Willi Lange/Saburo Okamura/Joachim Scharloth

Grundwortschatz Deutsch als Fremdsprache: Ein datengeleiteter Ansatz[1]

The paper gives an overview over the different criteria and corpus measures applied in the research project 'Basic German Vocabulary for Foreign Language Learners: A data-driven Approach'. Taking their start from the intersection and differences of the vocabularies in different learners' dictionaries, the authors discuss different criteria for the selection of lexemes for basic vocabularies, give a critical account of frequency-based approaches, and introduce and discuss their own criteria, namely frequency, stability and productivity. After that, the authors present the results of their data-driven approach to basic German. Special focusses lie on the questions, whether data-driven criteria for the size of such a basic vocabulary can be found and to what extent different basic communicative constellations influence the findings.

I Zur Heterogenität von Grundwortschätzen

Hans-Heinrich Plickat sah im Grundwortschatz ein „linguistisch ungelöstes und vielleicht im absoluten Sinne nicht lösbares Problem".[2] Nicht nur die Variation hinsichtlich des Zwecks und der Adressaten des Grundwortschatzes, sondern auch die Vielfalt potentieller empirischer Grundlagen machten es unmöglich, den lexikalischen Kernbestand einer Sprache einheitlich zu bestimmen. An diesem Befund hat sich auch 35 Jahre später nichts geändert.

Grundwortschätze sind eine Teilmenge des Gesamtwortschatzes einer (Standard-)Sprache, die sich einer Selektion durch die Autorinnen und Autoren verdankt. Auch wenn sie vorgeben, „nur die deutschen Wörter, die zum modernen Deutsch gehören, das jedermann verwendet",[3] zu enthalten, so zeigt eine Analyse unterschiedlicher Grundwortschätze doch, dass diese Alltagssprache sehr unterschiedlich konstruiert wird. Die Autoren dieses Beitrags haben in einer kleinen Studie die Lexeme in sieben Grund- bzw. Lernerwortschätzen mit einander verglichen.

---

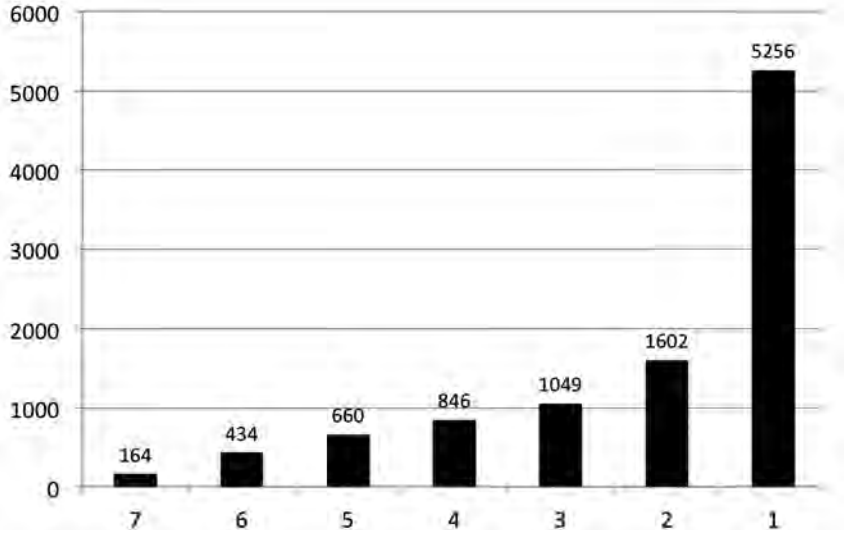
1 Das Forschungsprojekt ‚Basic German Vocabulary for Foreign Language Learners: A data-driven Approach' (コーパス駆動型研究に基づく学習用ドイツ語語彙) wurde finanziert durch einen Grant-in-Aid for Scientific Research (Kaken-B) der Japan Society for the Promotion of Science (JSPS) 2011-2015.

2 Plickat 1980, S. 10 f.

3 Lübke 2008, S. 4.

- Baldegger, Markus/Müller, Martin/Schneider, Günther: *Kontaktschwelle Deutsch als Fremdsprache.* Langenscheidt: Berlin u. a. 1993.
- Feuerle, Lois M./Schmidt, Conrad J./Weiss, Edda: *Schaum's Outline of German Vocabulary.* Mcgraw Hill o.O. 2009.
- Hiratsuka, Shigeo/Hatori, Hisahiro: *4000 Wörter Deutsch zum praktischen Gebrauch.* Hakusui-sha: Tokyo 1969.
- James, Carol/James, Charles: *Basic German Vocabulary.* Langenscheidt: Berlin u. a. 1991.
- Lübke, Diethard: *Lernwortschatz Deutsch. Deutsch-Englisch.* Hueber: Ismaning 2008.
- Reimann, Monika/Dinsel, Sabine: *Großer Lernwortschatz Deutsch als Fremdsprache. Deutsch-Englisch.* Hueber: Ismaning 2006. (hier nur die Asterix-Wörter)
- Tschirner, Erwin: *Deutsch als Fremdsprache.* Grund- und Aufbauwortschatz nach Themen. Cornelsen: Berlin 2008.

Abb. 1: Anzahl der Wörter (y-Achse), die in n Grundwortschätzen (x-Achse) vorkommen.



Insgesamt enthielten die Lehrwerke rund 10.000 unterschiedliche Lexeme. Wie Abbildung 1 zeigt, kommen mehr als die Hälfte (5.256 Lexeme) von ihnen nur in einem einzigen Grundwortschatz vor. Gerade einmal 164 Lexeme werden in allen sieben Grundwortschätzen eingeführt. Dies ist ein deutliches Indiz dafür, dass die Wortschatzselektion entweder nach sehr unterschiedlichen Kriterien erfolgt

ist, oder dass dieselben Kriterien sehr unterschiedlich angewendet wurden bzw. keine einheitlichen Kriterien zur Anwendung kamen.

Dieser Befund bestätigt sich auch dann, wenn man die Schnittmengen von je zwei Grundwortschätzen vergleicht.

Tabelle 1: Schnittmengen im Vokabular von sieben unterschiedlichen Grundwortschätzen für Deutsch als Fremdsprache

	Baldegger Kontakt-schwelle	*Hirat-suka / Hatori 4000: Wörter*	*Langen-scheidt: Basic German Vocabu-lary*	*Lübke: Lernwort-schatz Deutsch*	*Reimann / Dinsel: Großer Lernwort-schatz*	*Tschirner: Grund- und Aufbauwort-schatz*	*Schaum's Outline of German Vocabu-lary*
Baldegger Kontakt-schwelle	100 %	43.3 %	71.8 %	67,9 %	69,2 %	57.3 %	27.7 %
Hiratsuka / Hatori: 4000 Wörter	27.6 %	100 %	48.3 %	42.1 %	47.1 %	37.1 %	16.4 %
Langen-scheidt: Basic German Vocabulary	33.2 %	35.1 %	100 %	60.2 %	58.8 %	66 %	17.6 %
Lübke: Lern-wortschatz Deutsch	37.8 %	36.8 %	72.4 %	100 %	67.4 %	63.2 %	20.4 %
Reimann / Dinsel: Großer Lern-wortschatz	22.1 %	23.7 %	40.7 %	38.8 %	100 %	37.3 %	13.8 %
Tschirner: Grund- und Aufbauwort-schatz	24.6 %	25 %	61.2 %	48.7 %	49.9 %	100 %	13.2 %
Schaum's Outline of German Vocabulary	32.5 %	30,3 %	44.7 %	43.1 %	50.7 %	36.1 %	100 %

Tabelle 1 zeigt, dass die Übereinstimmungen in einem Bereich zwischen 13 % und 73 % liegen. Dies hat seine Ursache zwar teilweise darin, dass die Umfänge der Wortschätze sehr unterschiedlich sind, dennoch verweisen diese Ergebnisse insgesamt darauf, dass es offenbar an belastbaren und reproduzierbar einsetzbaren Kriterien

sowie einer als Referenzkorpus fungierenden Datengrundlage für die Zusammenstellung von Grundwortschätzen fehlt.

II Ansätze zur Bestimmung des zentralen Wortschatzes[4]

Grundwortschätze wollen Lernenden jene Lexeme einer Standardsprache näher bringen, die dazu befähigen, sich möglichst schnell mit den Angehörigen einer Sprachgemeinschaft, die Trägerin der zu erlernenden Fremdsprache ist, zu verständigen. Doch Lernerinnen und Lerner können nicht alle Wörter einer Sprache erlernen, Lehrwerke und Grundwortschätze müssen eine Auswahl treffen. Das Kriterium, das bei der Auswahl fast immer implizit zur Begründung dient, ist die Wahrscheinlichkeit, mit der ein Lerner bzw. eine Lernerin mit einem Wort in Kontakt kommt. Doch wie bestimmt man die Wahrscheinlichkeit, mit der man mit einem Wort einer Fremdsprache konfrontiert wird?

Der *kommunikativ-pragmatische Ansatz* geht von in Sprachgemeinschaften typischen kommunikativen Situationen und Sprechintentionen aus, denen dann die sprachlichen Mittel – und somit auch der Wortschatz – zugeordnet werden können. Peter Kühn hat die bei der Selektion leitende Frage wie folgt formuliert: „Welches lexikalische Material [...] benötigt ein Sprecher/Schreiber [...], um in der Situation [...] über das Thema [...] in der Rolle [...] die kommunikative Intention [...] mithilfe des Kommunikationsmodus [...] erfolgreich durchzuführen?“[5] Für das Deutsche bilden *Zertifikat Deutsch als Fremdsprache* (1972, Neubearbeitung 1992)[6], *Kontaktschwelle Deutsch* (1980)[7] und die deutsche Ausarbeitung des europäischen Referenzrahmens in *Profile* (2005)[8] Meilensteine des kommunikativ-pragmatischen Ansatzes. Insbesondere *Profile* hat sich zu einem Quasi-Standard für Lehrbücher entwickelt. So plausibel dieser Ansatz auch erscheint, so wenig empirisch fundiert ist er: er beruht nicht auf einer Erhebung oder gar Quantifizierung des Sprachgebrauchs in typischen Alltagssituationen. Der Situationsbegriff ist theoretisch ebenso wenig hinreichend bestimmt wie das Alltagskonzept. Zudem sind die sprachlichen Selektionsverfahren intransparent.

---

4 Für uns ist im Folgenden der Begriff *Zentraler Wortschatz* der Oberbegriff für zwei verschiedene Typen von begrenzender Wortschatzbeschreibung. Während *Kernwortschatz* eine zweckfreie Beschreibung bezeichnet, wird *Grundwortschatz* für alle Formen der Beschreibung verwendet, die eine sprachdidaktische Zielsetzung haben. Dabei ist zunächst unerheblich, ob die Zielsetzung muttersprachlich oder fremdsprachlich ist.

5 Kühn 1989, S. 230-239, hier S. 20.

6 Deutscher Volkshochschulverband/Goethe-Institut 1985.

7 Baldegger 1980.

8 Glaboniat, Manuela u. a.: Profile deutsch. 2005.

Der *frequenzorientierte Ansatz* bestimmt die Wahrscheinlichkeit, mit der man mit einem Wort einer Fremdsprache konfrontiert wird, indem er große Korpora auf die Häufigkeit des Auftretens von Lexemen hin untersucht. Für das Deutsche sind neben frühen Ausarbeitungen von Pfeffer[9] und Rosengren in jüngerer Zeit mit Jones/Tschirner[10] und Tschirner[11] neue frequenzbasierte Versuche der Bestimmung eines Grundwortschatzes getreten. In ihnen ist die Häufigkeit eines Wortes das Hauptkriterium der Selektion. Zwar geht dieser Ansatz empirisch vor, allerdings ist die Wahl des Korpus bzw. dessen Zusammenstellung und Umfang von entscheidender Bedeutung für das Ergebnis. Die vorhandenen Korpora freilich sind meist sehr selektiv im Hinblick auf die von ihnen abgedeckten Kommunikationsbereiche und bilden die gesprochene Sprache nur äußerst fragmentarisch ab. Zudem kann man am frequenzorientierten Ansatz kritisieren, dass *Häufigkeit* und *Wichtigkeit* von Lexemen verkürzend gleichgesetzt wird und dass wegen der starken Formbezogenheit Bedeutungsgesichtspunkte und die kommunikative Funktion von Wörtern generell vernachlässigt werden. Gleichwohl haben frequenzorientierte Ansätze den Vorteil, dass sie überhaupt eine empirische Grundlage haben, ihre Ergebnisse folglich reproduzierbar sein müssen und somit die Möglichkeit eröffnen, intersubjektiv nachvollziehbare Maßstäbe in die Wortschatzselektion einzubringen.

Der *lexikographische Ansatz* schließlich nimmt eine Metaperspektive ein: Auf der Basis von vorhandenen Wörterbüchern oder Wortschatzsammlungenversucht man, einen Kern bzw. ein Zentrum zu herauszudestillieren. Repräsentanten dieses Ansatzes sind Schnörch (2002) und Haderlein (2008).[12]

Nicht immer werden zentrale Wortschätze aber konsequent einem der Ansätze folgend konstruiert. Häufig werden die Ansätze gemischt, wie im Fall des *Basic German Vocabulary* von James und James:[13]

> „Langenscheidt's Basic Vocabulary selects the most important words for a student to learn and use. The Basic Vocabulary is based on evaluation of numerous lists of basic German vocabulary published in Germany, Austria, Switzerland and other countries. All the important sources of information on word frequency in written and spoken German were considered. [...] The choice of words was not based only on frequency. Factors such as how familiar and useful a word is in everyday conversation were also considered. Langenscheidt's experience in producing dictionaries and teaching materials also helped."

9 Pfeffer 1970.
10 Jones/Tschirner 2006.
11 Tschirner 2008.
12 Haderlein 2008. Schnörch 2002.
13 James/James, Basic German Vocabulary, S. VII.

Das Vermischen unterschiedlicher Ansätze ist nicht grundsätzlich zu kritisieren. Problematisch ist in diesem Fall (und analog in anderen Fällen) jedoch, dass intransparent bleibt, wann und mit welcher Begründung von der Frequenzorientierung Abstand genommen und wie die Gebräuchlichkeit und Nützlichkeit eines Wortes in Alltagskonversation bestimmt wurden.

III Das Projekt „Datengeleiteter Grund- und Aufbauwortschatz Deutsch"

Im Forschungsprojekt „Datengeleiteter Grund- und Aufbauwortschatz Deutsch" verfolgen wir einen frequenzorientierten Ansatz. Dieser ermöglicht es, die Entscheidung, warum ein Wort zum Kernwortschatz des Deutschen gehört und deshalb in einen Grundwortschatz für Lernende aufgenommen werden sollte, methodisch transparent und damit nachvollziehbar zu machen. Wir vertreten dabei ähnlich wie Tschirner (2008) einen radikal frequenzorientierten Ansatz, d. h. dass wir Lemmafrequenzen nicht nur in Zweifelsfällen als Entscheidungshilfe heranziehen, sondern sie prinzipiell zur Grundlage der Berechnung des Kernwortschatzes machen. Unser Vorgehen möchten wir daher als datengeleitet[14] (im Gegensatz zu datenbasiert) bezeichnen.

An den bisherigen frequenzbasierten Ansätzen schienen uns folgende Aspekte problematisch:

1. Sie operieren mit einem zu engen Verständnis von Frequenzorientierung. Diese wird gleichgesetzt mit einer Berechnung der Rangfolge der relativen Frequenzen von Lemmata in einem Korpus.
2. Sie beruhen meist auf zu kleinen Korpora, deren Repräsentativität für „die deutsche Sprache" fragwürdig ist.

(Ad 1.) Anders als etwa für Tschirner bedeutet Frequenzorientierung für uns jedoch nicht ausschließlich, eine Rangfolge der relativen Frequenzen von Lemmata in einem Korpus als Kriterium für die Aufnahme in den Kernwortschatz zu wählen. Wir verstehen *frequenzorientiert* allgemeiner im Sinn von *die Distribution von Lexemen/lexikalischen Morphemen betreffend* und differenzieren den Frequenzaspekt in die Dimensionen (1) Häufigkeit, (2) Stabilität und (3) Produktivität.

Zum Kernwortschatz zählen wir demnach jene Lexeme, die (1) häufig vorkommen, die (2a) über einen längeren Zeitraum gleichmäßig häufig auftreten (also keine Modewörter sind), (2b) nicht bzw. kaum themenaffin sind (d. h. in Texten

14 Tognini-Bonelli 2001.

unterschiedlicher thematischer Prägung gleichmäßig distribuiert sind), die (3a) als lexikalische Morpheme in vielen Ableitungen und Zusammensetzungen (Types) auftreten, die (3b) als Lexeme selbst häufig sind (Tokens), und (3c) die als lexikalische Morpheme häufiger als Zweitglied in Komposita verwendet werden. Die Frequenzdimensionen wurden mittels der in Tabelle 2 dargestellten Werte operationalisiert.

Tabelle 2: Übersicht über die Operationalisierung der Frequenzdimensionen

Dimension	*Spezifizierung*	*Berechnungsbasis*	*Wert*	*Gewichtung*
Häufigkeit	Frequenz	gesamtes Korpus	Häufigkeitsklasse[15]	3
Stabilität	temporale Stabilität	jahresspezifische Subkorpora	Gries' DP[16]	2
	thematische Stabilität	Rubriken / Teilforen als Subkorpora	Gries' DP	2
Produktivität	Anzahl unterschiedlicher Ableitungen und Komposita	Types	absolute Frequenz	1
	Frequenz des Auftretens der Ableitungen und Komposita	Token	absolute Frequenz	1
	Anzahl von Ableitungen und Komposita in niedrigen Häufigkeitsklassen	Distribution der Ableitungen und Komposita über die Häufigkeitsklassen	Entropie	1
	Frequenz in Funktion als Determinatum	auf der Basis der Types	relative Frequenz	1

Die so berechneten Werte wurden normalisiert (teilweise logarithmiert), gewichtet und mit Hilfe eines Vektordistanzmodells nach ihrem Abstand zum Idealvektor in eine Rangfolge gebracht.

(Ad 2.) Aus unserer Sicht ist der Versuch, ein Textkorpus zusammenzustellen, das alltagsweltlich relevante kommunikative Gattungen, Register und Stile abbil-

15 Vgl. Perkuhn/Keibel/Kupietz 2012, S. 80-82.
16 Gries 2008, S. 403-437.

det, sowie hinsichtlich regionaler und altersmäßiger Verteilung der Autorinnen und Autoren ausgewogen ist, zum Scheitern verurteilt. Zwar waren beispielsweise Jones und Tschirners sehr sorgfältig bei der Zusammenstellung ihres Leipzig/BYU Corpus of Contemporary German, über das sie schreiben[17]

> „It is a balanced, structured, and integrated corpus, meaning that it was carefully planned to achieve representation of genre, register, style, geography, and age group. lt consists of one million words each of spoken language, literature, newspapers, and academic texts, and 200,000 words of instructional language."

Doch sind die Annahmen darüber, welche kommunikativen Gattungen, Register und Stile für das Gegenwartsdeutsch relevant sind, spekulativ, denn hierzu gibt es keine empirisch gesättigten linguistischen Untersuchungen. Wenn aber die Grundgesamtheit unbekannt ist, dann ist auch *Repräsentativität* im Sinne einer strukturellen Analogie zwischen Sample und Grundgesamtheit nicht erreichbar.[18]

Bei der Zusammenstellung des Textkorpus, auf dessen Basis der Kernwortschatz berechnet wurde, gingen wir daher von zwei kommunikativen Grundkonstellationen aus: Einerseits mehrfachadressierende und konzeptionell schriftliche Texte, andererseits aber auch Texte, die persönlich adressierend und konzeptionell mündlich sind. Um die Stabilität messen zu können, sollte das Korpus zudem mehrere Jahre abdecken. Zur validen Messung von Stabilität und Produktivität ist zudem ein umfangreiches Korpus notwendig.

Für mehrfachadressierende und konzeptionell schriftliche Texte griffen wir auf Zeitungs- bzw. Zeitschriftentexte (Print und Online) zurück, für persönlich adressierende und konzeptionell mündliche Texte auf Diskussionsforen aus dem Internet, weil nur in ihnen zeitlich hinreichend rückläufige Massendaten zur Verfügung stehen. Insgesamt umfasst unser Korpus rund 845 Millionen Wörter aus Online-Diskussionsforen aus den Jahren 1998 bis 2012 (rund 475 Millionen laufende Wortformen, siehe Tabelle 3) sowie aus Zeitungstexten der Jahre 1990 bis 2012 (370 Millionen laufende Wortformen, siehe Tabelle 4).

Tabelle 3: Übersicht über das Foren-Teilkorpus (persönlich adressiert und konzeptionell mündlich)

	Beiträge	*Wörter*
seniorentreff.de	1.005.159	68.514.967
bfriends.brigitte.de	1.719.564	141.686.509

17 Jones/Tschirner 2006, S. 2.

18 Zudem ist das Korpus mit 4.2 Millionen laufenden Wörtern sehr klein.

	Beiträge	*Wörter*
politikforen.net	3.260.363	263.866.105
Gesamt Foren:	5.985.086	474.067.581

Tabelle 4: Übersicht über das Zeitungs-Teilkorpus (mehrfachadressiert, konzeptionell schriftlich)

	Beiträge	*Wörter*
SPON	374.253	151.852.627
Spiegel Print 1990-2011	139.578	87.156.665
ZEIT 1995-2011	114.109	86.915.216
FOCUS 1993-2012	106.400	43.349.229
Gesamt Zeitungen:	734.340	369.273.737

Die Korpora wurden mit dem TreeTagger[19] lemmatisiert und mit Part-of-Speech-Informationen annotiert. Für die morphologische Analyse kam Morphisto, der auf dem SFST-Toolkit beruht, mit der morphologischen Komponente SMOR[20] zum Einsatz. Alle anderen Berechnungen wurden mit eigenen Softwareentwicklungen realisiert. Der Kernwortschatz wurde sowohl für das gesamte Korpus als auch für die beiden Kommunikationsbereiche getrennt berechnet.

Ein Beispiel: Die Lexeme *Frucht* und *Futter* sind beide in der Häufigkeitsklasse 11, einer Häufigkeitsklasse, die üblicherweise den Grenzbereich für frequenzbasierte Grundwortschätze bildet. Sucht man nun nach weitergehenden frequenzbasierten Kriterien, welches der beiden Wörter bessere Chancen haben sollte, in einen Grundwortschatz aufgenommen zu werden, dann können wir zunächst die Stabilität über Themen (geringere Themenaffinität) und Stabilität in der Zeit in den Blick nehmen. Das Lexem *Frucht* weist jeweils den niedrigeren Wert für Gries' DP auf (Zeit: 0.074, Themen: 0.201) als das Lexem *Futter* (Zeit: 0.084, Themen: 0.233), wobei beide eine relativ starke Themenaffinität erkennen lassen. Auf der Basis der Stabilitätswerte sollte also das Lexem *Frucht* die besseren Chancen haben, Teil eines Grundwortschatzes zu werden.

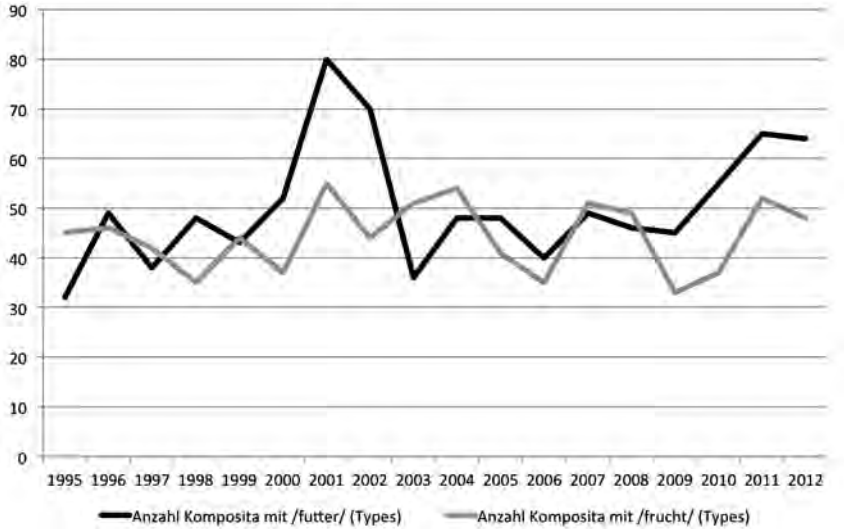
Hinsichtlich der Produktivität scheint zunächst keinem der beiden Lexeme der Vorzug zu gebühren: Im Zeitungskorpus finden sich 985 unterschiedliche

19 Vgl. Schmid 1995. und Schmid 1994.

20 Schmid/Fitschen/Heid 2004, S. 1263-1266.

Komposita (Types) und Ableitungen mit dem lexikalischen Morphem /futter/, 964 für /frucht/, ihr Auftreten ist darüber hinaus auch relativ stabil verteilt, lediglich für das Jahr 2001 ist ein signifikanter Anstieg der Komposita mit /futter/ zu beobachten (Abbildung 2).

Abb. 2: Anzahl unterschiedlicher Komposita und Ableitungen (Types) mit den lexikalischen Morphemen /futter/ und /frucht/



Betrachtet man hingegen die Anzahl der Komposita- und Ableitungstoken (Abbildung 3), dann zeigt sich, dass Zusammensetzungen mit /frucht/ deutlich häufiger verwendet werden als Zusammensetzungen mit /futter/. Ein Blick auf die Distribution der Komposita und Ableitungen über die Häufigkeitsklassen (Abbildung 4) zeigt, dass – auch wenn die allermeisten Lemmata in die höchste Häufigkeitsklasse fallen – einige Komposita und Ableitungen mit Frucht selbst relativ häufig vorkommen. Dazu zählen insbesondere *fruchtbar*, *unfruchtbar* und *Fruchtbarkeit*, *Fruchtsaft* und *fruchtig*. Die häufigsten Komposita mit /futter/, *Tierfutter*, *Hundefutter*, *Futtermittel*, *Kraftfutter* und *Kanonenfutter*, kommen hingegen vergleichsweise selten vor.

Die Unterschiede in den Dimensionen Stabilität und Produktivität führen im Ergebnis dazu, dass in unserer Berechnung des zentralen Wortschatzes *Frucht* auf Rang 3673, *Futter* hingegen auf Rang 4729 landet, obwohl beide zur gleichen Häufigkeitsklasse gehören.

Abb. 3: Relative Frequenz von Komposita- und Ableitungstoken mit den lexikalischen Morphemen /futter/ und /frucht/ je 100.000 Wörter

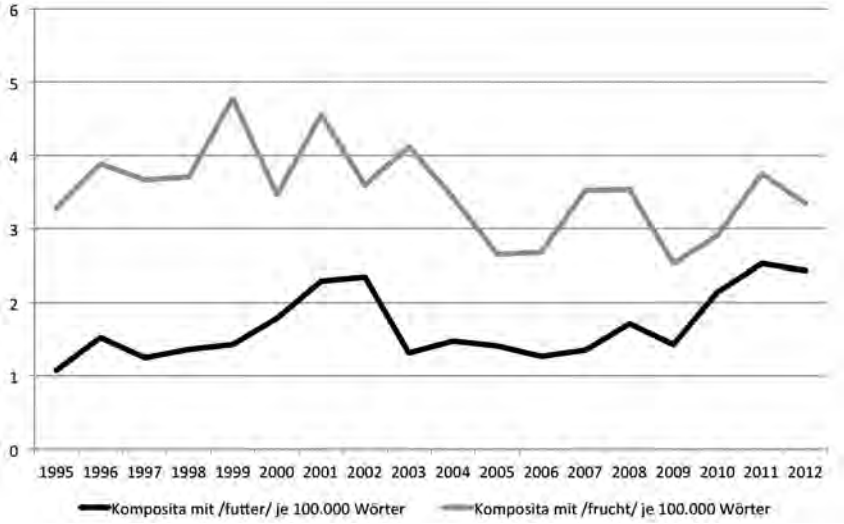
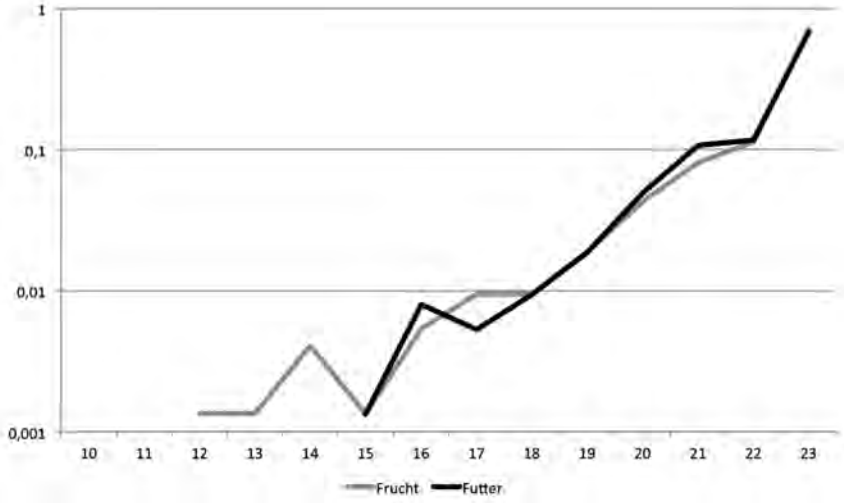


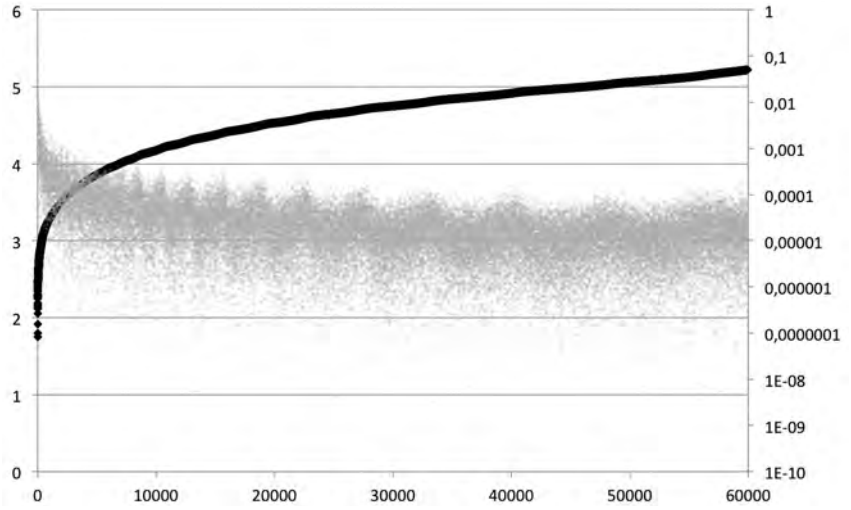
Abb. 4: Anteil der Häufigkeitsklassen in der Menge der Ableitungen und Komposita mit /frucht/ und /futter/ (normalisierte Werte und logarithmierte y-Achse)



IV Ergebnisse der datengeleiteten Analyse

Das Ergebnis unserer Berechnungen ist eine nach Distanz zum Idealvektor (höchste Frequenz, höchste Stabilitätswerte, höchste Produktivität) geordnete Liste von Lexemen. Abbildung 5 zeigt die Distribution der Vektordistanzen (schwarze Kurve, linke y-Achse) und die Differenzen der Vektordistanzen zwischen aufeinander folgenden Lexemen (graue Punkte, rechte Sekundärachse, logarithmiert). Sie illustriert, dass die Distanzen zunächst groß sind, immer kleiner werden und sich schließlich bei einem Wert zu stabilisieren scheinen. Die Abbildung zeigt auch, dass es – zumindest aus dieser Perspektive – kein datengeleitetes Kriterium für die Abgrenzung eines zentralen Wortschatzes von einem Fachwortschatz gibt: keine Wendepunkte, keine anderen Veränderungen der Eigenschaft der Kurve erlauben eine Abgrenzung. In der Konsequenz bedeutet dies, dass der Umfang von frequenzbasierten Grundwortschätzen für Deutsch als Fremdsprache sich ausschließlich nach didaktischen Kriterien richten kann.

Abb. 5: Aufsteigend geordnete Vektordistanzen der top 60.000 Lexeme (schwarz, Primärachse) und Differenzen der Vektordistanzen zwischen aufeinander folgenden Lexemen (graue Punkte, logarithmierte Sekundärachse)

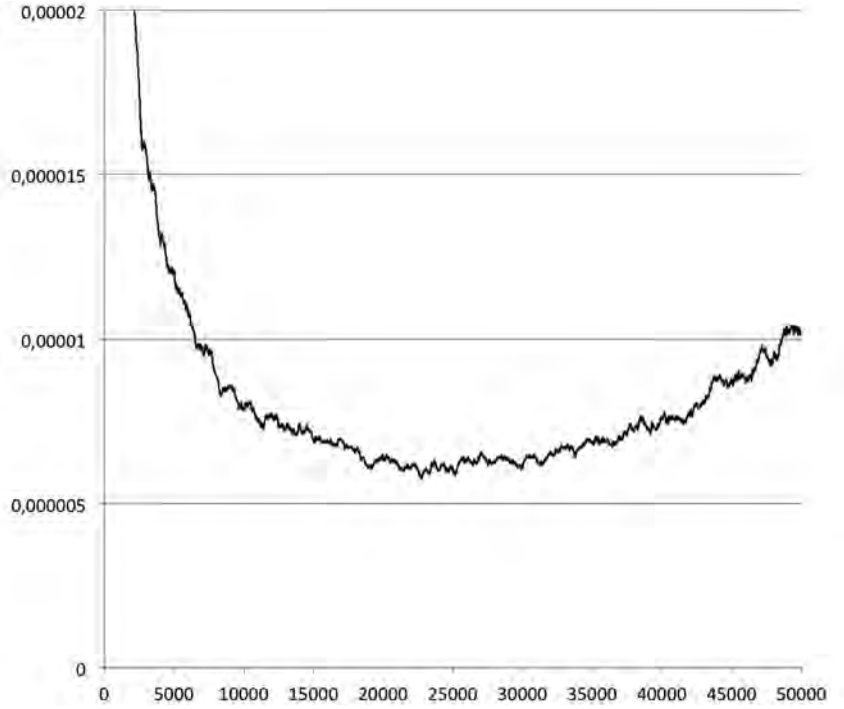


Vor

Zwei Indizien für eine Begrenzung des Kernwortschatzes lassen sich dennoch ausmachen. Zum einen ist hier die Verteilung der Stabilitätswerte zu nennen. Abbildung 6 zeigt die Differenzen zwischen den aufeinander folgenden Werten

im Stabilitätsranking (gleitende Durchschnitte). Diese Differenzen nehmen bis zum Rang 23.000 ab und nehmen dann wieder zu. Dies bedeutet, dass jenseits der Marke von 23.000 der Wortschatz von immer größerer Instabilität in der Zeit geprägt ist und zugleich themenspezifischer wird.

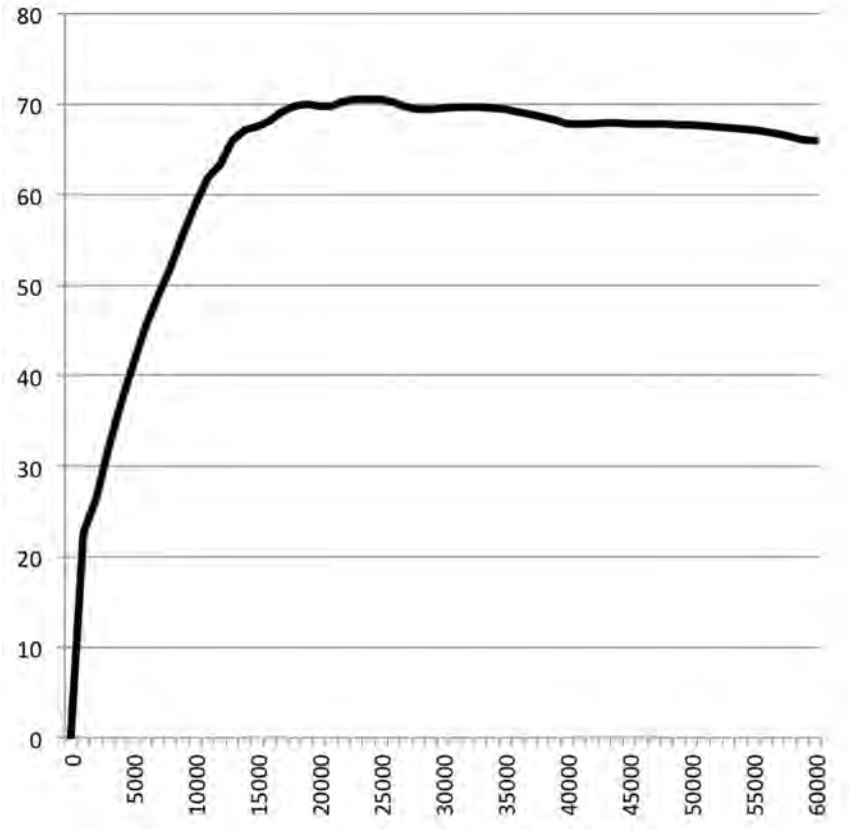
Abb. 6: Differenzen aufeinander folgender Werte im Stabilitätsranking (gleitende Durchschnitte)



Ein weiteres Indiz liefert die getrennte Berechnung der Grundwortschätze für das Foren- und das Zeitungskorpus. Vergleicht man, wie groß der Anteil gemeinsamer Lexeme bei Wörtern mit dem Rang von 1 bis n in beiden Listen ist, erhält man für unterschiedliche n sehr unterschiedliche Schnittmengengrößen. Abbildung 7 zeigt die Größe der Schnittmenge, wenn man n in 1000er-Schritten erhöht. Die Schnittmenge wächst zunächst auf etwas über 70 % an, nimmt dann aber ab Rang 23.000 leicht ab. Dies bedeutet, dass sich der Wortschatz stärker diversifiziert. Wie beim Stabilitätsranking scheint auch hier eine Grenze bei 23.000 Wörtern zu liegen. Auch wenn dieser Befund interessant ist, ist er bei der Zusammenstellung von Grundwortschätzen für Deutschlernende doch wenig hilfreich. Denn sollte

der zentrale Wortschatz des Deutschen auch 23.000 Lexeme umfassen, könnte ein solch umfangreiches Lexikon in einem Unterricht in herkömmlicher Form und Umfang doch unmöglich erarbeitet werden.

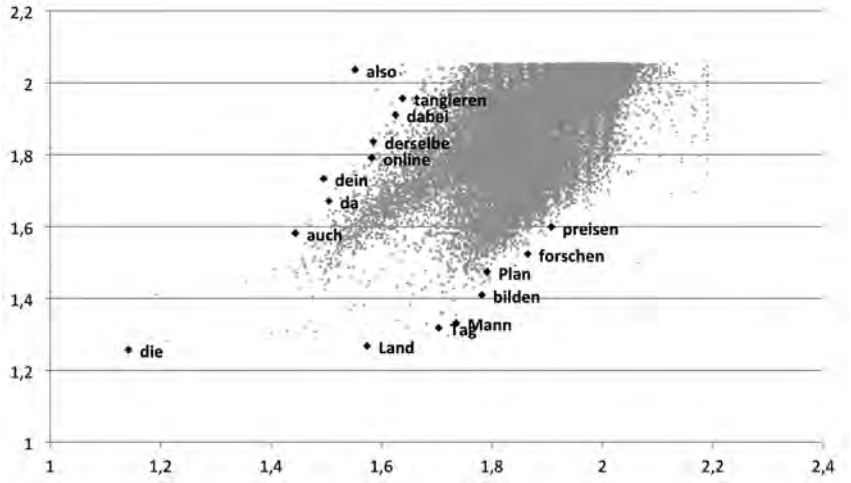
Abb. 7: Größe der Schnittmengen in Prozent bei 0 bis n Wörtern in Foren- und Zeitungsgrundwortschatz in Prozent



Die getrennte Berechnung des Grundwortschatzes für mehrfachadressierende, konzeptionell schriftliche Texte einerseits und für persönlich adressierende, konzeptionell mündliche Texte andererseits erlaubt auch einen Vergleich hinsichtlich der lexikalischen Unterschiede beider Kommunikationsbereiche. Abbildung 8 visualisiert in einer Matrix die Vektordistanzen aus beiden Berechnungen für jene Wörter, die in beiden Korpora mit einer maximalen Häufigkeitsklasse von 17 vorkommen.

Untersucht man, welche Lexeme die größten Differenzen in den Vektordistanzen aufweisen, so zeigt sich folgendes Muster: Während im Forenkorpus vor allem Funktionswörter, Adverbien und interaktionskoordinierende Lexeme eine geringe Distanz zum Idealvektor aufweisen, sind es im Zeitungskorpus vorwiegend Inhaltswörter. Im Forenkorpus sind demnach Kommunikation eröffnende oder schließende Wörter (*Hallo, hallo, Gruß, wünschen, freundlich, grüßen, Hi, Moin, Spaß*), Wörter, die auf Sprechakte verweisen (*Dank, danken, Glückwunsch, bitte, bitten, leid, Entschuldigung, entschuldigen, Bitte*), Partikel (vor allem Modal- und Gradpartikel: *ach, mal, ja, wohl; sehr, möglichst*) und Adverbien, vor allem Kommentaradverbien (*glücklicherweise, angeblich, sicherlich, hoffentlich, möglicherweise, leider*) und spezielle Temporaladverbien (*neulich, übermorgen, damals*) häufiger, stabiler und produktiver.

Abb. 8: Vektordistanzen (ungewichtet) für Zeitungskorpus (y-Achse) und Forenkorpus (x-Achse), beschriftet sind einige Datenpunkte mit großer Vektordistanzdifferenz



Folgende grundwortschatzrelevanten Lexeme haben hingegen im Zeitungskorpus eine deutlich geringere Distanz zum Idealvektor als im Forenkorpus: *fahren, Mann, Tag, führen, Zeit, bilden, bauen, desolat, nehmen, geben, leiten, Haus, fliegen, forschen, landen, Werk, Stand, üben, Programm, kommen, Zentrum, bilanzieren, Plan, Direktor, Rat, wählen, bergen, Schule, enden, ziehen, managen, Firma, beraten, preisen, Smog, Dienst, Land, Ministerium, Beginn, Vorsitz, abwiegeln, Stelle, Fall, raten, planen, mögen* etc.

V Fazit

Mit dem im Rahmen des Forschungsprojektes ‚Basic German Vocabulary for Foreign Language Learners: A data-driven Approach' berechneten zentralen Wortschatz des Deutschen liegt erstmals eine empirische Basis für die Erstellung von Grundwortschätzen und Lehrwerken für Deutsch als Fremdsprache vor, die nicht nur die Häufigkeit eines Lexems als reproduzierbares Kriterium für die Wortschatzselektion berücksichtigt, sondern auch die Produktivität lexikalischer Morpheme und die thematische und diachrone Stabilität. Die exemplarische Analyse in diesem Aufsatz konnte zeigen, dass das Einbeziehen der Selektionskriterien der Produktivität und der Stabilität einen großen Einfluss auf das Ranking der Wörter hat. Zudem ist es mit der von uns erarbeiteten Datengrundlage möglich, eine an kommunikativen Grundkonstellationen orientierte Binnendifferenzierung des zentralen Wortschatzes vorzunehmen.

Die Schwächen des frequenzorientierten Ansatzes liegen auf der Hand: Homonyme mit gleicher Schreibweise, unterschiedliche Verwendungsweisen eines Wortes werden durch den oberflächlichen Zugriff über die Form des Lexems nicht sichtbar. Dies betrifft auch die Berechnung der Produktivität: sie setzt stillschweigend voraus, dass Komposita und Ableitungen semantisch transparent sind, was zwar in vielen Fällen der Fall ist, aber längst nicht in allen. Auch künftige Grundwortschatz- und Lehrbuchautoren werden also nicht auf eine differenzierte Wortschatzbeschreibung verzichten können.

Literatur

Baldegger, Markus/Müller, Martin/Schneider, Günther: *Kontaktschwelle Deutsch als Fremdsprache.* Langenscheidt: Berlin u. a. 1993.

Deutscher Volkshochschulverband/Goethe-Institut: *Das Zertifikat Deutsch als Fremdsprache.* 3. Aufl. o.V.: Bonn/Frankfurt 1985.

Feuerle, Lois M./Schmidt, Conrad J./Weiss, Edda: *Schaum's Outline of German Vocabulary.* Mcgraw Hill: o.O. 2009.

Glaboniat, Manuela u. a.: *Profile deutsch.* Langenscheidt: Berlin 2005.

Gries, Stefan Th.: „Dispersions and adjusted frequencies in corpora", in: *International Journal of Corpus Linguistics* 13:4 (2008), S. 403-437.

Haderlein, Veronika: *Das Konzept zentraler Wortschätze. Bestandsaufnahme, theoretisch-methodische Weiterführung und praktische Untersuchung.* Diss. München 2008.

Hiratsuka, Shigeo/Hatori, Hisahiro: *4000 Wörter Deutsch zum praktischen Gebrauch.* Hakusui-sha: Tokyo 1969.

James, Carol / James, Charles: *Basic German Vocabulary.* Langenscheidt: Berlin u. a. 1991.

Jones, Randall L./ Tschirner, Erwin: *A Frequency Dictionary of German. Core vocabulary for learners.* Routledge: London, New York 2006.

Kühn, Peter: „Pragmatische Aspekte der Grundwortschatzbestimmung." In: *Neuphilologische Mitteilungen* 81, 1989, S. 230-239

Lübke, Diethard: *Lernwortschatz Deutsch. Deutsch-Englisch.* Hueber: Ismaning 2008.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc: *Korpuslinguistik.* Fink: Paderborn 2012. (Reihe LIBAC – Linguistik für Bachelor 3433)

Pfeffer, Allan J.: *Grunddeutsch. Basic (Spoken) German Dictionary.* Prentice-Hall: Englewood Cliffs, N.J. 1970.

Plickat, Hans-Heinrich: *Deutscher Grundwortschatz. Wortlisten und Wortgruppen für Rechtschreibunterricht und Förderkurse. Unter Mitarbeit von Reiner Herden.* Beltz Verlag: Weinheim, Basel 1980.

Reimann, Monika/Dinsel, Sabine: *Großer Lernwortschatz Deutsch als Fremdsprache.* Deutsch-Englisch. Hueber: Ismaning 2006.

Rosengren, Inger: *Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung.* 2 Bde. LiberLäromedel/Gleerup: Lund 1970-1977.

Schnörch, Ulrich: *Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und lexikografischen Aufarbeitung.* Narr: Tübingen 2002.

Schmid, Helmut: „Improvements in Part-of-Speech Tagging with an Application to German", In: *Proceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland 1995.

Schmid, Helmut: „Probabilistic Part-of-Speech Tagging Using Decision Trees", In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK 1994.

Schmid, Helmut/Fitschen, Arne/Heid, Ulrich: „SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection", In: *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal 2004, S. 1263-1266.

Tognini-Bonelli, Elena: *Corpus linguistics at work.* Benjamins: Amsterdam 2001. (= Studies in corpus linguistics 6).

Tschirner, Erwin: *Deutsch als Fremdsprache. Grund- und Aufbauwortschatz nach Themen.* Cornelsen: Berlin 2008.