Joachim Scharloth / Saburo Okamura / Willi Lange

Gibt es einen Kernwortschatz?[1] Datengeleitete Perspektiven auf die Erstellung von Grundwortschätzen für Deutsch als Fremdsprache

Ein Grundwortschatz ist eine Teilmenge des Gesamtwortschatzes einer (Standard-) Sprache, die von Lernerinnen und Lernern einer Fremdsprache zuerst gelernt werden soll. Es handelt sich also um eine zu didaktischen Zwecken getroffene Auswahl, die auf unterschiedliche Art begründet werden kann. Die mit der Idee eines Grundwortschatzes verknüpfte Vorstellung lässt sich dabei stets wie folgt explizieren: Es gibt einen Kernbestand an lexikalischen Einheiten, mit dessen Hilfe es möglich ist, in einer Sprachgemeinschaft, die Trägerin der zu erlernenden Fremdsprache ist, sprachlich zu handeln und ggf. mangelndes lexikalisches Wissen selbst zu erschließen bzw. sich im Kommunikationsprozess anzueignen. Die Idee eines Grundwortschatzes für Fremdsprachenlerner fußt damit auf der Annahme eines Kernwortschatzes in der zu erlernenden Sprache, der in allen Kommunikationssituationen zur Anwendung kommt und eine grundlegende Verständigung sichert.

So plausibel diese Annahme auf den ersten Blick erscheint, so wenig geklärt sind ihre theoretischen, methodologischen und empirischen Grundlagen. Nicht abschließend beantwortet oder teilweise weitgehend unbeantwortet sind beispielsweise die Fragen:

1. Anhand welcher Kriterien kann der Kernwortschatz ermittelt werden?
2. Gibt es tatsächlich einen abgrenzbaren lexikalischen Kern des Wortschatzes einer Sprache / einer Sprachgemeinschaft oder doch eher ein Kontinuum, das sich in einzelne Soziolekte verästelt?
3. Ist dieser Kern unabhängig von den Kommunikationszwecken und sozialen Merkmalen der Sprecherinnen und Sprecher?

Diese Fragen sollen im Rahmen des vorliegenden Aufsatzes diskutiert werden und anhand empirischer Studien Hinweise auf mögliche Antworten gefunden werden.

[1] Der Aufsatz beruht auf Ergebnissen aus dem 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), das durch einen Grant-in-Aid for Scientific Research (Kaken-B) der Japanese Society for the Promotion of Science (JSPS) 2011-2015 finanziert wurde.

1. Kriterien zur Bestimmung des Kernwortschatzes

Die Kriterien zur Bestimmung des Kernwortschatzes einer Sprache beruhen größtenteils auf einer oft nur implizit formulierten Korrespondenzhypothese, wie im Fall des *Lernwortschatzes Deutsch* von Diethard Lübke. Er motiviert die Auswahl aus dem Gesamtwortschatz damit, dass die betreffende Teilmenge „nur die deutschen Wörter" umfasse, „die zum modernen Deutsch gehören, das jedermann verwendet".[2] Der Grundwortschatz ist in seiner Darstellung damit eine Abstraktion des tatsächlichen Sprachgebrauchs; und zwar nicht des Sprachgebrauchs in einzelnen Domänen, Medien oder einzelner Gruppen, sondern jener Anteile des Sprachgebrauchs jedes Einzelnen, die sich bei allen anderen auch finden, mithin die Schnittmenge. Doch wie kann man bestimmen, welche Wörter tatsächlich im Sprachgebrauch von „jedermann" vorkommen? Hierfür lassen sich drei Ansätze unterscheiden.

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. Für das Deutsche bilden die Bücher *Das Zertifikat Deutsch als Fremdsprache* (1972, ^{3}1985, Neubearbeitung 1992),[3] *Kontaktschwelle Deutsch* (^{3}1993 [1980])[4] und die deutsche Ausarbeitung des *Gemeinsamen europäischen Referenzrahmens für Sprachen* in *Profile deutsch* (2005)[5] Meilensteine des kommunikativ-pragmatischen Ansatzes. Insbesondere *Profile* hat sich zu einem Quasi-Standard für Lehrbücher entwickelt. So plausibel dieser Ansatz auch klingt, 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.

Mit dem *frequenzorientierten Ansatz* wird das Ziel verfolgt, die Wahrscheinlichkeit zu bestimmen, mit der man mit einem Wort einer Fremdsprache konfrontiert wird. Hierfür wird die Distribution von Lexemen in großen Korpora analysiert. Für das Deutsche sind neben frühen Ausarbeitungen von Pfeffer (1970)[6] und Rosengren (1972-1977)[7] in jüngerer Zeit mit Jones / Tschirner (2006)[8] und Tschirner (2008)[9] 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

---

2 Lübke 2008, 4.

3 Deutscher Volkshochschulverband / Goethe-Institut 1985.

4 Baldegger / Müller / Schneider ^{3}1993 [1980].

5 Glaboniat / Müller / Rusch / Schmitz / Wertenschlag 2005.

6 Pfeffer 1970.

7 Rosengren 1970-1977.

8 Jones / Tschirner 2006.

9 Tschirner 2008.

und Umfang von entscheidender Bedeutung für das Ergebnis. Die vorhandenen Korpora sind freilich meist (insbesondere bei den früheren Grundwortschätzen) 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 Lehrwerkerstellung einzubringen.

Der *lexikographische Ansatz* sucht durch Kombination und Kollationierung von vorhandenen Wörterbüchern und / oder Wortschatzsammlungen einen zentralen Wortschatz zu identifizieren. Neuere Repräsentanten dieses Ansatzes sind die Arbeiten von Schnörch (2002) und Haderlein (2008).[10] Der lexikographische Ansatz geht davon aus, dass durch die Bildung von Schnittmengen von je zweckgebundenen Wortlisten, sich ein zweckabstraktes lexikalisches Zentrum einer Sprache zeigt.

In der lexikographischen Praxis kommen häufiger mehrere Auswahlkriterien zum Einsatz, etwa in *Langenscheidt's Basic Vocabulary*,[11] das zunächst angibt, dem lexikographischen Prinzip zu folgen: „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."[12] Im Anschluss aber verdeutlichen die Autoren, dass auch Frequenzargumente („All the important sources of information on word frequency in written and spoken German were considered.")[13] und kommunikativ-pragmatische Aspekte („Factors such as how familiar and useful a word is in everyday conversation were also considered.")[14] berücksichtigt wurden. Daneben floss noch die Erfahrung des Verlags in den Auswahlprozess ein („Langenscheidt's experience in producing dictionaries and teaching materials also helped.").[15] Ob es eine Kriterien-Hierarchie gab bzw. in welchen Fällen welches Kriterium zur Anwendung kam, darüber schweigen die Autoren.

Dieser Mangel an konsistenter Befolgung von Auswahlkriterien führt dazu, dass die Schnittmengen zwischen unterschiedlichen Grundwortschätzen nicht sehr groß sind, wie wir in einer früheren Publikation[16] zeigen konnten. Darin untersuchten wir die folgenden Grundwortschätze auf Schnittmengen im Wortschatz:

---

[10] Schnörch 2002, Haderlein 2008.

[11] James / James 1991.

[12] James / James 1991, VIII.

[13] James / James 1991, VIII.

[14] James / James 1991, VIII.

[15] James / James 1991, VIII.

[16] Bubenhofer / Lange / Okamura / Scharloth 2016.

- Baldegger, Markus / Martin Müller / Günther Schneider ([3]1993 [1980]), *Kontaktschwelle Deutsch als Fremdsprache.* Berlin (u.a.): Langenscheidt.
- Feuerle, Lois M. / Conrad J. Schmidt / Edda Weiss (2009), *Schaum's Outline of German Vocabulary.* New York (u.a.): Mcgraw Hill.
- Hiratsuka, Hisahiro (1969), *4000 Wörter Deutsch zum praktischen Gebrauch.* Tokyo: Hakusuisha.
- James, Carol / Charles James (1991), *Basic German Vocabulary*. Berlin (u.a.): Langenscheidt.
- Lübke, Diethard (2008), *Lernwortschatz Deutsch. Deutsch-Englisch.* Ismaning: Hueber.
- Reimann, Monika / Sabine Dinsel (2006), *Großer Lernwortschatz Deutsch als Fremdsprache. Deutsch-Englisch.* Ismaning: Hueber. (Hier wurden ausschließlich die als Bestandteil der Wortliste des *Zertifikat Deutsch* gekennzeichneten Lemmata erfasst.)
- Tschirner, Erwin (2008), *Deutsch als Fremdsprache. Grund- und Aufbauwortschatz nach Themen.* Berlin: Cornelsen.[17]

Abb. 1 zeigt, dass mehr als die Hälfte (5.256 Lexeme) der Lexeme nur in einem einzigen Grundwortschatz vorkommen. 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 Kriterien zur Anwendung kamen.

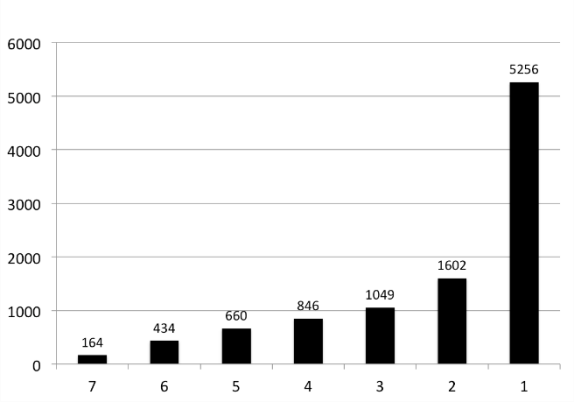

Abb. 1: Anzahl der Wörter (y-Achse), die in n Grundwortschätzen (x-Achse) vorkommen.

[17] Diese Auswahl deckt wichtige aktuelle Grundwortschätze ab (Lübke 2008; Reimann / Dinsel 2006; Tschirner 2008), Meilensteine in der Geschichte der DaF-Lexikographie (Baldegger / Müller / Schneider [3]1993 [1980]; James / James 1991) sowie Grundwortschätze, die für Lernende aus einer spezifischen Sprachgemeinschaft konzipiert wurden (Feuerle / Schmidt / Weiss 2009; Hiratsuka 1969).

Im Hinblick auf die in der Einleitung formulierte Frage nach den Kriterien zur Identifizierung eines Kernwortschatzes kann also festgehalten werden, dass die Forschung zwar unterschiedliche Ansätze erarbeitet hat, diese jedoch nicht zu konsistenten Ergebnissen führen. Ob dies seine Ursache in der inkonsequenten Kriterienanwendung oder in der mangelnden Validität des Konstrukts „Kernwortschatz" hat, soll im nächsten Abschnitt diskutiert werden.

2. Gibt es Grenzen des Kernwortschatzes?

Um die Frage, ob es tatsächlich einen abgrenzbaren lexikalischen Kern des Wortschatzes einer Sprache oder doch eher ein Kontinuum gibt, das sich in einzelne Soziolekte verästelt, sollen im Folgenden die Ergebnisse einer frequenzbasierten Studie dargestellt werden, die wir anhand sehr großer Korpora durchgeführt haben. Die Frage nach der Abgrenzbarkeit eines Kernwortschatzes wird damit zwar nur aus der Perspektive eines Ansatzes heraus untersucht; allerdings fiel die Wahl dabei auf jenen Ansatz, der objektivierbare Befunde am ehesten erwartbar macht. In unserer Studie vertreten wir ähnlich wie Tschirner[18] einen radikal frequenzorientierten Ansatz, das heißt, dass wir Lemmafrequenzen nicht nur in Zweifelsfällen als Entscheidungshilfe heranziehen, sondern sie prinzipiell zur Grundlage der Berechnung des Kernwortschatzes machen. Unser Vorgehen bezeichnen wir daher als datengeleitet[19] (im Gegensatz zu datenbasiert).

An den bisherigen frequenzbasierten Ansätzen schienen uns zwei Aspekte problematisch. *Erstens* operieren frequenzbasierte Ansätze mit einem zu engen Verständnis von Frequenzorientierung. Diese wird gleichgesetzt mit einer Berechnung der Rangfolge der relativen Frequenzen von Lemmata in einem Korpus. Im Gegensatz dazu bedeutet Frequenzorientierung für uns 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 (das heißt in Texten 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 Tab. 1 dargestellten Werte operationalisiert.

---

[18] Tschirner 2008.

[19] Tognini-Bonelli 2001.

Dimension	*Spezifizierung*	*Berechnungsbasis*	*Wert*	*Gewichtung*
Häufigkeit	Frequenz	gesamtes Korpus	Häufigkeitsklasse[20]	3
Stabilität	temporale Stabilität	jahresspezifische Subkorpora	Gries' DP[21]	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

Tab. 1: Übersicht über die Operationalisierung der Frequenzdimensionen.

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

Zweitens arbeiten frequenzorientierte Ansätze meist mit zu kleinen Korpora, deren Repräsentativität für „die deutsche Sprache" bzw. die deutsche Standardsprache fragwürdig ist. Aus unserer Sicht ist der Versuch, ein Textkorpus zusammenzustellen, das alltagsweltlich relevante kommunikative Gattungen, Register und Stile abbildet, sowie hinsichtlich regionaler und altersmäßiger Verteilung der Autorinnen und Autoren ausgewogen ist, zum Scheitern verurteilt. Zwar waren beispielsweise Jones und Tschirner sehr sorgfältig bei der Zusammenstellung ihres Leipzig / BYU Corpus of Contemporary German, über das sie schreiben:[22] „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

[20] Vgl. Perkuhn / Keibel / Kupietz 2012, 80-82.

[21] Gries 2008, 403-437.

[22] Jones / Tschirner 2006, 2.

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.[23] 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 diachrone 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 Tab. 2) sowie aus Zeitungstexten der Jahre 1990 bis 2012 (370 Millionen laufende Wortformen, siehe Tab. 3).

	Beiträge	*Wörter*
seniorentreff.de	1.005.159	68.514.967
bfriends.brigitte.de	1.719.564	141.686.509
politikforen.net	3.260.363	263.866.105
Gesamt Foren:	*5.985.086*	*474.067.581*

Tab. 2: Übersicht über das Foren-Teilkorpus
(persönlich adressiert und konzeptionell mündlich).

	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*

Tab. 3: Übersicht über das Zeitungs-Teilkorpus
(mehrfachadressiert, konzeptionell schriftlich).

Die Korpora wurden mit dem *TreeTagger*[24] 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*[25] zum Einsatz. Alle

[23] Zudem ist das Korpus mit 4,2 Millionen laufenden Wörtern sehr klein.

[24] Vgl. Schmid 1994.

[25] Schmid / Fitschen / Heid 2004, 1263-1266.

anderen Berechnungen wurden mit eigenen Softwareentwicklungen realisiert. Der Kernwortschatz wurde sowohl für das gesamte Korpus als auch für die beiden Kommunikationsbereiche getrennt berechnet.

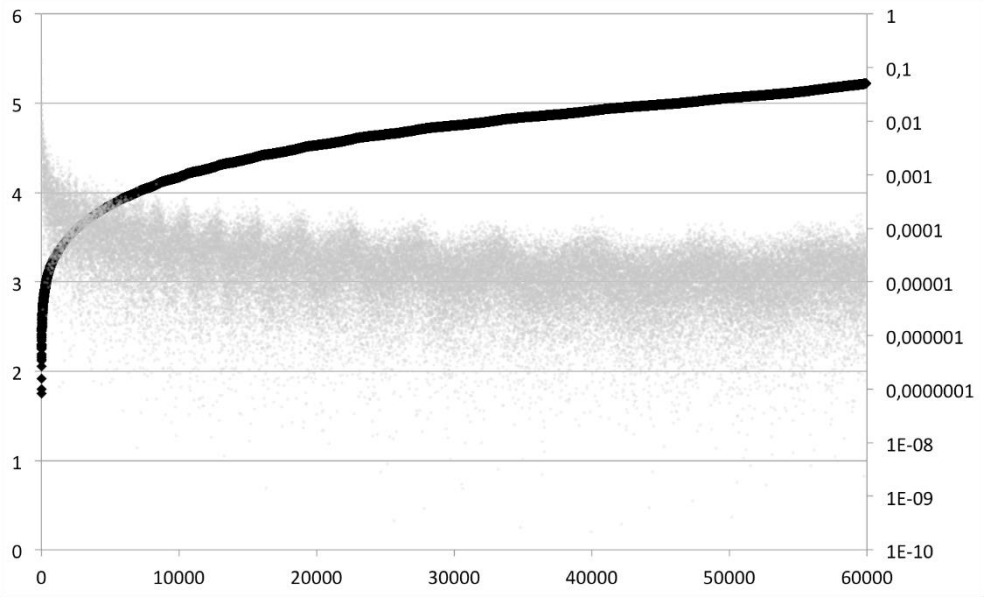

Abb. 2: Aufsteigend geordnete Vektordistanzen der top 60.000 Lexeme (schwarz, Primärachse) und Differenzen der Vektordistanzen zwischen aufeinander folgenden Lexemen (graue Punkte, logarithmierte Sekundärachse).

Das Ergebnis unserer Berechnungen ist eine nach Distanz zum Idealvektor (höchste Frequenz, höchste Stabilitätswerte, höchste Produktivität) geordnete Liste von Lexemen. Abb. 2 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. Sie zeigt damit, dass es kein datengeleitetes Kriterium für die Abgrenzung eines zentralen Wortschatzes von einem Bildungs- oder Fachwortschatz gibt: Keine Wendepunkte, keine anderen Veränderungen der Kurve erlauben eine Grenzziehung. In der Konsequenz bedeutet dies, dass der Umfang von Grundwortschätzen für Deutsch als Fremdsprache sich guten Gewissens ausschließlich nach didaktischen Kriterien richten kann.

Zwar suggeriert das datengeleitete Verfahren eine Homogenität in den Daten, gleichwohl bleibt die Frage noch unbeantwortet, wie groß die kommunikationsbereichsspezifische Variation des so berechneten Kernwortschatzes ist.

3. Zur Zweckgebundenheit von Grundwortschätzen

Zum Konzept des Grundwortschatzes gehört wie eingangs beschrieben die Vorstellung, dass alle Domänen und Kommunikationsbereiche gleichermaßen von ihm durchdrungen sind und die betreffenden Lexeme überall die Verständigung sichern. Mit der Idee des Kernwortschatzes einher geht also die Vorstellung seiner Kommunikationszweck- und Kontextabstraktheit. Doch wie homogen sind die Ergebnisse von datengeleiteten Grundwortschatzanalysen, wenn man sie auf Korpora aus unterschiedlichen Kommunikationsbereichen mit unterschiedlichen Kommunikationszwecken anwendet? Dies soll im Folgenden anhand von drei exemplarischen Berechnungen überprüft werden.

Mithilfe der in Abschnitt 2 entwickelten Methoden wurden für die beiden kommunikationsbereichsspezifischen Teilkorpora (massenmediale, konzeptionell schriftliche vs. persönlich adressierende, konzeptionell mündliche Kommunikation, vgl. Tab. 2 und 3) sowie für ein Kinderbuch-Korpus eine nach Vektordistanzen geordnete Liste berechnet. Das Kinderbuch-Korpus bestand aus 1067 Kinder- und Jugendbüchern (Originalwerke und Übersetzungen) aus dem 20. und 21. Jahrhundert mit zusammen 39.460.099 Wortformen. Kinderbücher können als Repräsentanten für Texte mit einfacher Sprache gelten. Sowohl im Hinblick auf Wortlänge als auch im Hinblick auf die Differenziertheit des Wortschatzes.

Um die Homogenität der Grundwortschätze zu ermitteln, wurde gemessen, wie groß der Anteil gemeinsamer Lexeme bei Wörtern mit dem Rang von 1 bis n in den unterschiedlichen Vektordistanz-Rankings ist. Wie Abb. 3 zeigt, variiert die Schnittmengengröße für unterschiedliche n.

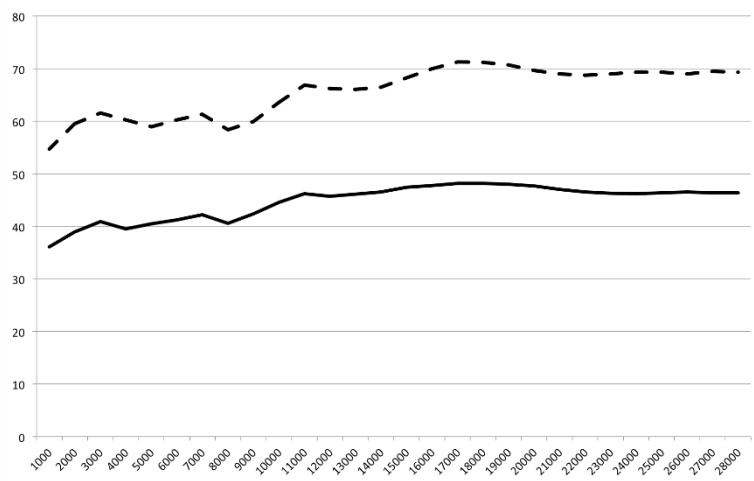

Abb. 3: Größe der Schnittmenge in den ersten n Wörtern in den Rankings von Foren- und Zeitungskorpus (gestrichelte Linie) sowie Foren-, Zeitungs- und Kinderbuchkorpus (durchgehende Linie).

Die Schnittmenge des Foren- und Zeitungskorpus wächst zunächst auf etwas über 70 % an, nimmt dann aber ab Rang 18.000 leicht ab. Die gleiche Entwicklung ist auf (erwartbar) niedrigerem Niveau für die Schnittmenge der Rankings auf der Basis aller drei Korpora – also des Foren-, Zeitungs- und Kinderbuchkorpus – sichtbar.

Die Grafik ist für die im vorliegenden Aufsatz thematisierte Fragestellung in mehrfacher Hinsicht aussagekräftig. Zum einen wird sichtbar, dass das Konstrukt eines abgrenzbaren Kernwortschatzes generell fragwürdig ist. Wenn der gemeinsame Wortschatz unterschiedlicher Kommunikationsbereiche lediglich zwischen 60 und 70 Prozent liegt und bei Hinzuziehung eines weiteren Kommunikationsbereichs (fiktional-narrativ) signifikant auf unter 50 Prozent sinkt, dann ist fragwürdig, ob diese Schnittmenge als Kernwortschatz angesehen werden kann. Vielmehr ist davon auszugehen, dass die Schnittmenge sich bei Hinzuziehung weiterer Kommunikationsbereiche weiter signifikant verkleinert und sich damit auch der vermeintliche Kernbestand des Wortschatzes weiter verflüchtigt. Darüber hinaus ist die Schnittmenge insbesondere bei den für den Grundwortschatz für Fremdsprachenlerner relevanten n besonders klein und liegt lediglich zwischen 35 und 40 Prozent, wenn man alle drei Rankings in die Analyse einbezieht. Dabei wären im unteren Bereich der Rankings, also im Bereich der frequentesten, produktivsten und stabilsten Lexeme, eigentlich die höchsten Übereinstimmungen zu erwarten, wenn sich im Sprachgebrauch ein Kernwortschatz manifestieren würde. Dies ist jedoch nicht der Fall.

Die dem lexikographischen Ansatz folgenden Analysen dieses Abschnitts deuten damit darauf hin, dass es keinen vom Kommunikationszweck unabhängigen Kernwortschatz gibt, sieht man einmal vom hochfrequenten Funktionswortschatz ab. Der Wortschatz diversifiziert sich vielmehr schon im Bereich der hochfrequenten, produktiven und stabilen Lexeme. Der Wortschatz ist weniger ein Baum, dessen Krone (spezielle Wortschätze) auf einem Stamm (Kernwortschatz) ruht, sondern eher ein Busch, der sich nahe am Boden verzweigt.

4. Fazit

Auf der Basis der vorgestellten Untersuchungen lässt sich die Leitfrage dieses Beitrags nach der Existenz eines Kernwortschatzes wie folgt beantworten. Zwar existieren mit dem kommunikativ-pragmatischen, dem frequenzorientierten und dem lexikographischen Ansatz unterschiedliche Kriterien für die Bestimmung eines Kernbereichs des Wortschatzes. In der lexikographischen Praxis werden die einzelnen Kriterien aber nicht konsequent und konsistent angewendet. Dies führt dazu, dass sich in unterschiedlichen Grundwortschätzen sehr unterschiedlicher Wortschatz findet und die Schnittmengen gering sind. Die Frage, ob dies nur eine Konsequenz der (inkonsistenten) Anwendung unterschiedlicher Methoden ist oder seine Ursache darin hat, dass es empirisch keinen abgrenzbaren, homogenen Kernwortschatz gibt, wurde im Rahmen des vorliegenden Beitrags im frequenzorientierten Paradigma untersucht. Dabei konnte festgestellt werden, dass sich datengeleitet kein Kriterium für einen Kernbereich des

Wortschatzes identifizieren lässt. Darüber hinaus zeigte ein Vergleich von Kernwortschätzen, die anhand unterschiedlicher kommunikationsbereichsspezifischer Korpora berechnet wurden, dass auch hier die lexikalischen Einheiten stark variieren und die Schnittmengen entsprechend klein waren. Dies lässt den Schluss zu, dass die Frequenz, Produktivität und Stabilität von Lexemen abhängig von den Kommunikationszwecken variiert und zwar auch und gerade bei den hochfrequenten, hochproduktiven und sehr stabilen Lexemen.

Für die Frage nach dem Kernwortschatz bedeutet dies, dass zumindest aus frequenzorientier Sicht mehr gegen seine Existenz spricht als dafür. Dies bedeutet freilich nicht, dass sich keine Lexeme identifizieren ließen, die häufiger, produktiver und stabiler verwendet würden. Nur bilden diese Lexeme keinen abgrenzbaren, zweckabstrakten Teil innerhalb des Gesamtwortschatzes. Dies hat auch Konsequenzen für die Erstellung von Grundwortschätzen für Fremdsprachenlernerinnen und Fremdsprachenlerner: Hinsichtlich der Wortschatzselektion ist für sie eine Orientierung am Kommunikationszweck sinnvoll, ihr Umfang kann anhand didaktischer Kriterien begrenzt werden.

Literaturverzeichnis

Baldegger, Markus / Martin Müller / Günther Schneider (^{3}1993 [1980]), *Kontaktschwelle Deutsch als Fremdsprache.* Berlin (u.a.): Langenscheidt.

Bubenhofer, Noah / Willi Lange / Saburo Okamura / Joachim Scharloth (2016), „Wortschatz in Lehrwerken für Deutsch als Fremdsprache: ein frequenzorientierter Ansatz." In: Jana Kiesendahl / Christine Ott (Hg.): *Linguistik und Schulbuchforschung.* Göttingen: V&R unipress.

Deutscher Volkshochschulverband / Goethe-Institut (^{3}1985), *Das Zertifikat Deutsch als Fremdsprache.* Bonn / Frankfurt a.M.: Deutscher Volkshochschulverband.

Feuerle, Lois M. / Conrad J. Schmidt / Edda Weiss (2009), *Schaum's Outline of German Vocabulary.* New York (u.a.): Mcgraw Hill.

Glaboniat, Manuela / Martin Müller / Paul Rusch / Helen Schmitz / Lukas Wertenschlag (2005), *Profile deutsch*. Berlin (u.a.): Langenscheidt.

Gries, Stefan Thomas (2008), „Dispersions and Adjusted Frequencies in Corpora". In: *International Journal of Corpus Linguistics* 13/4, 403-437.

Haderlein, Veronika (2008), *Das Konzept zentraler Wortschätze. Bestandsaufnahme, theoretisch-methodische Weiterführung und praktische Untersuchung*, Dissertation. Ludwig-Maximilians-Universität München, München.

Hiratsuka, Hisahiro (1969), *4000 Wörter Deutsch zum praktischen Gebrauch.* Tokyo: Hakusuisha.

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

Jones, Randall L. / Erwin Tschirner (2006), *A Frequency Dictionary of German. Core Vocabulary for Learners.* London / New York: Routledge.

Lange, Willi / Saburo Okamura / Joachim Scharloth (2015), „Grundwortschatz Deutsch als Fremdsprache: Ein datengeleiteter Ansatz." In: Jörg Kilian / Jan Eckhoff (Hg.): *Deutscher Wortschatz – beschreiben, lernen, lehren. Beiträge zur Wortschatzarbeit in Wissenschaft, Sprachunterricht, Gesellschaft.* Frankfurt a.M. (u.a.): Peter Lang, 203-219.

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

Perkuhn, Rainer / Holger Keibel / Marc Kupietz (2012), *Korpuslinguistik*. Paderborn: Fink.

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

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

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

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

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

Schmid, Helmut / Arne Fitschen / Ulrich Heid (2004), „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, 1263-1266.

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

Tognini-Bonelli, Elena (2001), *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

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