wörterbuch stehen – vor.

Nicht gerechnet sind dabei Mehrworteinheiten *wie ins Gras beißen* oder *jemandem ein Ohr abkauen*, die ihrer Bedeutung nach (*sterben* bzw. *zutexten*) streng genommen auch den Status einer lexikalischen Einheit haben (Burger et al. 2007 gibt umfassend über solche Mehrwortwörter Auskunft; die spezifischen Probleme bei ihrer lexikographischen Erfassung werden in Ďurčo 2010 diskutiert). Es sei hier noch einmal daran erinnert, dass die meisten Lexeme mehrere Bedeutungen haben können. Wenn man *Absatz* seinen vier verschiedenen Bedeutungen entsprechend nicht als eine lexikalische Einheit zählt, sondern als vier, dann enthält man entsprechend mehr lexikalische Einheiten für das Deutsche.[19]

Die obigen Werte sind berechnet. Der für die 3. Zeitscheibe angegebene Wert lässt sich jedoch in der Tat anhand eines realen Korpus gleicher Länge überprüfen. Wie oben im Abschnitt Korpora bemerkt, besteht dieses Grüne Korpus zum größten Teil aus Zeitungstexten der Gegenwart, enthält aber auch zu etwa 10% Texte, die über das ganze Jahrhundert verteilt sind. Die Zahl der lexikalischen

19 Einen Durchschnitt von vier Bedeutungen pro Lexem anzusetzen mag hoch sein. Es sei daran erinnert, dass, wie oben bemerkt, *Le Grand Robert* für das Französische nach eigenen Angaben 100 000 Stichwörter mit 350 000 Bedeutungen beschreibt.

Einheiten sollte daher ein wenig unter der nach Heaps' Gesetz berechneten Zahl für ein reines Gegenwartskorpus gleicher Größe liegen. Dies ist in der Tat so – im Grünen Korpus liegt der Wert bei knapp unter 5 Millionen Lexemen gegenüber dem berechneten Wert von 5,3 Millionen. Für die beiden ersten Zeitscheiben gibt es keine vergleichbaren Korpora dieser Größe; bis zum Beweis des Gegenteils gibt es jedoch keinen Grund anzunehmen, dass das Heaps'sche Gesetz für sie nicht gilt. Dies bringt uns zu dem zweiten zentralen Ergebnis:

Der deutsche Wortschatz hat im Verlauf des 20. Jahrhunderts um etwa ein Drittel – und wenn man eine Sicherheitsmarge annimmt, ein Viertel – zugenommen.

Dabei ist der Anstieg in der ersten Jahrhunderthälfte deutlicher als in der zweiten. Dieses Anwachsen der deutschen Lexik in den letzten hundert Jahren wird man nicht als eine Verarmung betrachten wollen.

Worin im Einzelnen liegt der Zuwachs? Genaue Zahlen dafür gibt es nicht und wird es auch so schnell nicht geben. Dazu müsste man sich die Lexeme im einzelnen ansehen, eine Aufgabe, die, selbst wenn man die grundsätzlichen Probleme bei „Fremdwörtern" ausklammert, eigene aufwendige Untersuchungen verlangen würde (fünf Millionen Wörter sind nicht so rasch überprüft). Eine erste Durchsicht zeigt aber sofort, dass eigenständige neue einfache Wörter, beispielsweise *rödeln* oder *mosern,* zwar durchaus vorkommen, aber selten sind. Die Zahl der Übernahmen aus anderen Sprachen, so sehr sie ins Auge fallen mögen, wird überschätzt (siehe dazu den Beitrag von Eisenberg in diesem Band). Der weitaus größte Teil des Zuwachses entfällt auf Wortbildungen aus bestehenden Wörtern – also auf Ableitungen und Komposita. Beides findet sich auch in anderen Sprachen, aber selten in dem Ausmaß, in dem man es im Deutschen vor allem bei den Komposita beobachten kann. Das wirft die schwierige Frage auf, ob man denn Ableitungen wie *Zocker* oder *geldmäßig* und Komposita wie *Nacktbackverbot* oder *Vorsorgeuntersuchung* wirklich als „neue Wörter" ansehen soll, die den Ausdrucksreichtum der deutschen Lexik ändern. Schließlich beruhen sie ja auf bekannten Bestandteilen und sind zumindest der Form nach gemäß festen Regeln gebildet. Dieser sehr schwierigen Frage soll hier kurz nachgegangen werden, da sie für eine Einschätzung des lexikalischen Ausdrucksreichtums sehr wichtig ist. Dabei beschränke ich mich auf Komposita, die den weitaus größten Teil des Aufwuchses ausmachen; für Ableitungen gilt cum grano salis dasselbe (die Produktivität von Suffixableitungen wurde – teils mit überraschenden Ergebnissen – systematisch erstmals in Schneider-Wiejowski 2011 quantitativ untersucht).

Als wichtigstes Kriterium dafür, ob man ein Kompositum als eigenes Lexem zählen soll, gilt in der Wortbildungslehre (etwa Fleischer & Barz 2012: 42ff.) der Grad der „Kompositionalität", d. h. das Ausmaß, in dem sich die Bedeutung des zusammengesetzten Worts aus der Bedeutung seiner Teile ergibt. Es gibt in der Tat viele Komposita, deren Bedeutung sich relativ klar aus der ihrer Bestand-

teile ablesen lässt, wie *Bäckerlehrling* oder *graublau*; man nennt solche Komposita oft „vollmotiviert". Für andere ist das aber durchaus nicht der Fall: wer nur *rubbeln* und *fest* kennt, weiß noch nicht, was *rubbelfest* bedeutet; immerhin denkt man sich, dass es etwas mit *rubbeln* zu tun hat – *rubbelfest* ist „teilmotiviert". In anderen Fällen sieht man überhaupt keinen Zusammenhang zwischen Teilen und Ganzem, beispielsweise bei dem in den letzten Jahren häufig verwendeten Wort *Herdprämie*, das daher als „unmotiviert" gilt. Eine andere in dieser Hinsicht nebulöse Neuerung ist das Adjektiv *zeitnah*, das anscheinend so viel bedeutet wie *rasch* – eine Deutung, die sich nicht unmittelbar aus der Bedeutung von *Zeit* und *nah* herleiten lässt. Diese dreistufige Scheidung ist verbreitet und durchaus auch sinnvoll; sie wird jedoch der tatsächlichen Komplexität der Kompositabedeutung nicht gerecht. Zum einen ist es nämlich so, dass die Teilwörter ihrerseits oft mehrere Bedeutungen haben, von denen ein Kompositum nur eine herausgreift. Bei *Absatzkurve* etwa ist dies – wahrscheinlich – die Bedeutung III „Verkauf, Vertrieb", bei *Absatzlänge* hingegen eher die Bedeutung II „Erhöhung der Schuhsohle unter der Hacke". Zum andren führt die Zusammensetzung oft auf eine ganze Gruppe von Bedeutungen, von denen aber nur eine tatsächlich verwendet wird.

Beide Probleme hängen oft zusammen. Man kann sie sich an einem so gängigen Wort wie *Parklücke* vor Augen führen. Die Verbindung von *Park* und *Lücke* könnte vieles bedeuten – eine Lücke in einem Park, eine Stelle in der Bebauung, die für einen Park freigelassen ist, der Abstand, den ein geparktes Auto von einem anderen anstandshalber einhalten sollte, und anderes mehr. Tatsächlich verwendet wird *Parklücke* aber nur für eine freie Stelle, an der man sein Fahrzeug abstellen kann. Ebenso könnte man rein aufgrund der Bauform von *Führerschein* darunter auch eine Gloriole um einen Führer verstehen (analog zu *Heiligenschein*), man tut es aber nicht. Man muss daher unterscheiden zwischen der Bedeutung, die ein Kompositum rein aufgrund seiner Zusammensetzung haben könnte, und der Bedeutung, in der es tatsächlich verwendet wird. Letztere kann man nur in Grenzen aus der Bedeutung der Bestandteile ablesen. Darüber täuscht man sich leicht hinweg, denn wir treffen uns bisher unbekannte Komposita gewöhnlich in einem bestimmten Textzusammenhang an, der uns Aufschluss darüber gibt, was denn nun tatsächlich gemeint ist. Anders gesagt, wir verstehen unvertraute Komposita nicht allein aufgrund ihrer sprachlichen Form, sondern auch aufgrund der Information, die uns aus dem jeweiligen Kontext zukommt. Ebenso verstehen wir auch oft einfache Wörter, die uns bislang nicht bekannt waren.[20] Anders als man

20 Besonders schön sehen kann man dies, wenn man einen Text in einer anderen Sprache liest und auf ein Wort trifft, das man nicht kennt oder das man vielleicht einmal gelernt hat, aber dessen Bedeutung einem entfallen ist.

zu glauben geneigt ist, sind neue Komposita daher in der Tat zumeist eine echte Erweiterung der Lexik einer Sprache. Eine ganz andere Frage ist, ob man jedes Kompositum in ein Wörterbuch aufnehmen soll (Schippan 1992; Schlaefer 2009). Traditionell geschieht dies nicht, und das mit guten Gründen. Zum einen leiden gedruckte Wörterbücher unter Raumbeschränkungen; zum andern ist es angesichts des Zwecks der Wörterbücher oft auch nicht nötig (Engelberg & Lemnitzer 2009). Wörterbücher haben nicht so sehr die Aufgabe, die spezifische Bedeutung eines Wortes erschöpfend zu beschreiben – das gelingt fast nie. Vielmehr sollen sie in erster Linie ihren Benutzern helfen, einen Text, in dem das Wort vorkommt, zu verstehen. Bei einem Kompositum kann der Benutzer allein schon aufgrund seiner Kenntnis der Bestandteile eine gewisse Vorstellung von der Verwendungsweise des ganzen Wortes gewinnen; diese Vorstellung reicht dann aus, um das Wort im Kontext richtig zu verstehen.

Unterschiede in den Textsorten

Bislang haben wir den Umfang des Gesamtwortschatzes betrachtet, so wie er sich in Korpora bestimmter Länge niederschlägt. Im Folgenden wird nun nach den vier Textsorten Belletristik, Zeitungen, Gebrauchstexte und wissenschaftliche Texte differenziert.[21] Anders als im vorigen Abschnitt betrachten wir dabei keine hochgerechneten Korpora, sondern die tatsächlichen drei Zeitscheiben mit einem Umfang von jeweils 10 Millionen Textwörtern, weil es uns hier weniger auf die Gesamtgröße ankommt als auf die relativen Unterschiede.

Eine erste Frage ist hier, wie viele Lexeme sich in allen vier Textsorten über alle Zeiten hinweg finden, was also die Konstanten über das 20. Jahrhundert sind. Es sind dies nur 8% der Lexeme – mit anderen Worten, die verschiedenen Textsorten haben eine sehr hohe Spezifik im Wortschatz über die Zeit hinweg. Diese 8% gemeinsamer Wörter decken jedoch über 90% aller Wortvorkommen ab. Das liegt daran, dass es sich dabei großenteils um inhaltsarme Wörter (Funktionswörter) wie *dass, die, in, nach, so, und, weil* handelt. Sie sind daher nicht textspezifisch. Das heißt allerdings nicht unbedingt, dass sie auch in allen Texten gleich

21 Diese Einteilung folgt den Vorgaben der zugrundeliegenden Korpora. Sie ist offenkundig grob; Zeitungen beispielsweise setzen sich im Grunde aus sehr verschiedenen Textsorten zusammen. Es ist aber im Prinzip aufgrund der Angaben in den Korpora möglich, hier feiner zu differenzieren, indem man die einzelnen Texte mit entsprechenden „Metadaten", d. h. Schlagwörtern, die den Text näher kennzeichnen, versieht. Der praktische Aufwand ist allerdings doch erheblich.

oft vorkommen; es mag sehr wohl sein, dass *dann* in einem narrativen Text viel häufiger ist als in einem wissenschaftlichen, während es sich bei *weil* umgekehrt verhält: *dann* ist charakteristisch für temporale Strukturen, *weil* für argumentative. Dieser Frage werden wir weiter unten exemplarisch nachgehen.

Im Deutschen gibt es etwa 200 solcher Funktionswörter, darunter die drei häufigsten deutschen Wörter überhaupt, nämlich *der* (mit allen Flexionsformen wie *die, dem, denen* usw.), *und* sowie *ein* (dies ebenfalls mit allen Flexionsformen). Unterscheiden sich die Textsorten nun in ihrem Anteil an solchen inhaltsarmen Wörtern? Ja. In der Belletristik machen sie 51% aller Wortvorkommen aus, in den Gebrauchstexten 48%, in der wissenschaftlichen Prosa 46%, in den Zeitungen 45%. Die Unterschiede sind nicht sehr groß, aber doch deutlich. Die Belletristik hat demnach eine leichte Neigung zu inhaltsarmen Wörtern. Dieselbe Tendenz zeigt sich auch, wenn man betrachtet, welche Wörter – diesmal nun inhaltsreiche und daher thematisch stärker gebundene – in nur einer einzigen Textsorte vorkommen: Belletristik 13%, Gebrauchstexte 16%, wissenschaftliche Literatur 22%, Zeitungen 23%, also fast doppelt so viel als in der Belletristik. Das mag daran liegen, dass das thematische Repertoire der Belletristik begrenzter ist als das der Zeitungen oder auch der wissenschaftlichen Literatur, die verschiedenen Disziplinen entstammt und daher auch einen stärker variierenden Wortschatz aufweist.[22]

Vergleichen wir nun die lexikalische Entwicklung der vier Textsorten über die Zeit. Die folgenden Zahlen geben an, wie viele Lexeme sich in einem Text von zehn Millionen Textwörtern finden. Sie sind bereits um Flexionsformen, Zahlen und Fehler bereinigt und auf 100 gerundet. Ausgeschlossen wurden auch – anders als im Beitrag von Eisenberg in diesem Band – alle Eigennamen. Eigennamen sind zwar Wörter, aber man würde sie nur in Ausnahmefällen als lexikalische Einheiten ansehen, die den Ausdrucksreichtum des Deutschen vergrößern.

	1905–1914	1948–1957	1995–2004
Belletristik	52 700	57 400	57 000
Zeitungen	66 500	68 500	84 800
Wissenschaftliche Prosa	64 800	70 800	76 200
Gebrauchstexte	54 500	66 800	75 900

Auch hier fällt auf, dass die Belletristik den geringsten Wortschatz aufweist, die Zeitungen – außer seltsamerweise in der zweiten Zeitscheibe – den reichsten. Ebenso entwickelt sich der Wortschatz in Romanen und Erzählungen über

22 Bei den Zeitungen – nicht allerdings bei der wissenschaftlichen Literatur – ist auch die Zahl der Autoren größer als bei der Belletristik, d. h., der reichere Wortschatz wird teilweise auch einer gewissen Autorenspezifik geschuldet sein.

das ganze 20. Jahrhundert am schwächsten; in der zweiten Jahrhunderthälfte gibt es sogar einen leichten Abfall. Den im Ergebnis stärksten Ausbau der Lexik finden wir wiederum bei Zeitungen. Das hat seinen Grund weniger darin, dass die Schriftsteller sprachlichen Neuerungen abhold sind, sondern darin, dass in Zeitungen immer neue Themen auftauchen, und die erfordern neue Wörter. Erstaunlich und zumindest mir rätselhaft ist allerdings, dass es bei der Belletristik seit den 1950er Jahren offenbar keine Zunahme gegeben hat. Wohlgemerkt: Es geht hier um die Zunahme der Zahl nach. Es ist keineswegs gesagt, dass es sich um dieselben rund 57 000 Wörter handelt; der Wortschatz, so wie er sich bei den hier dokumentierten Autoren zeigt, kann sich also sehr wohl geändert haben.

Die häufigsten deutschen Wörter

Manche Wörter kommen nur in bestimmten Texten vor, weil sie für eine bestimmte Thematik wichtig sind, andere in allen, weil sie thematisch nicht gebunden ist. Dies sind vor allem Funktionswörter; deshalb sind die häufigsten deutschen Wörter allesamt Funktionswörter. Als Kandidaten für den ersten Platz gelten *die* und *und*. Der Vergleich zwischen beiden ist allerdings aus zwei Gründen etwas problematisch. Erstens ist *die* Teil eines kleinen Paradigmas, zu dem auch *der, das, des, dem* und *den* zählen, ohne dass man sie in der Tradition als gewöhnliche Flexionsformen eines einzigen Lexems betrachten würde. Zweitens kann *die* (ebenso wie *der, das, dem, den*), in verschiedenen Funktionen auftreten: als Artikel (*die Maus*), als Relativpronomen (*eine Maus, die*) und als eine Art Personalpronomen (*der Kater wollte eine Maus fangen, aber die war …*). Im Folgenden ist angegeben, wie oft die drei reinen Wortformen *der, die, das* (in gleich welcher Funktion), die Konjunktion *und* und die gleichfalls sehr häufige Präposition *in* vorkommen, und zwar im Kernkorpus (100 Millionen Textwörter) und in drei Zeitungskorpora: der *Zeit* (etwa 460 000 Millionen), in der *Süddeutschen Zeitung* (ebenfalls etwa 460 000 Millionen) und in zwei Berliner Zeitungen (*Tagesspiegel* und *Berliner Zeitung*, zusammengenommen etwa 420 Millionen). Ferner sind die Häufigkeiten für *ein* in allen Flexionsformen zusammen angegeben; dann schiebt es sich nämlich den Vorkommen nach zwischen die genannten Wörter (alle Zahlen sind auf Tausend gerundet):

	Kernkorpus	*Zeit*	*Süddeutsche*	Berliner Zeitungen
der	1 909 000	7 796 000	8 365 000	7 418 000
die	1 899 000	8 719 000	8 000 000	7 301 000
das	779 000	3 688 000	2 868 000	2 640 000
und	1 834 000	6 449 000	6 417 000	5 546 000
in	1 226 000	5 076 000	5 220 000	4 593 000
ein-	1 490 000	6 306 000	6 289 000	5 667 000

Demnach trägt im Kernkorpus, in der *Süddeutschen Zeitung* und in den Berliner Zeitungen *der* die Krone; in der *Zeit* hingegen liegt *die* an der Spitze, und zwar mit klarem Abstand. Die Unterschiede sind deutlich, und bei der großen Zahl ist nicht plausibel, dass sie reiner Zufall sind. Ein Blick auf die Belege von *der* zeigt, dass sehr viele davon gar nicht dem Maskulinum (*der Löffel*) geschuldet sind, sondern dem Genitiv im Singular oder Plural (*Chor der Gefangenen, Königin der Nacht*). Aber soll man annehmen, dass die Autoren der *Zeit* weniger Genitivattribute verwenden? Es ist dies nicht die einzige merkwürdige Schwankung: Im Vergleich zu den anderen Korpora kommt in der *Zeit* das Wort *das* sehr oft vor. Was mag das für Gründe haben? Umgekehrt schließt im Kernkorpus, und nur dort, die Konjunktion *und* dicht zu den beiden Führenden auf; offenbar gibt es in reinen Zeitungskorpora weniger *und* – vielleicht weil die Sätze im Schnitt insgesamt kürzer sind.

Es ist schon bemerkenswert, dass selbst derart riesige Korpora kein einheitliches Bild ergeben, und dies wohlgemerkt bei Wörtern, die thematisch nicht gebunden sind. Sicher sagen kann man nur, dass entweder *der* oder *die* an der Spitze liegen und dass *und* an dritter Stelle folgt.

Nun sind Funktionswörter, so wichtig sie sein mögen, nicht die Hauptträger der Information in einem Text. Wie steht es mit gehaltreicheren Wörtern, insbesondere den drei flektierenden Wortklassen Nomen, Verb und Adjektiv, aber auch mit den Adverbien? Im Folgenden sind jeweils die 20 häufigsten Wörter in den drei Zeitscheiben zusammengestellt. Flexionsformen sind zusammengerechnet. Wir beginnen mit den Nomina, die unter allen Wörtern die stärkste thematische Bindung aufweisen. Die häufigsten deutschen Nomina sind mit einigen wenigen Ausnahmen allerdings in dieser Hinsicht relativ neutral – sie könnten in den meisten Textsorten auftauchen.

1905–1914		1948–1957		1995–2004	
Jahr	15 376	*Mensch*	12 395	*Jahr*	22 221
Herr	15 196	*Jahr*	11 906	*Zeit*	9 650
Zeit	14 520	*Zeit*	11 321	*Frau*	8 709
Frau	13 199	*Frau*	9 238	*Mensch*	7 654
Mensch	11 549	*Herr*	8 302	*Tag*	7 398
Tag	10 685	*Tag*	7 540	*Prozent*	6 031
Wort	10 418	*Leben*	7 495	*Kind*	5 946
Leben	9 662	*Welt*	7 465	*Leben*	5 638
Mann	8 734	*Frage*	7 004	*Vater*	5 456
Kind	8 431	*Staat*	7 001	*Haus*	5 379
Gott	7 616	*Mann*	6 643	*Mann*	5 170
Auge	7 120	*Regierung*	6 566	*Ende*	5 105
Haus	6 745	*Kind*	5 918	*Welt*	4 852
Welt	6 450	*Wort*	5 617	*Frage*	4 532
Mutter	6 008	*Art*	5 387	*Mutter*	4 309
Frage	5 878	*Hand*	5 053	*Million*	4 266
Teil	5 414	*Auge*	4 996	*Teil*	4 247
Hand	5 394	*Teil*	4 955	*Auge*	4 065
Art	5 340	*Partei*	4 884	*Herr*	4 064
Ding	5 107	*Arbeit*	4 845	*Seite*	4 063

Man sieht sofort, dass es sich im Großen und Ganzen um dieselben Wörter handelt. Bei näherem Hinschauen zeigen sich doch einige eigentümliche Verschiebungen: *Gott*, zu Beginn des 20. Jahrhunderts an elfter Stelle, taucht am Ende dieses Jahrhunderts nicht einmal unter den hundert häufigsten Wörtern auf; es ist nach wie vor viel von der *Welt* die Rede, nicht aber von *Gott*. Auch der *Kaiser*, in der 1. Zeitscheibe an 57. Stelle, ist aus den hundert häufigsten Wörtern herausgefallen. Stattdessen haben sich *Prozent* und *Million* ins Vorfeld geschoben; keines davon findet sich in den beiden ersten Zeitscheiben unter den hundert häufigsten Nomina. Die gleichbleibend hohe Verwendung von *Frau* und *Herr* erklärt sich, wie ein Blick in die Belege zeigt, vor allem aus ihrem Gebrauch in *Frau Müller* oder *Herr Lehmann*. Auffällig ist freilich, dass *Herr* in der 1. Zeitscheibe über 15 000-mal vorkommt, während es sich in der 3. Zeitscheibe zwar nach wie vor häufig findet, aber doch auf rund 4 000 Belege gefallen ist. Überhaupt zeigt ein Blick auf die Zahlen, dass die Verschiebungen komplexer sind, als es zunächst den Anschein hat. So findet sich *Zeit* in allen Zeitscheiben unter den ersten drei Wörtern, es ist aber von 14 520 auf 11 321 und schließlich auf 9 650 gefallen; merkwürdiger ist dies noch bei *Frau*, das vom vierten auf den dritten Rangplatz gestiegen, aber dennoch deutlich seltener geworden ist: statt rund 13 200-mal kommt es nur noch rund 8 700-mal vor. Es gibt eine Reihe weiterer Verschiebungen, die zeigen, dass es offenbar jenseits „neuer“ und „verlorener“ Wörter eine erhebliche Entwicklung im nominalen Wortschatz über das 20. Jahrhundert gegeben hat.

Werfen wir nun einen Blick auf die 20 häufigsten Verben:

1905–1914		1948–1957		1995–2004	
sagen	28 875	*sagen*	23 268	*sagen*	23 061
kommen	24 394	*geben*	19 189	*geben*	16 991
machen	20 112	*sehen*	18 545	*kommen*	16 399
sehen	18 986	*kommen*	17 885	*sehen*	15 866
geben	18 146	*machen*	15 553	*gehen*	14 738
gehen	16 872	*gehen*	14 182	*machen*	14 601
stehen	13 078	*stehen*	12 827	*stehen*	11 103
finden	11 781	*bleiben*	9 632	*finden*	8 368
nehmen	10 659	*nehmen*	9 236	*bleiben*	8 143
bleiben	9 918	*liegen*	9 037	*liegen*	7 632
liegen	9 308	*finden*	8 089	*nehmen*	7 240
bringen	8 587	*tun*	7 484	*stellen*	5 947
halten	8 556	*lassen*	7 294	*gelten*	5 458
tun	8 551	*bringen*	6 537	*lassen*	5 425
lassen	8 471	*sprechen*	6 332	*tun*	5 398
sprechen	7 936	*stellen*	6 329	*zeigen*	5 359
treten	6 313	*halten*	6 325	*bringen*	5 239
glauben	6 008	*zeigen*	5 471	*halten*	5 015
stellen	5 994	*denken*	5 216	*setzen*	4 785
denken	5 860	*setzen*	4 875	*fragen*	4 771

All dies sind Allerweltsverben, und anders als bei den Nomina gibt es recht wenig Verschiebungen. Das Wort *denken*, in der 1. Zeitscheibe an 20. und in der 2. Zeitscheibe an 19. Stelle, ist in der 3. Zeitscheibe zwar aus der Tabelle herausgefallen, steht aber auf Platz 22 und ist damit nach wie vor sehr häufig. Es fällt allerdings auf, dass die absoluten Zahlen im Schnitt deutlich zurückgegangen sind: das häufigste Verb *sagen* von 28 875 auf 23 061, das jeweils zweithäufigste von 24 394 auf 16 991, das jeweilige Verb auf Rangplatz 20 von 5 860 auf 4 771. Das kann zwei Gründe haben: Der Anteil der Verben insgesamt ist zurückgegangen – vielleicht zugunsten der Nomina; dies würde die Idee einer zunehmenden Nominalisierung, wie sie von manchen Sprachkritikern beklagt wird („Substantivitis"), stützen. Oder aber der Anteil der Verben bleibt gleich, es werden jedoch mehr verschiedene Verben benutzt. Dies spräche für eine größere Differenzierung im Bereich der Verben und damit für eine Entwicklung, die man eher als positiv werten würde. Der Umstand, dass sich der deutsche Wortschatz insgesamt über das 20. Jahrhundert deutlich ausgeweitet hat, spräche für letztere Deutung; allerdings mag die Erweiterung vorwiegend neuen Nomina zu verdanken sein. Die Daten, so wie sie hier vorliegen, sind mit allen Deutungen zu vereinbaren, und so müssen wir die Frage hier offenlassen.

Die dritte Kategorie sind die Adjektive. Hier ergibt sich folgendes Bild:

1905–1914		1948–1957		1995–2004	
gut	15 390	*ander*	15 958	*ander*	15 020
neu	13 575	*gut*	14 960	*neu*	14 319
ander	12 822	*neu*	12 856	*gut*	13 090
ganz	11 446	*deutsch*	10 784	*ganz*	11 465
klein	11 195	*klein*	9 945	*erst*	11 386
alt	10 341	*erst*	9 416	*deutsch*	10 653
erst	10 291	*alt*	7 766	*weit*	8 009
deutsch	9 499	*weit*	7 730	*klein*	7 312
weit	8 864	*ganz*	7 093	*alt*	7 070
letzte	6 217	*eigen*	6 771	*eigen*	5 731
allgemein	6 166	*letzte*	5 822	*letzte*	5 227
jung	5 773	*politisch*	5 491	*lang*	4 360
eigen	5 482	*wirklich*	5 004	*politisch*	3 988
lang	5 402	*gleich*	4 902	*zweit*	3 926
verschieden	5 251	*hoch*	4 607	*hoch*	3 911
schwer	4 987	*jung*	4 599	*kurz*	3 815
wirklich	4 663	*lang*	4 500	*einfach*	3 805
hoch	4 634	*frei*	4 239	*schnell*	3 589
einzeln	4 587	*allgemein*	4 228	*wichtig*	3 506
kurz	4 386	*verschieden*	3 903	*jung*	3 316

Auch hier ist der Bestand weitgehend derselbe; das Adjektiv *politisch* taucht in der 1. Zeitscheibe zwar nicht unter den 20 häufigsten auf, steht dort aber immerhin auf Rang 40. Die drei häufigsten Adjektive sind *ander, gut, neu*. Gold, Silber und Bronze wechseln, aber die Häufigkeiten bleiben fast gleich. Bemerkenswert ist jedoch, dass auch hier die absoluten Werte im Schnitt deutlich zurückgehen, wenn auch weniger als bei den Verben. Wie bei diesen kann das daran liegen, dass der Anteil der Adjektive insgesamt in den verschiedenen Texten zurückgeht, oder aber daran, dass bei den Adjektiven stärker differenziert wird.

Kommen wir zum Schluss auf die Adverbien, deren Zuordnung zu Funktionswörtern oder Inhaltswörtern etwas schwankend ist. In der folgenden Liste sind nur „genuine“ Adverbien verzeichnet (wobei man sich bei manchen Wörtern, etwa *nur*, darum streiten kann, ob man sie als Adverb oder als Partikel betrachten soll; die Grammatiker vertreten hier etwas unterschiedliche Auffassungen); adverbial verwendete Adjektive sind bei den Adjektiven berücksichtigt.

1905–1914		1948–1957		1995–2004	
so	60 846	*auch*	58 493	*auch*	50 159
auch	60 792	*so*	47 143	*so*	35 945
nur	41 278	*nur*	37 624	*noch*	30 383
noch	38 790	*noch*	36 051	*nur*	29 637
aber	34 193	*aber*	26 460	*dann*	18 752
dann	21 219	*dann*	21 649	*aber*	17 384
doch	19 378	*schon*	16 619	*schon*	15 570
schon	18 743	*wieder*	15 889	*wieder*	14 231
wieder	16 920	*immer*	15 393	*immer*	13 973
da	16 741	*mehr*	14 302	*mehr*	10 728
sehr	16 330	*hier*	14 120	*jetzt*	9 512
nun	15 644	*sehr*	13 014	*da*	9 124
immer	14 847	*doch*	12 810	*hier*	9 020
hier	14 478	*selbst*	12 091	*doch*	8 962
ganz	14 386	*nun*	11 842	*selbst*	8 315
selbst	13 360	*ganz*	10 471	*nun*	7 844
mehr	13 068	*also*	10 254	*einmal*	6 854
jetzt	11 641	*da*	9 930	*sehr*	6 710
ja	9 686	*jetzt*	9 889	*erst*	6 654
einmal	9 397	*einmal*	9 583	*also*	6 547

Der Bestand ist im Großen und Ganzen derselbe, und auch die Reihenfolge bleibt weitgehend gleich. Auffällig sind wiederum einige quantitative Entwicklungen, die zu Fragen Anlass geben. Das Wort *aber* liegt in der 1. Zeitscheibe auf dem fünften Platz und in der 3. Zeitscheibe auf dem sechsten – kein großer Unterschied also; aber in der 1. Zeitscheibe kommt es 34 193-mal vor, in der 3. Zeitscheibe nur 17 384-mal, also kaum mehr als halb so oft. Offenbar war man vor dem Ersten Weltkrieg eher geneigt, Gegensätze herauszustellen, denn das ist ja die hauptsächliche Funktion von *aber*. Und wie bei Verben und Adjektiven liegt auch bei den Adverbien der Durchschnitt deutlich niedriger. Es gibt noch viel Stoff für Dissertationen, bei denen die Gefahr eines Plagiats sehr gering ist.

2 Verluste

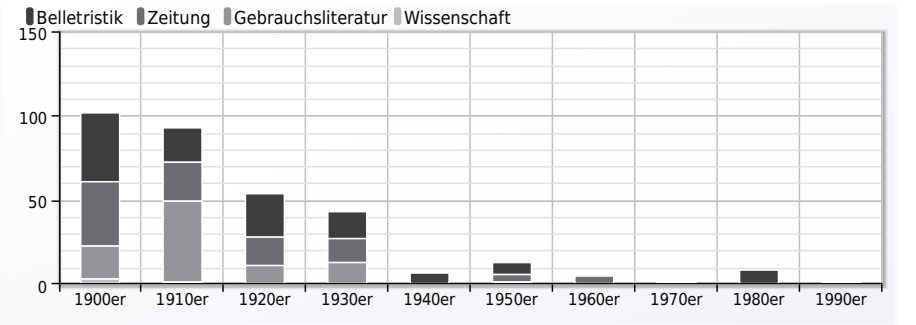
Insgesamt ist der deutsche Wortschatz im Verlaufe des 20. Jahrhunderts deutlich gewachsen. Dies schließt natürlich nicht aus, dass in diesen Jahren auch Wörter außer Gebrauch geraten sind oder dass ihnen dieses Schicksal bevorsteht, wenn sich niemand ihrer erbarmt und sie wieder benutzt.[23] Das heißt freilich nicht,

23 Mrozek (2008) ist eine unterhaltsame Sammlung solcher Wörter; viele davon sind allerdings

dass sie aus dem Wortschatz verschwunden sind, denn sie werden ja oft noch verstanden, und sie können benutzt werden, um dem Gesagten ein altertümliches Gepräge zu geben. In den gängigen Wörterbüchern heißt es oft, dass ein bestimmtes Wort „altertümlich, veraltet, veraltend" ist. Solche Angaben beruhen zumeist auf subjektiven Einschätzungen der Artikelverfasser (durchaus nichts Schlechtes), nicht auf Untersuchungen der tatsächlichen Verwendung. Im Folgenden wird das exemplarisch für einige typische Kandidaten nachgeholt. Dazu benutzen wir nicht nur die drei Zeitscheiben, sondern Wortverlaufskurven über das gesamte Kernkorpus hinweg (siehe www.dwds.de), eingeteilt in Intervalle von zehn Jahren: 1900–1909, 1910–1919, ..., 1990–1999.

Das erste Beispiel ist das Wort *Droschke*. Es kommt im Kernkorpus insgesamt 284-mal vor, und die Entwicklung zeigt genau das, was man erwarten würde – das Wort wird zunehmend seltener, so wie die Sache auf den Straßen.

Abb. 1 Wortverlauf für „Droschke" im DWDS-Kernkorpus



Ein Blick in umfängliche Zeitungskorpora zeigt allerdings, dass es auch heute noch verwendet wird, in der *Zeit* seit ihren Anfängen etwa 100-mal (gegenüber 4 700 Vorkommen von *Taxi* im selben Korpus). Schaut man sich nun die Belege selbst an, so sieht man rasch, dass es durchweg in historisierenden Kontexten oder in spezifischer Gestaltungsabsicht benutzt wird. Ein typisches Beispiel ist die folgende Stelle vom 16. 7. 2008:

> Schulz benutzt einzelne Erinnerungsbilder. Besonders wichtig war ihm das einer Droschke „mit aufgesetzten Kasten und brennenden Laternen", die in einen Wald hinausfährt. „Mir scheint", schreibt er 1935, „daß der ganze Rest des Lebens damit vergeht, diese Einblicke zu interpretieren ...".

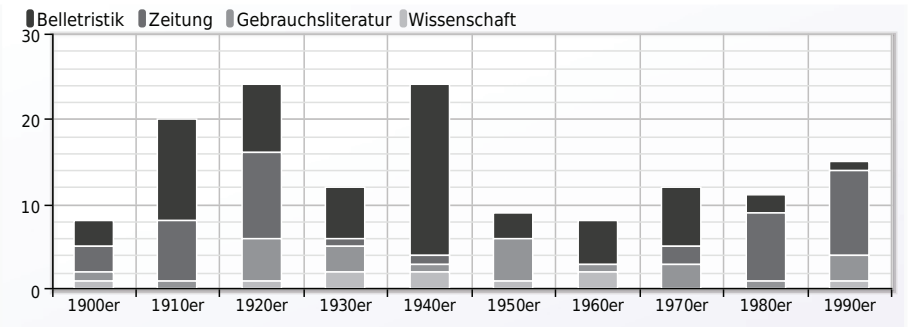
durchaus noch in neueren Texten zu finden. Systematische Untersuchungen auf breiter Datengrundlage gibt es meines Wissens bislang nicht.

Dasselbe beobachtet man für ein Wort wie *alldieweil*. Wo es überhaupt vorkommt – im Kernkorpus viermal, in der *Zeit* 25-mal –, geschieht dies durchweg in ironisierender oder historisierender Absicht. Hier ein Beispiel vom 25. 3. 1996:

> Biolek läßt mit pc-gebügelter „Ich verstehe alles"-Huld ein paar Nönnchen aufmarschieren, alldieweil ein geplagter Peter Rühmkorf bei *3 nach 9* erfahren darf, daß Hella von Sinnen einen Hammerzeh hat oder die unvermeidliche Hera Lind dem literarischen Rezept „Ich schreibe nicht mit dem Kopf, sondern aus dem Bauch" folgt.

Ähnliches gilt für Wörter wie *sintemalen, zuvörderst* oder *weiland*. Letzteres ist allerdings auffällig häufig. Im Kernkorpus findet es sich 109-mal, mit einer eigentümlichen Verlaufskurve.

Abb. 2 Wortverlauf für „weiland" im DWDS-Kernkorpus



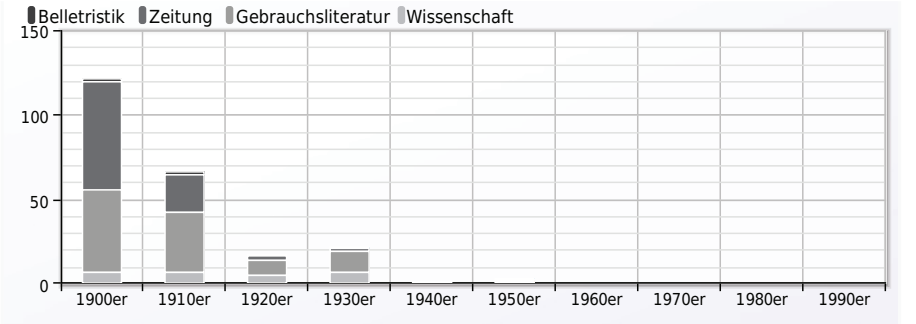
Da die Zahl der Belege insgesamt klein ist, kann sich schon ein einzelnes längeres Werk deutlich auswirken. Das ist hier für die 1940er Jahre der Fall: Die meisten Belege von *weiland* in diesem Jahrzehnt stammen aus Wolf von Niebelschütz' Roman *Der blaue Kammerherr*, der 1949 erschienen ist. Bemerkenswerterweise ist *weiland* wieder im Kommen. In der *Zeit* findet es sich beispielsweise nicht weniger als 1 800-mal, das ist etwa viermal auf eine Million Wörter. Hier ein typischer Beleg vom 12. 7. 2009:

> Mr. Steele redet sich derart um Kopf und Kragen, dass er einem schon fast wieder Leid tut. Er beschäftigt die Comedyshows wie weiland Sarah Palin.

Auch in anderen überregionalen Zeitungen zeigt sich diese Tendenz, wenn auch vielleicht weniger ausgeprägt: in der *Süddeutschen Zeitung* und der *Welt* jeweils etwa 1,8-mal auf eine Million, in *Bild* immerhin einmal auf zehn Millionen.

Ein letztes Beispiel ist die Präposition *behufs*, die für das ganze 20. Jahrhundert im Kernkorpus immerhin 220-mal belegt ist – allerdings mit einem klaren Abfall in den ersten vier Jahrzehnten.

Abb. 3 Wortverlauf für „behufs“ im DWDS-Kernkorpus



Vereinzelte Belege finden sich noch in den 1950er Jahren; sie sind jedoch zu selten, um im Diagramm sichtbar zu sein. So ist *behufs* sicher ein gutes Beispiel für ein Wort, das aktiv kaum noch gebraucht wird, während es zu Beginn des 20. Jahrhunderts in wissenschaftlicher Literatur und Gebrauchstexten durchaus noch rege benutzt wurde. Es scheint aber nicht vermisst zu werden. Es ist nämlich so: Wörter, die man braucht, verschwinden nicht.[24]

Wir haben hier nur einige wenige Beispiele für „Verluste“ im deutschen Wortschatz betrachtet. Sie machen aber deutlich, dass die Vorstellung verloren gehender Wörter oft trügerisch ist. Selbst Wörter wie *sintemalen, weiland* und auch *behufs* bleiben uns lange erhalten. Sie eröffnen besondere Gestaltungsmöglichkeiten, die es nicht gäbe, wenn die Wörter weiterhin fleißig gebraucht würden: Der Verlust ist auch ein Gewinn.

Gewinne

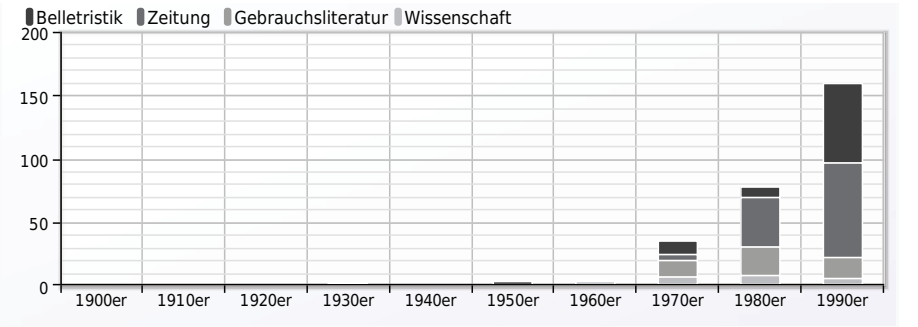
Sie sind, wie die statistischen Fakten zeigen, reich, und auch wenn die absoluten Zahlen sich bei weiteren Untersuchungen nicht genau so erhärten lassen sollten – grundsätzlich kann kein Zweifel bestehen, dass der deutsche Wortschatz sehr viel umfangreicher geworden ist. Nur sehr wenige Sprachen weisen einen solchen Reichtum in ihrer Lexik auf. Im Folgenden soll nun exemplarisch

24 An dieser Stelle gerät man ins Grübeln, warum ein Wort wie *behufs* mehr oder minder außer Gebrauch geraten ist. Braucht man es anders als früher nicht mehr? Es ist ja nicht so wie bei dem Wort *Droschke*, dessen Bedeutung aus unseren Straßen verschwunden ist. Ist *behufs* durch ein anderes Wort verdrängt worden, und weshalb das?

die Entwicklung einiger Wörter über das ganze letzte Jahrhundert betrachtet werden.[25]

Dabei schenken wir uns Wörter für Dinge, die es erst seit kurzem gibt und für die man erst seither ein Wort braucht, beispielsweise *Handy*, das erstmals in den 1980er Jahren verwendet wird und seit den 1990er Jahren in allen Korpora reich belegt ist. Ein interessanterer Fall ist *Sex*.

Abb. 4 Wortverlauf für „Sex“ im DWDS-Kernkorpus



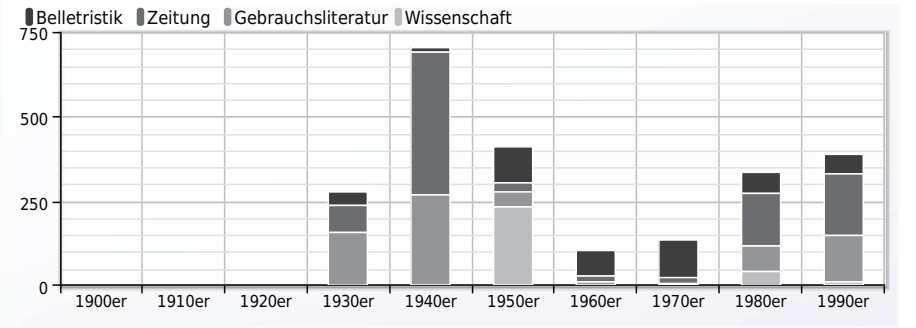
Hier ist seit den 1970er Jahren in Belletristik und Zeitungen ein neuer Ausdrucksbedarf entstanden, den man mit einem aus dem Englischen übernommenem Wort befriedigt. Es zeigt in charakteristischer Weise die ambivalente Natur solcher Übernahmen: Es ist eigentlich ja ein lateinisches Wort, und Wörter wie *Sexualität* tauchen schon lange vorher im Deutschen auf. Interessanterweise findet sich auch *sex appeal* schon wesentlich früher, wie es in dem Chanson von Friedrich Hollaender und Marcellus Schiffer aus den 1920er Jahren heißt: „Fast bin ich schon vom Sex Appeal das Geigentiel“. In seinen grammatischen und lautlichen Eigenschaften zeigt *Sex* keine Besonderheiten des Englischen, und für jemanden, der die Herkunft nicht kennt, ist es ein einfaches deutsches Wort wie jedes andere. Nicht anders ist es letztlich mit dem Wort *cool*, außer dass wir hier durch die Schreibweise an den Ursprung erinnert werden; für ein vierjähriges Kind ist, was wir *cool* schreiben, genauso ein deutsches Wort wie *toll*, und

25 Es gibt zwei umfassendere Dokumentationen des Zugewinns: Herberg et al. (2004) beschreiben in einer am Institut für Deutsche Sprache entstandenen Untersuchung rund 700 Neologismen der 1990er Jahre. Lothar Lemnitzer hat in seiner seit 2000 betriebenen „Wortwarte“ (www.wortwarte.de) bislang gut 25 000 Neologismen dokumentiert, die aus einer weitaus größeren Zahl von in verschiedenen Zeitungen belegten Wörtern ausgewählt sind; siehe auch Lemnitzer (2008).

wenn ein vierjähriges Kind schreiben könnte, würde es wahrscheinlich auch *kuhl* schreiben; man fragt sich, für wie viele Erwachsene, die kein Englisch gelernt haben, dies auch gilt.

Eine andere Art von nicht zusammengesetzten neuen Wörter sind Verkürzungen, wie etwa *Nazi*, dessen Gebrauch, so wie er sich in den Texten niederschlägt, ein etwas eigentümlicher ist.

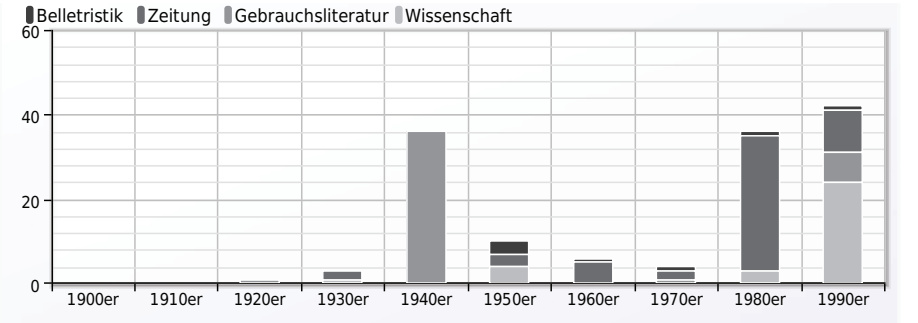
Abb. 5 Wortverlauf für „Nazi“ im DWDS-Kernkorpus



Es gibt bereits Belege aus den 1920er Jahren, die im Diagramm, da zu selten, nicht aufscheinen; in den 1930er und 1940er Jahren nimmt das Wort einen gewaltigen Aufschwung, wobei für letztere Dekade ein großer Teil den Jahren nach Kriegsende entstammt; in der Nazizeit war zumindest in den Zeitungen weitaus weniger von *Nazis* die Rede als danach. Dann sinkt *Nazi* stark im Vorkommen, um in den 1980er Jahren wieder eine Belebung zu erfahren. Über die Ursachen dieser eigentümlichen Entwicklung kann man nur spekulieren, insbesondere auch darüber, weshalb sie bei den verschiedenen Textarten so unterschiedlich ist. Hier müsste man sich die Texte im Einzelnen ansehen.

Sex ebenso wie *Nazi* sind kurze neue Wörter; die Mehrheit des Zuwachses entfällt aber, wie oben bemerkt, auf Ableitungen und Komposita. Unter allen deutschen Neubildungen, die das 20. Jahrhundert gezeitigt hat, zeigt wohl keines die Kluft zwischen dem, was ein Kompositum aufgrund seiner Bestandteile bedeutet, und dem, was man damit tatsächlich meint, so krass wie jenes, das aus den einfachen deutschen Wörtern *Ende* und *Lösung* besteht.

Abb. 6 Wortverlauf für „Endlösung“ im DWDS-Kernkorpus



Die seltenen frühen Belege (der älteste stammt noch aus den 1920er Jahren) haben nicht die Bedeutung, unter der wir alle dieses Wort heute verstehen. In dieser tritt es erst nach dem Krieg auf; die große Häufigkeit in den 1940er Jahren rührt großenteils aus den „Nürnberger Protokollen“, und in den 1 321 Belegen aus der *Zeit* seit 1946 hat es nur noch diese Bedeutung. Bemerkenswert ist übrigens, dass in der *Süddeutschen Zeitung* und in den beiden Berliner Zeitungen (*Tagesspiegel* und *Berliner Zeitung*), für die wir nur Daten seit den 1990er Jahren haben, das Wort viel seltener geworden ist.

3 Rätsel

Es gibt, wie schon bemerkt, Wörter, die thematisch gebunden sind, und solche, die wenig eigenen Inhalt haben und sich daher in allen Texten finden – Funktionswörter wie *der, die, das, so, wenn, auf, dass* und dergleichen mehr. Dies schließt nicht aus, dass es dennoch Schwankungen in der Häufigkeit gibt. Für narrative Texte, auch dies wurde schon bemerkt, ist vielleicht *dann* eher typisch als für wissenschaftliche Prosa, für *weil* mag das Umkehrte gelten, weil es bei ersteren eher um temporale, bei letzteren eher um kausale Zusammenhänge geht. Wie ist es nun tatsächlich? Der Leser möge einen Augenblick innehalten und nachdenken, bevor er sich das Diagramm anschaut.

Es ist nämlich nicht so: sowohl *dann* als auch *weil* sind in der Belletristik weitaus häufiger. Ist dies schon eigentümlich, so ist noch weitaus rätselhafter, wieso beide über das 20. Jahrhundert so unterschiedlich gern gebraucht werden. Die 1940er Jahre waren in der Belletristik eine *dann*-arme Zeit, umgekehrt zeigen die Journalisten in den 1920er Jahren und noch einmal in den 1980er Jahren eine gewisse Liebe für das Kausale. Bei kleinen Belegzahlen können schon einzelne

Texte einen erheblichen Unterschied ausmachen; das ist hier aber nicht der Fall: Es geht jeweils um Schwankungen von mehreren Tausend Vorkommen.

Abb. 7 Wortverlauf für „dann“ im DWDS-Kernkorpus

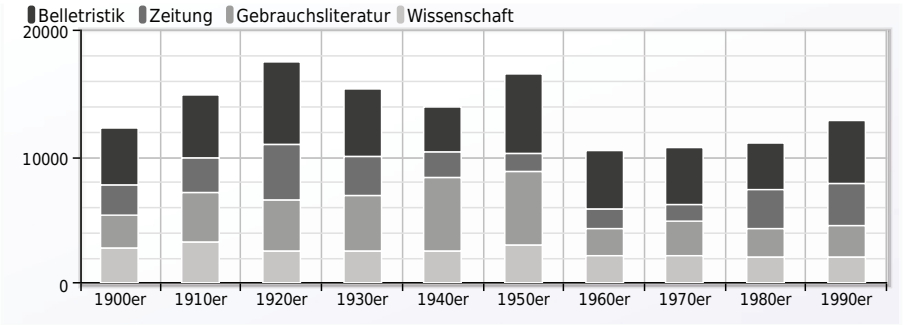
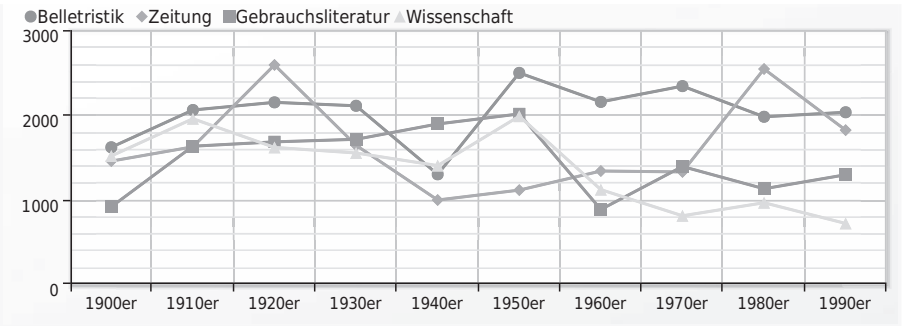


Abb. 8 Wortverlauf für „weil“ im DWDS-Kernkorpus

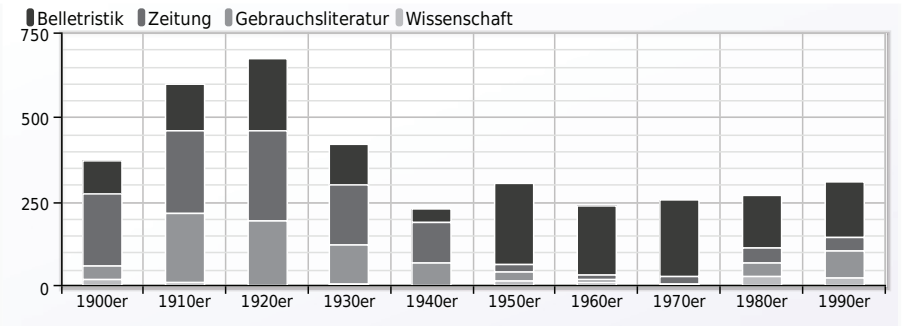


Dies ist einer von vielen merkwürdigen Befunden, auf die man stößt, wenn man das Schicksal einzelner Wörter über die Zeit verfolgt. Ein weiteres Rätsel wirft beispielsweise das sehr häufige Funktionswort *so* auf; es ist kein Grund zu erkennen, weshalb es in den einzelnen Textsorten unterschiedlich häufig vorkommen sollte. In literarischen Texten ist es aber weitaus häufiger als in allen anderen Textsorten. Noch merkwürdiger ist, dass von 1900 bis 1930 pro Jahrzehnt im Schnitt 40 000-mal vorkommt, von 1960 bis 1990 im Schnitt hingegen 20 000-mal. In den 1990er Jahren steigt die Zahl dann wieder. All dies gilt für ähnlich zusammengesetzte Textkorpora, die jeweils etwa 10 Millionen Wörter lang sind. Diese Zahlen sind viel zu hoch, als dass man von einer zufallsbedingten Schwankung sprechen könnte.

Ich schließe diese kleine Rätselsammlung mit einem letzten kuriosen Faktum. Diesmal geht es nicht um ein Funktionswort, sondern ein Inhaltswort, das jeder

kennt und von dem man nicht annimmt, dass es thematisch stark gebunden ist. Als vor fast 60 Jahren einige französische Sprachforscher erstmals versucht haben, den „Grundwortschatz“ einer Sprache nicht allein aufgrund des Bauchgefühls zu bestimmen, sondern auszuzählen, wie oft die Wörter tatsächlich verwendet werden, da hat sich zu ihrem Erstaunen ergeben, dass man im Französischen das Wort *la gare* offenbar nicht benutzt; es kam nämlich in den durchaus planvoll zusammengestellten Korpora kaum vor (Gougenheim et al. 1956). Wie ist das mit dem deutschen Wort *Bahnhof*?

Abb. 9 Wortverlauf für „Bahnhof“ im DWDS-Kernkorpus



Es kommt vor, im Schnitt 370-mal pro Jahrzehnt. Relativ selten ist es nur in der wissenschaftlichen Prosa. Merkwürdig ist nun jedoch, dass es sich in der ersten Jahrhunderthälfte in den Zeitungen sehr oft findet, in der zweiten hingegen sehr wenig: Der Abfall beginnt in den 1930er Jahren. In der Belletristik ist es hingegen sowohl am Ende wie am Beginn sehr häufig – nur von 1930 bis 1950 nicht. Besonders in den 1940er Jahren ist es eher selten. Man beachte, dass es sich dabei um den Durchschnitt handelt – es kann also sein, dass es nach Kriegsende (*Der Zug war pünktlich* erschien 1949) schon wieder auf dem Anstieg war.

4 Schluss

Wie in den einleitenden Abschnitten erläutert wurde, wissen wir bislang sehr wenig Gesichertes über den tatsächlichen Reichtum des deutschen Wortschatzes. Das ist kein Zufall, und so sind denn die hier berichteten Ergebnisse mit mancherlei Unsicherheiten behaftet, die sich wohl so bald nicht werden beseitigen lassen. Die wichtigsten Hemmnisse, die sich wohlfundierten Aussagen in den Weg stellen, sind in den ersten Abschnitten genannt worden. Das, was in den weiteren

Abschnitten dann doch gesagt wird, ist nicht zuletzt auch als eine Aufforderung zum Tanz anzusehen: Mögen andere kommen und das hier Gesagte verfeinern oder auch widerlegen, wie es der erwünschte Gang der Wissenschaften ist.

Immerhin, zwei Befunde zeichnen sich sehr klar ab, und sie werden sich auch bei weiteren Untersuchungen nicht ändern:

1. Die heutige deutsche Sprache verfügt über einen überaus reichen Wortschatz, der weit jenseits dessen liegt, was je in einem Wörterbuch beschrieben worden ist.

2. Der Wortschatz, so wie er in seinem Gebrauch in großen Textkorpora fasslich wird, ist im Verlauf der letzten hundert Jahre um mindestens eine Million Wörter angewachsen.

Da der Ausdrucksreichtum einer Sprache letztlich auf ihrem Wortschatz fußt, muss man schließen, dass sich das Deutsche in dieser Zeit zu einem immer mächtigeren Instrument entwickelt hat. Wenn es uns bisweilen so scheint, als würde unsere Sprache verarmen, dann liegt das nicht an der deutschen Sprache, sondern an denen, die von ihr Gebrauch machen. Es reicht nicht, einen Bösendorfer in der Stube stehen zu haben; man muss ihn auch spielen können.

5 Literatur

Augst, Gerhard (2009): *Wortfamilienwörterbuch der deutschen Sprache*. 2. Aufl. Tübingen: Niemeyer.

Best, Karl-Heinz (2006): *Quantitative Linguistik. Eine Annäherung*. 3. Aufl. Göttingen: Peust & Gutschmidt.

Burger, Harald et al. (Hrsg.) (2007): *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin, New York: de Gruyter.

Didakowski, Jörg & Alexander Geyken (2013): From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions. In: *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network „Internet Lexicography"*. Mannheim: Institut für Deutsche Sprache (OPAL X/2012), S. 43–52.

Ďurčo, Peter (Hrsg.) (2010): *Feste Wortverbindungen und Lexikographie. Kolloquium zur Lexikographie und Wörterbuchforschung*. Berlin, New York: de Gruyter.

Engelberg, Stefan & Lothar Lemnitzer (2009): *Lexikographie und Wörterbuchbenutzung*. 4. Aufl. Tübingen: Stauffenberg.

Fleischer, Wolfgang & Irmhild Barz (2012): *Wortbildung der deutschen Gegenwartssprache*. 4. Aufl. Berlin, New York: de Gruyter.

Gougenheim, Georges et al. (1956): *L'élaboration du français élémentaire. Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.

Le Grand Robert de la langue française (2001). Paris: Le Grand Robert.

Jacob Grimm (1848): *Geschichte der deutschen Sprache*. Bd. 1. Leipzig: Weidmannsche Buchhandlung.

Das große Wörterbuch der deutschen Sprache (1999). 3. Aufl. Mannheim: Duden.

Haß-Zumkehr, Ulrike (2001): *Deutsche Wörterbücher – Brennpunkt von Sprach- und Kulturgeschichte*. Berlin, New York: de Gruyter.

Heaps, Harold Stanley (1978): *Information Retrieval. Computational and Theoretical Aspects*. New York: Academic Press.

Herberg, Dieter et al. (2004): *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. Berlin, New York: de Gruyter.

Herdan, Gustav (1960): *Type-token mathematics*. Den Haag: Mouton.

Klein, Wolfgang & Harald Zimmermann (1971): *Lemmatisierter Index zu Georg Trakls Werken*. Tübingen: Niemeyer.

Köhler, Reinhard (2005): Statistische Methoden in der Lexikologie. In: D. Alan Cruse et al. (Hrsg.): *Lexikologie. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*. 2. Halbbd. Berlin: de Gruyter, 953–963.

Köhler, Reinhard et al. (Hrsg.) (2008): *Quantitative Linguistik*. Berlin, New York: de Gruyter.

Küpper, Heinz (1955–1970): *Wörterbuch der deutschen Umgangssprache*. 6 Bde. Hamburg: Claassen.

Lemnitzer, Lothar (2008): *Hirndiebstahl im Sparadies: Was so (noch) nicht im Duden steht*. Mannheim: Bibliographisches Institut.

Mrozek, Bodo (2008): *Das große Lexikon der bedrohten Wörter*. 2 Bde. Hamburg: Rowohlt.

Ruoff, Arno (1990): *Häufigkeitswörterbuch gesprochener Sprache. Gesondert nach Wortarten alphabetisch, rückläufig-alphabetisch und nach Häufigkeit geordnet*. 2. Aufl. Tübingen: Niemeyer.

Schares, Thomas (2006): *Untersuchungen zu Anzahl, Umfang und Struktur der Artikel der Erstbearbeitung des Deutschen Wörterbuchs von Jacob Grimm und Wilhelm Grimm*. Phil. Diss. Universität Trier.

Schippan, Thea (1992): *Lexikologie der deutschen Gegenwartssprache*. 2. Aufl. Tübingen: Niemeyer.

Schlaefer, Michael (2009): *Lexikologie und Lexikographie. Eine Einführung am Beispiel deutscher Wörterbücher*. 2. Aufl. Berlin: Schmidt.

Schneider-Wiejowski, Karina (2011): *Produktivität in der deutschen Derivationsmorphologie*. Phil. Diss. Bielefeld.

Splett, Jochen (2009): *Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes*. Berlin, New York: de Gruyter.

Tudjman, Miroslav (2005): Gesetz zur Bestimmung des Wortschatzumfangs von Texten. Das Heaps'sche Gesetz und die Bestimmung der Wortschatzgröße in kroatischen Texten [kroatisch mit deutscher Zusammenfassung]. *Društvena istraživanja* 14, No.1-2, 227–250.

5.1 Websites

Im Folgenden geben wir jeweils die Startseite an, unter der weitere, spezifischere Links zu finden sind; im Text ist jeweils angegeben, wann eine Website abgerufen wurde.

Corpora from the Web an der FU Berlin:
http://hpsg.fu-berlin.de/cow/

Deutscher Wortschatz an der Universität Leipzig
http://wortschatz.uni-leipzig.de/

Digitales Wörterbuch der deutschen Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften
http://www.dwds.de/

Goethe-Wörterbuch
http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/gwb/

Institut für Deutsche Sprache, Mannheim
http://www1.ids-mannheim.de/start/

Oxford English Dictionary
http://www.oed.com/

Wortwarte (deutsche Neologismen)
http://www.wortwarte.de/