Häufigkeitsverteilungen im Deutschen und ihr Einfluss auf den Erwerb des Deutschen als Fremdsprache

Prof. Dr. Erwin Tschirner
Herder-Institut
Universität Leipzig
Beethovenstraße 15
04107 Leipzig

Abstract
The last major effort to develop a frequency dictionary of German was made in the 1960s when Alan Pfeffer developed his Basic (Spoken) German Word List. Most vocabulary compilations for German are still based on the first, and until recently only, frequency dictionary based on a structured, representative corpus of German in the German speaking countries published in 1897. This article will discuss the development of a new frequency dictionary of German based on a modern, spoken and written corpus of the DACH countries. It will then focus on three questions related to the teaching of German. (1) What is the relationship between the new frequency dictionary and other basic vocabulary lists? (2) What text coverage is provided by the most frequent 4000 words for text genres such as spoken language, newspaper, literary, and academic writing? (3) What insights can frequency studies provide for the teaching of German grammar?

1 Einführung

Seit geraumer Zeit interessieren sich Sprachwissenschaftler und Fremdsprachenerwerbsforscher (Ellis 1996, Pawley/Syder 1983, Sinclair 1991, Weinert 1995) und zunehmend auch Fremdsprachendidaktiker (Nattinger/DeCarrico 1992, Nation 2001) für Häufigkeitsverteilungen in zu lernenden Sprachen. Zum einen beeinflussen Häufigkeitseffekte den Zweitsprachenerwerb in gesteuerten und ungesteuerten Erwerbssituationen, sowohl im Lexikerwerb wie im Aufbau fremdsprachlicher Phonologie und Syntax (Ellis 2002), zum anderen beeinflusst die Variable "Wortschatzgröße" nicht nur mit wie viel Verständnis und wie fließend gelesen werden kann, sondern auch ob unbekannte Wörter aus dem Kontext geraten oder implizit gelernt werden können (Nation 2001). Um so erstaunlicher ist es, dass das letzte (und erste) empirisch erarbeitete Häufigkeitswörterbuch der gesamten deutschen Sprache (Kaeding 1891) bereits mehr als 100 Jahre alt ist. Die für Deutsch gültigen Grund- und Aufbauwortschätze basieren zum großen Teil auf den Ergebnissen von Kaeding und bedürfen dringend einer Anpassung an die heutige gesprochene und geschriebene Sprache in den deutschsprachigen Ländern. Dieser Beitrag befasst sich auf der Grundlage eines neuen Häufigkeitswörterbuch der deutschen Sprache (Jones/Tschirner 2006) zusammen mit dem Kor-

pus, auf dem es basiert, mit drei Fragestellungen. 1. Welche Übereinstimmungen gibt es zwischen den Standardgrund- und Aufbauwortschätzen des Deutschen als Fremdsprache und dem neuen Häufigkeitswörterbuch für Deutsch als Fremdsprache? 2. Welchen Wortschatzbedarf weisen einzelne Textsorten im Deutschen auf? 3. Welche Grammatikprogressionen für Lehrwerke lassen sich auf der Grundlage von Korpusdaten erarbeiten?

Ich gehe zuerst auf einige Aspekte einer häufigkeitsorientierten Fremdsprachendidaktik ein, insbesondere im Hinblick auf das Lesen, und befasse mich dabei in Sonderheit mit dem Wortschatzbedarf unterschiedlicher Textsorten und mit der Frage des direkten und indirekten Lernens von Wörtern. Als nächstes stelle ich das Herder/BYU-Korpus (Tschirner/Jones 2005) vor und skizziere die Erarbeitung des darauf aufbauenden Häufigkeitswörterbuchs (Jones/Tschirner 2006). Im Weiteren vergleiche ich die für die Lehrwerkerstellung des Deutschen maßgeblichen Grund- und Aufbauwortschatzes von Langenscheidt mit dem neuen Häufigkeitswörterbuch und gehe auf einige Probleme in diesem Zusammenhang ein. Anschließend stelle ich den Wortschatzbedarf unterschiedlicher Textsorten im Deutschen vor und vergleiche diesen mit dem aus der Literatur bekannten Wortschatzbedarf für das Englische. Schließlich spreche ich in einem Ausblick weitere Fragestellungen im Zusammenhang mit Häufigkeitsverteilungen an, vor allem solche, die mit Fragen der Grammatik im Lexikon und damit zusammenhängend mit dem Grammatikerwerb zu tun haben.

2 Aspekte einer häufigkeitsorientierten Fremdsprachendidaktik

Häufigkeitseffekte spielen eine große Rolle im Spracherwerb, sowohl im Erstsprach- wie im Fremdsprachenerwerb. Häufigere Wörter und Strukturen werden schneller und fehlerfreier erkannt, früher von Kindern erworben und sie sind weniger von Störungen bei Aphasie betroffen (Fenk-Oczlon, 2001: 435). Im Fremdsprachenerwerb beeinflussen sie den Lexikerwerb wie den Aufbau fremdsprachlicher Phonologie und Syntax (Ellis 2002). Eine Frage, mit der man sich schon seit geraumer Zeit beschäftigt, ist, wie viele der häufigsten Wörter man beherrschen muss, um unbekannte Texte relativ zügig lesen zu können.

Dass häufige Wörter den Großteil eines geschriebenen (oder gesprochenen) Textes ausmachen, ist nicht neu. Carroll/Davies/Richman (1971) erstellten für ihr Häufigkeitswörterbuch des amerikanischen Englisch ein Korpus von 86.741 laufenden Wörtern (Tokens)[1] und fanden heraus, dass die häufigsten 2000 Wortformen (Types) etwas über 80 Prozent der gesamten Wortformen ihres Korpus ausmachten und die häufigsten 5000 Wortformen fast 90 Prozent. Nach Nation (2001) erfassen die häufigsten 2000 Lexeme ca. 90 Prozent der Tokens von Alltagsgesprächen, ca. 87% von literarischen Texten und ca. 80% von Zeitungstexten. Fügt man die 570 Lexeme des Akademischen Wortschatzes (Coxhead 2000) hinzu, erreicht man ca. 84% von Zeitungstexten und 87% von Fachtexten.

Entgegen den Annahmen der Lesedidaktik der 70-er und 80-er Jahre des letzten Jahrhunderts haben empirische Untersuchungen ergeben, dass nicht nur 60-70 Prozent der Tokens

[1] Unter Tokens versteht man die laufenden Wörter eines Textes (die Gesamtzahl der Wörter) und unter Types die individuell vorkommenden Wortformen. Lexeme sind mit Wörterbucheinträgen vergleichbar und beinhalten alle dazugehörigen konjugierten und deklinierten Formen.

eines Textes bekannt sein müssen, um ihn zu verstehen, sondern 95% und mehr (Carver 1994, Laufer 1997, Hu/Nation 2000, Qian 2002). Um Wörter aus dem Kontext erraten zu können und um unbekanntes Vokabular auf implizite Art und Weise zu lernen – wichtige Ziele der fremdsprachlichen Lesedidaktik – müssen sogar mindestens 97% der Tokens eines Textes verstanden werden (Swanborn/de Glopper 1999). Nach Laufer (1997) benötigt man für das Englische einen Wortschatz, der die 5000 häufigsten Lexeme umfasst, um 95% der Tokens eines durchschnittlichen Zeitungstextes oder Fachaufsatzes zu verstehen und um ihn damit mit Verständnis und ohne größeren Zeitverlust lesen zu können.

Muttersprachler erwerben zwar ab einem gewissen Grundwortschatz von ca. 5000 Lexemen die meisten weiteren Wörter ihrer Muttersprache – ca. 1000 pro Jahr – über das Lesen, doch ist dies mit einem hohen Leseaufwand verbunden. Auf Grund von Studien, die ergeben, dass ca. 15% der unbekannten Wörter eines Textes gelernt werden, aber nur wenn sie ca. 16 Mal innerhalb einer bestimmten Zeitspanne angetroffen werden (Swanborn/de Glopper 1999), schätzt Nation (2001), dass man ca. 1 Mio. Tokens lesen muss, um 1000 Wörter indirekt, über das Lesen, zu lernen. Das sind etwa 10-12 Unterhaltungsromane pro Jahr. Hinzu kommt, dass man bereits einen Grundwortschatz von ca. 5000 Lexemen haben muss, den sich Fremdsprachenlerner ja erst anlernen müssen. Aus diesem Grunde argumentiert Nation, dass die häufigsten Lexeme einer Sprache direkt gelernt werden und gleichzeitig in allen Modalitäten in handlungsorientierten Zusammenhängen erfahren werden müssen, damit sich das fremdsprachliche Lexikon so weit entwickeln kann, dass wie bei Muttersprachlern ein Großteil des weiteren Wortschatzerwerbs implizit über das Lesen vonstatten gehen kann.

Für den Wortschatzerwerb besteht nach Nation (2001) der ideale Unterricht deshalb aus vier zeitlich relativ gleich umfangreichen Blöcken: direktes Lernen, kommunikative Erfahrungen im Hören und Sprechen, kommunikative Erfahrungen im Lesen und Schreiben und Flüssigkeitstraining. Unter Flüssigkeitstraining versteht Nation zusätzliche, handlungsorientierte, kommunikative Erfahrungen mit Wörtern, die man bereits kennt, in neuen, ungewohnten Kontexten oder unter Zeitdruck. Beim direkten Lernen ist für ihn das wichtigste Kriterium der Kosten-Nutzen-Vergleich, d. h. wie viel kostet das Lernen eines Wortes und wie groß ist der Nutzen, den man von einem bestimmten Wort hat. Aus diesem Grund bezeichnet er es als wenig sinnvoll, Wörter willkürlich zu lernen, sondern schlägt vor, dass man sich vor allem auf die häufigsten Wörter einer Sprache konzentrieren sollte. Dazu bedarf es zuverlässiger und aktueller Häufigkeitswörterbücher.

3 Häufigkeitswörterbücher und Grundwortschätze des Deutschen (als Fremdsprache)

Das älteste Häufigkeitswörterbuch der deutschen Sprache geht auf Kaeding (1897) zurück, der ein für damalige Verhältnisse kaum fassbares Korpus von zehn Mio. Tokens erstellte, um für das Deutsche ein Stenographiesystem zu entwickeln. Das Korpus bestand ausschließlich aus schriftlichen Texten und enthielt literarische Werke, Zeitungstexte, amtliche Bekanntmachungen und Verordnungen, Geschäftsbriefe und Fachliteratur. Das Wörterbuch enthält eine Häufigkeitsliste der Wortformen und nicht der Lexeme, da es Kaeding vor allem darum ging, häufige Graphemverbindungen zu erkennen, um ein stenographisches Kürzelsystem zu entwickeln.

Ortmann (1975) entwickelte auf der Basis von Kaedings Wortliste seine hochfrequenten deutschen Wortformen, die als Grundlage des Zertifikats Deutsch als Fremdsprache (später Zertifikat Deutsch) die Lexikauswahl deutscher DaF-Lehrwerke beeinflusste. Selbst für *Profile Deutsch* (Glaboniat, Müller, Rusch 2002) und damit für die jetzige Generation von DaF-Lehrwerken wurde keine neue Untersuchung gestartet, sondern die alten Wortschatzlisten des Goethe-Instituts weiterverwendet. Somit beruht ein Großteil der Lexikauswahl aktueller Grund- und Aufbauwortschätze wie auch von Lehrwerken auf einer Zählung, die mehr als 100 Jahre zurückliegt.

Neben Kaeding (1897) gibt es weitere Häufigkeitswörterbücher der geschriebenen Sprache, deren Korpora aber deutlich weniger Tokens enthalten und die keinen Anspruch auf eine repräsentative Auswahl von Texten deutscher Sprache erheben können. Rosengren (1972) und Swenson (1967) erarbeiteten Häufigkeitslisten der Zeitungssprache, Scherer (1965) der deutschen Kurzgeschichte.

Pfeffer erarbeitete in den Jahren 1960-61 ein strukturiertes, repräsentatives Korpus der überregionalen Umgangssprache der Bundesrepublik Deutschland, der Deutschen Demokratischen Republik, Österreichs und der Schweiz, bestehend aus 398 Texten mit ca. 650.000 Tokens. Daraus entwickelte er ein Häufigkeitswörterbuch der gesprochenen Sprache (Pfeffer 1964; 1975), das insbesondere in den USA Grundlage didaktischer Entscheidungen in Deutschlehrwerken wurde, dessen Ergebnisse aber auch Eingang in die Wortlisten des Goethe-Instituts fanden und damit auch in DaF-Lehrwerke. Eine weitere Häufigkeitswortliste der gesprochenen Sprache wurde von Ruoff (1981) entwickelt. Seine Wortliste basiert auf mehreren umfangreichen Dialektkorpora des schwäbischen Dialekts (ca. 1200 Texte), aufgenommen in den Jahren 1955-1974. Aus diesem Gesamtkorpus wurde ein Subkorpus von 500.000 Tokens ausgewählt. Diese wurden lemmatisiert und in eine Häufigkeitsreihenfolge gebracht.

Zusammenfassend kann man sagen, dass fast alle Korpora der deutschen Sprache, auf denen Häufigkeitswörterbücher und Grund- und Aufbauwortschätze basieren, veraltet sind und viele davon wenig repräsentativ für die deutsche Sprache in ihrer Gesamtheit im gesamten deutschsprachigen Raum sind. Das folgende Kapitel beschäftigt sich mit dem seit Kaeding ersten Versuch, ein ausgewogenes Korpus der aktuell verwendeten Sprache des gesamten deutschen Sprachraums zu entwickeln, und mit dem daraus abgeleiteten Häufigkeitswörterbuch (Jones/Tschirner 2006).

4 Das Herder/BYU-Korpus und das Frequency Dictionary of German

Das Herder/BYU-Korpus (Tschirner / Jones 2005) ist ein aktuelles, ausgewogenes und für die deutschsprachigen Länder repräsentatives Korpus mit einem Umfang von 4,2 Mio. Tokens. Es weist jeweils 1 Mio. Tokens aus Zeitungstexten, Sach- und Fachtexten, literarischen Texten und gesprochener Sprache auf und 200.000 aus Gebrauchstexten. Es wurde darauf geachtet, dass in allen Subkorpora das Verhältnis zwischen deutschen, österreichischen und Schweizer Texten ca. 70: 20: 10 besteht. Die vier Hauptsubkorpora teilen sich in jeweils 100 Texte zu je 10.000 Tokens auf. Bei längeren Texten, wie z.B. Büchern, wurden jeweils die ersten 3000, die mittleren 4000 und die letzten 3000 Tokens gesammelt.

Die Texte wurden teilweise aus dem Internet gewonnen, teilweise wurden sie eingescannt. Alle Texte wurden mehrmals Korrektur gelesen, um Rechtschreibfehler zu entfernen und um die Rechtschreibung zu vereinheitlichen. Anschließend wurden die Wörter des gesamten Korpus mit dem Stuttgarter Tree Tagger (Schmidt 1995) und dem Stuttgart-Tübingen Tag-Set (STTS) mit Wortartannotierungen ergänzt. Für alle weiteren Arbeitsschritte wurden die Wordsmith Tools Version 3 (Scott 1999) benutzt. Dazu gehörten vor allem die Disambiguierung von Wortformen und die Lemmatisierung. Beides wurde größtenteils per Hand erstellt. Ein großes Problem stellten dabei die trennbaren Präfixverben dar, die zeitaufwändig ebenfalls per Hand ermittelt wurden.

Insgesamt ergeben sich aus den 4,2 Mio. Tokens des Herder/BYU-Korpus 270.000 Types. Davon tauchen 154.430 (ca. 57%) der Types nur einmal auf. Lemmatisierung bewirkt für die häufigsten 10.000 Lexeme eine Reduzierung der Types um ca. 40 Prozent, von 16.500 auf 10.000. Eigennamen machen ca. 5,8 Prozent der häufigsten 4000 Lexeme aus, davon 1,6% der häufigsten 1000 Lexeme, 5% der Lexeme von 1000 bis 2000 und jeweils 8,2% der Lexeme von 2000-3000 und 3000-4000.

Das daraus resultierende Häufigkeitswörterbuch (Jones/Tschirner 2006) erfasst die häufigsten 4000 Lexeme der deutschen Sprache. Das 4000. Lexem kommt im Korpus dabei noch 68 mal vor. Um eine zuverlässige Liste der häufigsten 4.000 Lexeme des Deutschen zu erstellen, wurden die 8.000 häufigsten Types lemmatisiert. Dies ist nötig, da Lemmatisierung, vor allem bei Verben, aber auch bei Substantiven und Adjektiven eine deutliche Erhöhung der Häufigkeit und damit ein Hinaufrutschen in der Häufigkeitsliste nach sich zieht.

5 Vergleich Grund- und Aufbauwortschatz Langenscheidt und Häufigkeitswörterbuch

Eine Reihe von Studien haben gezeigt, dass ein Wortschatz von 4000-5000 der häufigsten Wörter von großer Bedeutung für das Lesen in der Fremdsprache und darüber hinaus für den weiterführenden impliziten Wortschatzerwerb ist, der vor allem über das Lesen stattfindet. Wichtig dabei ist, dass es sich bei diesen Wörtern um die häufigsten 4000-5000 Wörter einer Sprache handelt und nicht um x-beliebige, die zum Teil weit jenseits der Fünf- oder Zehntausenderschwelle angesiedelt sind und so selten sind, dass sie beim zweiten Auftauchen schon wieder vergessen wurden. Grund- und Aufbauwortschätze des Deutschen enthalten meist 4000 Wörter, so z.B. der Grund- und Aufbauwortschatz Deutsch von Langenscheidt (in der aktuellsten Version *Basic German Vocabulary* von 1991).

Ein Vergleich des neuen Häufigkeitwörterbuchs von Routledge (Jones/Tschirner 2006) mit dem häufig benutzten Grund- und Aufbauwortschatz Deutsch (in der aktuellsten Version *Basic German*) (Langenscheidt 1991), ergibt große Unterschiede, die in Sonderheit darauf zurückzuführen sind, dass *Basic German* nicht auf einer empirisch erfassten Häufigkeitsliste beruht, sondern auf einer Kombination der Listen von Kaeding (1897), Meier (1967), Ortmann (1975) u.a. (Langenscheidt 1991, S. VIII).

Fast 40 Prozent der häufigsten 4000 Wörter des Deutschen nach Jones/Tschirner (2006) sind nicht in *Basic German* enthalten. Tabelle 1 zeigt eine Auswahl der häufigsten 1000 Wörter des Deutschen, die nicht in *Basic German* enthalten sind:

Verben	aufnehmen, betreiben, darstellen, durchführen, erscheinen, gewinnen, leisten, reagieren, richten, stammen, umfassen, vergehen, verwenden, weisen, wirken
Substantive	Angabe, Ansatz, Beitrag, Bewegung, Ebene, Einsatz, Halt, Internet, Konzept, Kraft, Medien, Rahmen, Region, Universität, Verbindung
Adjektive	aktuell, erneut, kulturell, natürlich, speziell, toll, unmittelbar, vorhanden, weltweit
Andere Wortarten	daraus, mittlerweile, somit, stets, überhaupt, zuvor

Tabelle 1. Auswahl der häufigsten tausend Wörter nicht in *Basic German*

Ungefähr ein Drittel (32 Prozent) der Wörter in *Basic German* gehören nicht zu den häufigsten 4000 Wörtern des Deutschen nach Jones/Tschirner (2006).[2] Tabelle 2 stellt anhand von drei willkürlich gewählten Bereichen dar, in welche Tausendergruppe nach Jones/Tschirner die ersten 15 Wörter, die von Langenscheidt jeweils zur Gruppe 1-2000 der jeweiligen Liste gezählt werden, gehören.

	Körperpflege	*Gegenstände*	*Religion*
1000		Ding, Gegenstand, Gerät, Sache	glauben, Gott, Kirche
2000	Bad, sauber	gebraucht	christlich, Religion
3000	Fleck, putzen, reinigen	Griff, Kette, Messer	Bibel, Christ, Seele, Gewissen, Glaube
4000	schmutzig	Kiste	beten, Priester
5000	Creme		
6000	Dusche	Geschirr, Pfanne, Schachtel	Gebet, Sünde, Gottesdienst
7000	duschen, Handtuch, Kamm, kämmen, Schmutz	Schere	
10000		Klingel, Nadel	
15000	abtrocknen, Bürste		

Tabelle 2. Vergleich *Basic German* 1-2000 mit Jones/Tschirner (2006)

Insgesamt sind in diesen drei Bereichen nur ein Drittel der Wörter (25 von 75) der *Basic German* Liste 1-2000 unter den häufigsten 2000 Wörtern des Deutschen, ein weiteres Drittel (25) fällt unter die häufigsten 4000 Wörter und das letzte Drittel (25) gehört nicht einmal zu den häufigsten 4000 Wörtern des Deutschen. Man kann sicherlich argumentieren, dass alle Wörter, die in Basic German Vocabulary enthalten sind, nützliche Wörter sind, die gelernt werden sollen. Das Problematische ist aber der Umkehreffekt, nämlich wie viele der häufigsten 4000 Wörter nicht aufgenommen werden konnten, weil weniger häufige Wörter aufgenommen wurden. Dabei handelt es sich um fast 40 Prozent der häufigsten 4000 Wörter des Deutschen. Tabelle 3 zeigt eine Auswahl der häufigsten 1000 Wörter des Deutschen, die nicht in *Basic German* enthalten sind.

[2] Der Verlag gibt an, dass das Wörterbuch 4000 Wörter enthält, in Wirklichkeit sind es aber nur 3593. Die untersuchten Verhältnisse beziehen sich also nur auf diese knapp 3600 Wörter.

Verben	aufnehmen, betreiben, darstellen, durchführen, erscheinen, gewinnen, leisten, reagieren, richten, stammen, umfassen, vergehen, verwenden, weisen, wirken
Substantive	Angabe, Ansatz, Beitrag, Bewegung, Ebene, Einsatz, Halt, Internet, Konzept, Kraft, Medien, Rahmen, Region, Universität, Verbindung
Adjektive	aktuell, erneut, kulturell, natürlich, speziell, toll, unmittelbar, vorhanden, weltweit
Andere Wortarten	daraus, mittlerweile, somit, stets, überhaupt, zuvor

Tabelle 3. Auswahl der häufigsten tausend Wörter, die nicht in *Basic German* enthalten sind.

Ein Vergleich mit *Profile Deutsch* (Glaboniat u.a. 2002) kommt zu ähnlichen Ergebnissen. Die Schnittmenge zwischen den 2089 Wörtern der Niveaustufen A1-B1 (produktiv) und dem Häufigkeitswörterbuch von Routledge ist 60%, d.h. 40% dieser Wörter gehören nicht zu den häufigsten 4000 Wörtern des Deutschen. Was allerdings noch schwerer wiegen könnte, ist, dass 55% der häufigsten 1000 Wörter nach Jones/Tschirner (2006) nicht in den *Profile Deutsch* Listen für A1-B1 (produktiv) auftauchen. Dazu gehören Wörter wie: Idee, Künstler, Vorstellung, Bewegung, Form, Werk, Medien, Rahmen; auftreten, beteiligen, betreiben, klingen, stimmen, beschäftigen, umfassen; rund, breit, eng, unmittelbar, bewusst, gering, deutlich, notwendig, und plötzlich. Sieht man sich die häufigsten 2000 Wörter des Deutschen an, so sind 61% dieser Wörter nicht in *Profile Deutsch* A1-B1 (produktiv) enthalten.

Wenn man nun davon ausgeht – und die Wirklichkeit des Lehrwerkschreibens sieht in der Tat so aus – dass die Wortschatzvorgaben von *Profile Deutsch* von aktuellen Lehrwerken des DaF übernommen werden, dann wird noch einmal sehr deutlich, wie wichtig aktuelle Häufigkeitsuntersuchungen für das Lehren und Lernen von Deutsch als Fremdsprache geworden sind. Im nächsten Kapitel soll dargestellt werden, welche Textdeckung die häufigsten 4000 Wörter des Deutschen haben, um der Frage nachgehen zu können, welche Mindestwortschätze für unterschiedliche Textsorten anzusetzen sind.

6 Textdeckung der häufigsten 4000 Wörter des Deutschen für einige Textsorten

Je häufiger die Wörter sind, die man kennt, desto mehr tragen sie zur Textdeckung und damit zum Verständnis des Textes bei. Die bereits erwähnte Studie von Carroll, Davies und Richman (1971) ergab für das amerikanische Englisch, dass die häufigsten 1000 Wortformen ca. 74,1% der Tokens eines Textes abdecken und die häufigsten 2000 Wortformen ca. 81,3%. Die jeweils nächsten tausend Wortformen tragen immer weniger zur Textdeckung bei. Die Wortformen 2001-3000 z.B. fügen weitere 3,9% hinzu, die Wortformen 3001-4000 weitere 2,4 Prozent und die Wortformen 4001-5000 weitere 1,8 Prozent. Je nach Textsorte tragen die häufigsten Wörter allerdings jeweils unterschiedlich zur Textdeckung bei. Die häufigsten 2000 Lexeme des Englischen decken im Durchschnitt 90,3% der Tokens mündlicher Konversationen ab, 90,7% der Tokens der Kinder- und Jugendliteratur, 87,4% allgemeiner literarischer Texte, 80,3% von Zeitungstexten und 76,1% von Fachtexten (Nation 2001). Zählt man die 6,3% hinzu, die durchschnittlich von den Lexemen 3001-4000 abgedeckt werden, so kommt man mit den häufigsten 4000 Lexemen auf 96,3% von Konversationen, auf 97% der Kinder- und Jugendliteratur, auf 93,7% allgemeiner literarischer Texte, auf 86,6% von Zeitungstexten und auf 82,2% von Fachtexten. In Konversationen und beim Lesen von Kinder- und Jugendliteratur sind damit im Englischen mit einem rezeptiven Wortschatz der häu-

figsten 4000 Wörter die Textverständnisschwellen von 95-97% erreicht und beim Lesen allgemein belletristischer Texte fast erreicht. Wie sind die Verhältnisse im Deutschen? Um diese Frage zu beantworten, wurden ca. 10% der Texte des Leipzig/BYU Korpus (Tschirner/Jones 2005) getrennt nach Textsorten analysiert. Für jeden Einzeltext wurde ermittelt, welcher Prozentsatz seiner Tokens über die häufigsten 1000, 2000, 3000 und 4000 Lexeme des Deutschen nach Jones/Tschirner (2006) abgedeckt wird.

Textsorte	Texte	Tokens	Prozent
Gesprochene Sprache Interviews (BYU Korpus)	32	53.508	8%
Zeitungstexte (Kommentar, Politik) Frankf. Rundschau, Süddeutsche, Welt	84	50.372	5%
Belletristik: Anspruchsvolle Literatur, Abenteuer, Bestseller, Gesellschaft	22	219.499	20%
Fachtexte Fachzeitschriften, Uni-Einführungen	10	105.453	10%

Tabelle 4. Subkorpora aus Leipzig/BYU Korpus zur Erfassung der Textdeckung der häufigsten 4000 Wörter des Deutschen

Tabelle 4 zeigt ausgewählte Textsorten, die Anzahl der Texte, die Anzahl der Tokens in diesen Texten und die Größe des Subkorpus im Vergleich zum Gesamtkorpus der jeweiligen Textsorte im Leipzig/BYU Korpus in Prozent. Für die gesprochene Sprache wurden ca. 8% der Texte des BYU Korpus der gesprochenen Sprache (Jones 1997) gewählt, für die Zeitungssprache ca. 5% des Zeitungskorpus, die Bereiche Politik und Kommentar der Frankfurter Rundschau, der Süddeutschen Zeitung und der Welt. Für die literarische Sprache wurden ca. 20% des Literaturkorpus bestehend aus einfacher und anspruchsvoller Literatur untersucht und für die Fachsprache ca. 10% des Fachtextkorpus und zwar Fachzeitschriften und Uni-Einführungen, also die anspruchsvolleren Texte des fachsprachlichen Korpus.

	1000	2000	3000	4000	+ Namen
Gesprochene Sprache	**85,2**	**89,2**	**90,9**	**91,9**	**93,1**
Bestseller	77,8	83,0	85,3	87,1	
Abenteuerromane	73,2	79,0	81,9	83,6	
Gesellschaftsromane	73,7	79,0	81,9	83,8	
Anspruchsvolle Literatur	73,8	79,4	82,0	83,9	
Belletristik (Durchschnitt)	**74,5**	**80,0**	**82,7**	**84,5**	**88,6**
Zeitungstexte	**67,4**	**73,9**	**77,3**	**79,4**	**86,9**
Uni-Einführungen	68,9	75,9	79,6	81,9	
Fachzeitschriften	66,3	73,5	77,4	79,6	
Fachtexte (Durchschnitt)	**67,6**	**74,7**	**78,5**	**80,7**	**82,8**

Tabelle 5. Textdeckung ausgewählter Textsorten des Leipzig/BYU Korpus durch die häufigsten 4000 Wörter des Deutschen

Tabelle 5 zeigt die Ergebnisse der Textdeckungsstudie zu ausgewählten Textsorten des Leipzig/BYU Korpus. Neben den Hauptsubkorpora Gesprochene Sprache, Belletristik, Zei-

tungstexte und Fachtexte listet die Tabelle die Ergebnisse für einige Untergruppen der Belletristik und der Fachtexte auf. Ähnlich wie im Englischen decken die häufigsten 2000 Wörter des Deutschen ca. 90% der Tokens in Texten der gesprochenen Sprache ab. Ähnlich ist die Lage bei den Fachtexten, wo durch die häufigsten 2000 Wörter im Deutschen ca. 75% und im Englischen ca. 76% der Tokens abgedeckt werden. Für die anderen beiden Textsorten ergeben sich jedoch eklatante Unterschiede zwischen dem Deutschen und dem Englischen. So decken die häufigsten 2000 Wörter nur ca. 80% der Tokens eines literarischen Textes ab, im Englischen sind es 87%. In Zeitungstexten decken sie ca. 74% der Tokens ab, im Englischen dagegen ca. 80%.

Erinnern wir uns: Studien zum Englischen haben gezeigt, dass zügiges, verständnisvolles Lesen erst möglich ist, wenn mindestens 95% der Tokens eines Textes bekannt sind. Fügt man den häufigsten 4000 Wörter des Deutschen die im Durchschnitt in den untersuchten Texten vorhandenen Eigennamen hinzu, so kommt man in der gesprochenen Sprache schon relativ nahe an diese 95% heran. Bei belletristischen Texten und bei Zeitungstexten genügen die häufigsten 4000 Wörter allerdings nicht und bei Fachtexten fehlt noch Einiges. Diese Zahlen scheinen darauf hinzudeuten, dass im Deutschen ein größerer Wortschatz nötig ist als im Englischen. Hazenberg / Hulstijn (1996) stellten für das Holländische einen Bedarf von 10.000 Lexemen fest, um das Lesepensum an niederländischen Universitäten auf Holländisch zu bewältigen. Möglicherweise ist dies auch für Deutsch nötig.

Haben fortgeschrittene Lerner den zum Lesen nötigen Wortschatz? In einer Studie an der Universität Leipzig wurde ermittelt, wie groß der Wortschatz von Anglistikstudierenden zu Beginn ihres Studiums nach 8 Jahren Englischunterricht in der Schule ist. Das Resultat: 78% hatten einen Lesewortschatz von 2000 Wörtern, 28% von 3000 Wörtern und 21% von 5000 Wörtern (Tschirner 2004). Diese Ergebnisse deuten darauf hin, dass es dringend nötig ist, sich über Wortschatzbedarfe und Wortschatzerwerb Gedanken zu machen und dabei sicherlich auch Häufigkeitseffekte mitzubedenken.

7 Ausblick

Dieser Beitrag hat versucht zu zeigen, dass es viele gute Argumente für eine systematische Wortschatzarbeit, die auf aktuellen Häufigkeitsverteilungen in der deutschen Sprache aufbaut, gibt, insbesondere für die rezeptiven Fertigkeiten des Lesens und Hörverstehens. Aber auch für das Sprechen und Schreiben sind korpusbasierte Analysen wichtig, in Sonderheit Analysen, die sich mit Häufigkeitsverteilungen auseinandersetzen. Wenn man beim produktiven Lernen von einem Einheitenlernen (*chunk learning*) ausgeht, also einem Speichern von Wortketten, die bei einer genügend hohen Zahl gleicher oder ähnlicher Wortketten Anlass zur Grammatikkonstruktion geben, d.h. zum Aufbau intuitiver prozeduraler mentaler Grammatikregeln (z.B. Ellis 2002), dann könnte man solche im natürlichen Sprachgebrauch häufig vorkommenden Einheiten benutzen, um durch ihren Einsatz (und ggf. ihre Analyse) im Unterricht diesen Grammatikerwerb zu fördern bzw. zu beschleunigen.

In der Korpuslinguistik wird davon ausgegangen, dass zwischen Inhalts- und Funktionswörtern kein qualitativer Unterschied besteht, da beide durch typische Muster, in denen sie auftreten, chrakterisiert werden und in diesen Mustern grammatische und lexikalische Ele-

mente untrennbar vereint sind (Sinclair 1991). Lexeme weisen Selektionspräferenzen und -beschränkungen auf. Wie wir im Folgenden am Beispiel der Wechselpräpositionen sehen werden, sind die grammatischen Möglichkeiten von Funktionswörtern ebenfalls höchst ungleich verteilt. Dass dies lange Zeit nicht erkannt wurde, lässt sich vor allem darauf zurückführen, dass Muttersprachlerintuitionen nur bedingt mit der tatsächlichen Sprachverwendung übereinstimmen, und dass in Grammatikdarstellungen immer wieder die gleichen auffälligen klassischen Beispiele und Belege verwendet werden, die in einem kollektiven Beispielgedächtnis von Forschern und Lexikographen enthalten sind.

Eine Zusammenstellung der häufigsten einheitenbildenden Kollokationen des Herder/BYU-Korpus ergibt eine hohe Anzahl von Präpositionalphrasen. Die häufigste Kollokation *zum Beispiel* würde mit ihrer Häufigkeit von 2252 auf Platz 160 der häufigsten Ausdrücke des Deutschen kommen. *In der Regel* käme mit 408 Okkurrenzen auf Platz 881, *auf jeden Fall* auf Platz 1317, *in der Stadt* auf Platz 1803 und *auf der Straße* auf Platz 2458.[3] Insgesamt kommen in den Präpositionalphrasen, die unter die häufigsten 4000 Ausdrücke kommen würden, zehn unterschiedliche Präpositionen vor. Die Präposition *in* hat dabei mit 58% der Mehrwortausdrücke den Löwenanteil, die Präposition *auf* kommt auf 18%, die Präpositionen *an* und *mit* auf jeweils 5%. Die Präpositionen *für, gegen, nach* und *seit* sind Kern jeweils einer häufigen Kollokation, die Präposition *zu* – nur in der Form *zum* – ist Kern von zwei häufigen Kollokationen und die Präposition *vor* von dreien.

Interessant ist die Verwendung des Kasus bei den drei Wechselpräpositionen *an, auf* und *in*. Die Präposition *an* regiert in allen vier Fällen den Dativ: *am Ende* (Häufigkeit: 410), *an der Uni/versität* (213), *an dieser Stelle* (86) und *am nächsten Tag* (81). Die Präposition *auf* regiert in knapp der Hälfte der Fälle (6 von 13) den Dativ, immer mit einer lokalen Bedeutung – *auf dem Boden, Tisch, Weg; auf der Straße, Bühne, anderen Seite* – während sie, wenn sie den Akkusativ regiert, meist eine nicht-lokale, übertragene Bedeutung hat – *auf den ersten Blick, jeden Fall, keinen Fall; auf die Frage, diese Weise*. Nur in zwei Fällen wird eine grammatische Wahl sichtbar, wobei die Häufigkeiten bei *Weg* in Richtung Dativ tendieren (159 von 237) und bei *Tisch* in Richtung Akkusativ (102 von 171).

Die Präposition *in* wiederum tendiert in häufigen Mehrworteinheiten eindeutig in Richtung Dativ, nämlich in 93% aller Fälle, wobei die drei Akkusativeinheiten eine deutliche Richtungsperspektive beinhalten – *in die Hand, Stadt, Schule* – während die Dativeinheiten neben der lokalen Perspektive häufig eine temporale oder übertragene Bedeutung haben. Bei den zwei Substantiven, die sowohl mit dem Dativ wie mit dem Akkusativ auftreten – *Hand, Stadt* – überwiegt der Dativ bei *Hand* im Verhältnis von ca. 2: 1 und bei *Stadt* ca. 3: 1.

Diese Ergebnisse deuten darauf hin, dass die Grammatikdarstellung in DaF-Lehrwerken, insbesondere was die Progression angeht, wichtige Impulse von einer korpus- und häufig-

[3] Aus der Lernerperspektive man möglicherweise zwischen freien und idiosynkratischen Wortverbindungen gar nicht so richtig unterscheiden, da die freien Wortverbindungen ja nur aus einer innersprachlichen Perspektive „frei" erscheinen, aus einer kontrastiven Perspektive durchaus idiosynkratisch sein mögen (in der Stadt = in the city; in die Stadt = *to* the city). Wie im Folgenden zu sehen ist, sind aus einer Häufigkeitsperspektive auch die sogenannten freien Wortverbindungen nicht wirklich „frei". Damit stellt sich auch die Frage neu, was zur Grammatik zu rechnen ist und was zum Lexikon.

keitsbasierten Betrachtung auch grammatischer Phänomene gewinnen kann, da es sicherlich einleuchtend ist zu argumentieren, dass häufige Phänomene früher eingeführt werden sollten als weniger häufige und dass nicht komplette Paradigmen vorgestellt werden sollten, sondern in erster Linie Elemente, die relativ häufig vorkommen. Neben Grund- und Aufbauwortschätzen und Textdeckungsstudien scheinen deshalb Untersuchungen, die sich mit lexikalischen Grundlagen grammatischer Selektionsbeschränkungen auseinandersetzen, vielsprechende Einblicke in den Fremdsprachenerwerb geben zu können.

Bibliographie

A. Dictionaries

Carroll, J. B., Davies, P., Richman, B. (1971), *The American Heritage word frequency book.* New York, Houghton Mifflin.

Jones, R. / Tschirner, E. (2006), *Frequency dictionary of German. Core vocabulary for learners.* London, Routledge.

Kaeding, F. W. (1897/98), *Häufigkeitswörterbuch der deutschen Sprache.* Band 1, 2. Steglitz bei Berlin.

Langenscheidt (1991), *Basic German Vocabulary.* Berlin, Langenscheidt.

Meier, H. (1967), *Deutsche Sprachstatistik.* Hildesheim, Georg Olms.

Ortmann, W. D. (1975), *Hochfrequente deutsche Wortformen.* München, Goethe Institut.

Pfeffer, A. (1964), *Basic (Spoken) German Word List.* Englewood Cliffs, NJ, Prentice-Hall.

Rosengren, I. (1972/77), *Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung.* Band 1,2. Lund, Gleerup.

Ruoff, A. (1981), *Häufigkeitswörterbuch gesprochener Sprache.* Tübingen, Niemeyer.

B. Other Literature

Carver, R. P. (1994), 'Percentages of unknown words in text as a function of the relative difficulty of the text: Implications for instruction.' *Journal of Reading Behavior* 26, pp. 413-437.

Coxhead, A. (2000), 'A new Academic Word List.' *TESOL Quarterly* 34, pp. 213-238.

Ellis, N. (1996), 'Sequencing in SLA: Phonological memory, chunking, and points of order.' *Studies in Second Language Acquisition* 18, pp. 91-126.

Ellis, N. (2002), 'Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition.' *Studies in Second Language Acquisition* 24, pp. 143-188.

Fenk-Oczlon, G. (2001), 'Familiarity, information flow, and linguistic form' in Bybee J., Hopper P. (eds.), *Frequency and the emergence of linguistic structure.* Amsterdam, Benjamins, pp. 431-449.

Glaboniat, M., Müller, M., Rusch, P. (2002), *Profile Deutsch.* Berlin, Langenscheidt.

Hazenberg, S., Hulstijn, J. (1996), 'Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation.' *Applied Linguistics* 17, pp. 145-63.

Hu, Hsueh-Chao M., Nation, P. (2000), 'Unknown vocabulary density and reading comprehension.' *Reading in a Foreign Language* 13, pp. 403-30.

Jones, R. (1997), 'Creating and Using a Corpus of Spoken German' in Wichmann A. et al. (eds.), *Teaching and Language Corpora.* London, Longman, pp. 146-156.

Laufer, B. (1997), 'The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess' in Coady J., Huckin T. (eds.), *Second language vocabulary acquisition.* Cambridge, Cambridge University Press, pp. 20-34.

Nation, P. (2001), *Learning vocabulary in another language.* Cambridge, Cambridge University Press.

Nattinger, J., DeCarrico, J. (1992), *Lexical phrases and language teaching.* Oxford, Oxford University Press.

Qian, D. (2002), 'Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective.' *Language Learning*, 52, pp. 513-536.

Pawley, A., Syder, F. (1983), 'Two puzzles for linguistic theory: Native-like selection and native-like fluency' in Richards, J., Schmidt, R. (eds.), *Language and communication*. London, Longman, pp. 191-226.

Pfeffer, A. (1975), *Grunddeutsch. Erarbeitung und Wertung dreier deutscher Korpora. Ein Bericht aus dem "Institute for Basic German", Pittsburgh*. Tübingen, Narr.

Schmid, H. (1994), 'Probabilistic part-of-speech tagging using decision trees.' *Proceedings of the international conference on new methods in language processing*. Manchester, UK, pp. 44-49.

Scott, M. (1999), *Wordsmith Tools version 3*. Oxford, Oxford University Press.

Sinclair, J. (1991), *Corpus, concordance, collocation*. Oxford, Oxford University Press.

Swanborn, M., de Glopper, K. (1999), 'Incidental word learning while reading: A metaanalysis.' *Review of Educational Research* 69, pp. 261-285.

Scherer, G. (1965), *Final report of the director on word frequency in the modern German short story*. Boulder, Colorado. Unpublished manuscript.

Swenson, R. (1967), *A Frequency Count of Contemporary German Vocabulary based on Three Current Leading Newspapers*. Dissertation Abstracts 28: 2222A-2223A.

Tschirner, E. (2004), 'Der Wortschatzstand von Studierenden zu Beginn ihres Anglistikstudiums.' *Fremdsprachen Lehren und Lernen* 33, pp. 114-127.

Tschirner, E., Jones, R. (2005) *The Herder-BYU electronic corpus of contemporary German*. Leipzig, Herder-Institut.

Weinert, R. (1995), 'The role of formulaic language in second language acquisition: A review.' *Applied Linguistics* 16, pp. 180-205.