Preprint von: Bubenhofer, Noah / Lange, Willi / Okamura, Saburo / Scharloth, Joachim: Wortschätze in Lehrbüchern für Deutsch als Fremdsprache: Möglichkeiten und Grenzen frequenzorientierter Ansätze. In: Kiesendahl, J. / Ott, C. (Hrsg.): Linguistik und Schulbuchforschung. Gegenstände – Methoden – Perspektiven, Eckert. Göttingen: V&R unipress, 2015, S. 85–110.

Wortschätze in Lehrbüchern für Deutsch als Fremdsprache: Möglichkeiten und Grenzen frequenzorientierter Ansätze[1]

Noah Bubenhofer, Dresden / Willi Lange, Tokyo / Saburo Okamura, Tokyo / Joachim Scharloth, Dresden

Kommunikativ-pragmatischer und frequenzorientierter Ansatz

Wortschatzaufbau ist neben der Vermittlung grammatikalischer und pragmatischer Kompetenz die zentrale Aufgabe von Lehrbüchern für Deutsch als Fremdsprache. Doch welcher Wortschatz soll vermittelt werden? Die Antwort klingt zwar einfach, bringt aber viele Probleme mit sich: Vermittelt werden sollten jene Wörter, die es den Lernenden ermöglichen, sich verstehend und verständigend in der Sprachgemeinschaft, die Trägerin der zu erlernenden Fremdsprache ist, zu bewegen. Das eigentliche Problem liegt jedoch darin, die Mittel, die dazu befähigen, sich mit den Angehörigen einer Sprachgemeinschaft zu verständigen, exakt zu benennen. Sie lassen sich nicht nur aus der kommunikativen Praxis der Sprachgemeinschaft ableiten, sondern hängen auch von den Interessen und Lebenslagen der Lernenden ab. Dennoch müssen Lehrbücher eine Auswahl aus der großen Anzahl an Lexemen treffen, die zum Wortschatz von Standardsprachen gehören. Das Kriterium, das dabei häufig 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. Für das Deutsche bilden die Bücher *Zertifikat Deutsch als Fremdsprache* (1972, Neubearbeitung 1992), *Kontaktschwelle Deutsch* (1980) und die deutsche Ausarbeitung des Gemeinsamen Europäischen Referenzrahmens für Sprachen in *Profile* (2005) 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. Zu diesem Zweck werden große Korpora auf das Auftreten von Lexemen hin untersucht. Für das Deutsche sind neben frühen Ausarbeitungen von Pfeffer (1970) und Rosengren (1972-1977) in jüngerer Zeit mit Jones/Tschirner (2006) und Tschirner (2008) 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.

[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.

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.
In den folgenden Abschnitten wollen wir diskutieren, wie frequenzorientierte Ansätze für die Analyse von Lehrwerken nutzbar gemacht werden können. Wir beschränken uns hierbei auf die korpuslinguistisch recht einfach zu operationalisierenden Aspekte des Wortschatzes in Lehrwerken für Deutsch als Fremdsprache: die Distribution von Lexemen und den Wortschatzaufbau.

Analyse von Grundwortschätzen[2]

Zunächst nehmen wir den Lehrwerkstyp Grundwortschatz in den Blick, bei dem Wortschatzfragen im Zentrum stehen. Grundwortschätze sind in ihrer einfachsten Form Listen von Wörtern, die eine Auswahl aus dem zentralen Wortschatz einer Sprache zum Zweck ihres Erlernens repräsentieren. Das Auswahlkriterium ist die Wichtigkeit eines Wortes für die Verständigung im Medium der zu erlernenden Sprache. Oft sind Grundwortschätze nach Themen oder Situationen geordnete Listen, die auch lexikographische Informationen und Übersetzungen in die jeweilige Muttersprache der Lernenden enthalten. Diese Aspekte werden jedoch in der folgenden Untersuchung vernachlässigt. Hier soll es allein darum gehen, wie groß die lexikalischen Schnittmengen zwischen sieben Grundwortschätzen für Deutsch als Fremdsprache sind.
Die folgenden Grundwortschätze wurden in die Analyse einbezogen:

- Baldegger, Markus/Müller, Martin/Schneider, Günther (1993): *Kontaktschwelle Deutsch als Fremdsprache*. Berlin u. a.
- Feuerle, Lois M./Schmidt, Conrad J./Weiss, Edda (2009): *Schaum's Outline of German Vocabulary*. o. O.
- Hiratsuka, Hatori (1969): *4000 Wörter Deutsch zum praktischen Gebrauch*. Tokyo.
- James, Carol/James, Charles (o. J.): *Basic German Vocabulary*. Berlin u. a.
- Lübke, Diethard (2008): *Lernwortschatz Deutsch. Deutsch-Englisch*. Ismaning.
- Reimann, Monika/Dinsel, Sabine (2006): *Großer Lernwortschatz Deutsch als Fremdsprache. Deutsch-Englisch*. Ismaning. (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.

Diese Auswahl deckt wichtige aktuelle Grundwortschätze ab (Lübke, Reimann/Dinsel, Tschirner), Meilensteine in der Geschichte der DaF-Lexikographie (Baldegger, James/James) sowie Grundwortschätze, die für Lernende aus einer spezifischen Sprachgemeinschaft konzipiert wurden (Feuerle/Schmidt/Weiss, Hiratsuka). 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 Kriterien zur Anwendung kamen.

[2] 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.

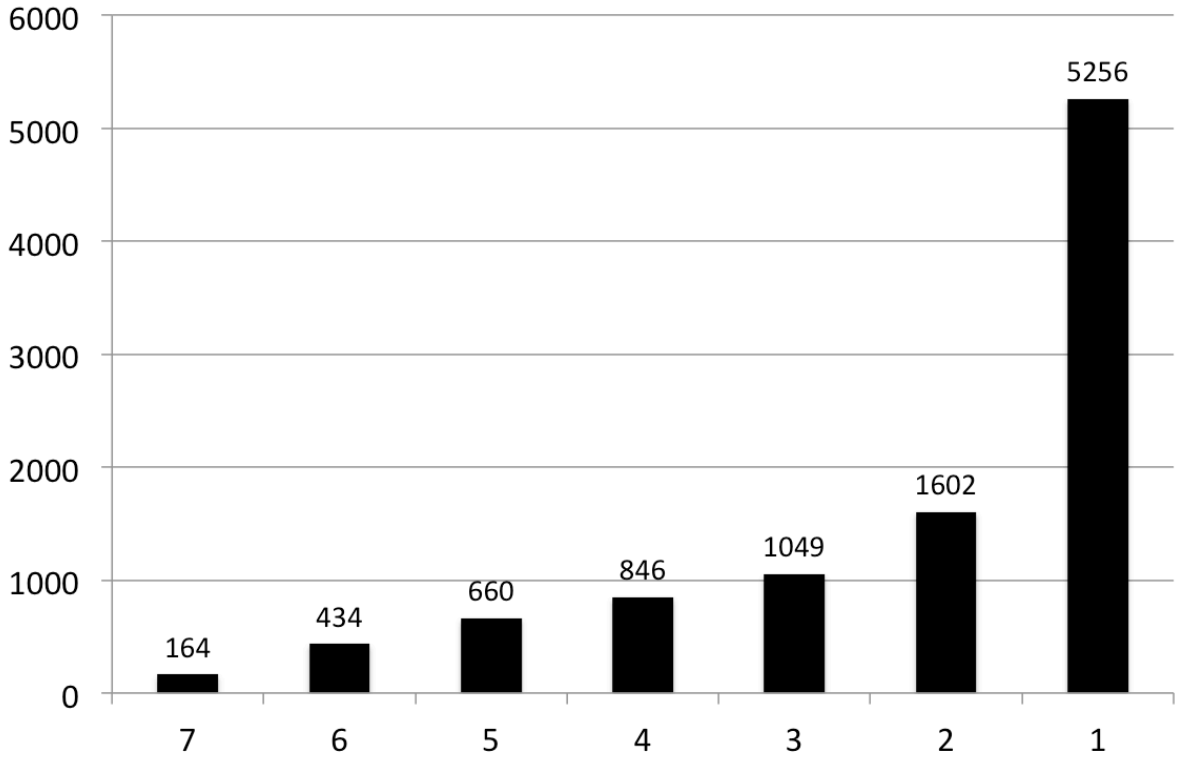

Abb. 1: Anzahl der Wörter (y-Achse), die in n Grundwortschätzen (x- Achse) vorkommen

Dies zeigt auch ein Blick auf die Schnittmengen zwischen den einzelnen Grundwortschätzen, die mittels einer selbst programmierten Software berechnet wurden (Tab. 1): Die Übereinstimmungen liegen in einem Bereich zwischen 13 % und 73 %. Die Selektion des Wortschatzes erfolgte also offenbar nach sehr unterschiedlichen Kriterien. Die Schwankungen in den Werten lassen sich zwar teilweise darauf zurückführen, dass die Umfänge der Wortschätze sehr unterschiedlich sind, dennoch verweisen die Ergebnisse insgesamt darauf, dass es offenbar an empirischen Grundlagen für die Zusammenstellung von Grundwortschätzen fehlt.

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

	Baldegger Kontakt-schwelle	*Hiratsuka 4000 Wörter*	*Langen-scheidt: Basic German Vocabulary*	*Lübke: Lernwort-schatz Deutsch*	*Reimann / Dinsel: Großer Lern-wortschatz*	*Tschirner: Grund- und Aufbauwort-schatz*	*Schaum's Outline of German Vocabulary*
Baldegger Kontakt-schwelle	100 %	43.3 %	71.8 %	67,9 %	69,2 %	57.3 %	27.7 %
Hiratsuka: 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 %

Noch deutlicher wird dieser Mangel an empirischer Grundierung, wenn man die thematische Gliederung der Grundwortschätze vergleicht. Hier scheint vollkommene Willkür zu herrschen. Erwin Tschirner beispielsweise gliedert seinen *Grund- und Aufbauwortschatz* (2008) grob nach folgenden Themen: *Allgemeine Begriffe, Arbeitswelt, Ausbildung, Einkaufen, Freizeit und Unterhaltung, Körper und Gesundheit, Personalien, Informationen zur Person, Persönliche Beziehungen und Kontakte, Politik und Gesellschaft, Reisen und Verkehr, Sprache, Strukturwörter, Umwelt, Verpflegung, Wahrnehmung und Bewegung, Wohnen, Öffentliche und private Dienstleistungen. Schaum's Outline of German Vocabulary* (2009) hingegen gliedert den Wortschatz wie folgt: *Anruf, Arzt, Badezimmer, Bank, Computer, Esszimmer, Friseur, Hausarbeit, Kaufhaus, Krankenhaus, Küche, Passkontrolle und Zoll, Post, Problem, Restaurant, Schlafzimmer, Sport, Theater, Wohnzimmer, Wäsche, am Bahnhof, am Flughafen, auf der Post, das Auto, die Familie, im Flugzeug, im Hotel, nach dem Weg fragen.* Während also Tschirners thematische Gliederung auf relativ abstrakten Kategorisierungen beruht (die in Unterkapiteln weiter differenziert und konkretisiert werden), ist *Schaum's Outline of German Vocabulary* schon auf der ersten Gliederungsebene an konkreten Kommunikationssituationen orientiert.
Der frequenzorientierte Ansatz erlaubt einen noch genaueren Blick auf die Zuordnung von Wörtern zu Themen bzw. Kommunikationssituationen in Grundwortschätzen. Tabelle 2 zeigt die Schnittmengen der Grundwortschätze von Tschirner (*Grund- und Aufbauwortschatz* 2008) und Lübke (*Lernwortschatz* 2008) auf der ersten thematischen Gliederungsebene als eine Heatmap: Je größer die Überschneidungen des Wortschatzes in den jeweiligen Themengebieten sind, desto dunkler sind die Felder der Tabelle unterlegt.

Tab. 2: Schnittmengen des Wortschatzes in der thematischen Gliederung von Tschirner (2008, Spalten) und Lübke (2008, Zeilen) nach deren thematischer Gliederungssystematik (erste Gliederungsebene)

	T: Allge-meine Begriffe	T: Arbeit swelt	T: Ausbi ldung	T: Einka ufen	T: Freize it und Unter-haltun g	T: Körper und Gesun dheit	T: Perso-nalien, Inform ationen zur Person	T: Persö nliche Bezie-hunge n und Kon-takte	T: Politi k und Gesell schaft	T: Reisen und Verke hr	T: Sprach e	T: Struktu rwörter	T: Umwe lt	T: Verpfl egung	T: Wahrnehmu ng und Bewegung	T: Wohne n	T: Öffent liche und privat e Diens t-leistu ngen
L: Allgemeine Begriffe	6.25	0	0	3.125	3.125	3.125	3.125	0	9.375	0	0	18.75	6.25	0	0	9.375	6.25
L: Beruf	3.676	20.588	13.235	2.205	2.941	2.205	5.147	3.676	4.411	2.941	0.735	5.147	1.47	0.735	0.735	0.735	4.411
L: Denken	7.943	1.401	10.28	0.467	0.467	1.869	7.476	2.803	8.411	0.467	25.7	10.747	0.467	0	1.869	0	3.738
L: Ernährung	1.333	0	0.444	2.222	0	2.222	0	0.444	0	0	0	0	0.444	21.333	2.222	0.888	0
L: Ethik, Religion	5.555	0	1.851	0	1.851	0	27.777	0	3.703	1.851	0	1.851	1.851	0	1.851	0	0
L: Freizeit	2.836	4.964	0.709	0.709	43.262	0	12.056	0	0	1.418	1.418	1.418	1.418	0	2.127	2.836	1.418
L: Gesundheit, Krankheit	2.409	1.204	1.204	0	1.204	40.963	6.024	0	0	0	0	0	0	0	0	0	1.204
L: Kleidung	1.204	0	0	28.915	0	1.204	1.204	0	0	0	0	0	0	0	1.204	6.024	0
L: Körper	6.451	0	0	0	1.935	27.741	4.516	0	0	0.645	1.29	1.29	3.87	0	12.258	2.58	0
L: Natur	3.076	1.538	0	0	3.076	0	1.538	0	3.076	3.076	0	0	49.23	0	1.538	0	0
L: Nicht Verwechse	0	0	0	0	0	0	0	0	33.333	0	33.333	0	0	0	0	0	0

ln																		
L: Privatleben	1.185	0.395	1.185	0	7.905	3.557	18.972	13.833	1.581	0.79	2.371	3.557	0.79	0	7.905	0.395	1.976	
L: Schulwesen	1.388	2.777	51.388	0	0	1.388	2.777	0	8.333	1.388	0	2.777	0	0	0	0	5.555	
L: Seelischer und geistiger Bereich, Gefühle	8.433	0	0	0	3.614	6.024	10.843	6.024	3.614	2.409	2.409	1.204	1.204	0	20.481	0	8.433	
L: Sprache	1.702	2.978	5.106	0	12.34	0.851	1.276	1.702	2.978	0.425	19.574	6.808	0	0	1.702	0	8.51	
L: Tiere, Pflanzen	0	1.123	0	1.123	1.123	0	2.247	0	0	1.123	0	0	29.213	1.123	0	1.123	0	
L: Umwelt	8.571	0	0	0	0	0	0	0	0	0	0	0	45.714	2.857	8.571	0	0	
L: Verkehr	1.557	3.115	0.623	1.246	19.937	0	5.607	0	0.311	17.757	0.934	2.492	1.246	0	3.115	2.18	2.18	
L: Wille	5.882	3.529	0	1.176	0	1.176	5.882	7.058	12.941	2.352	21.176	12.941	0	0	1.176	0	4.705	
L: Wirtschaft	1.65	6.6	2.31	19.471	2.97	1.65	0.99	1.98	6.27	2.64	2.31	1.65	0.99	1.65	2.31	4.95	6.6	
L: Wohnen	1.986	1.986	0.662	3.311	0.662	0	1.986	0	1.324	0.662	0	1.324	5.298	0.662	3.311	37.086	2.649	
L: Zeit, Raum, Menge	45.301	0.24	1.445	2.65	4.096	0.722	2.168	1.445	2.65	6.506	2.168	4.337	1.445	0.24	3.132	1.204	0.963	
L: Öffentliches Leben	1.041	3.645	0.52	0.52	4.166	0.52	4.166	0.52	36.979	0	0	0	1.562	0.52	0.52	0.52	16.145	

Die größte Übereinstimmung im Wortschatz besteht demnach in der Schnittmenge der Themen *Schulwesen* (Lübke) und *Ausbildung* (Tschirner) mit gerade einmal 51,4 %. Es zeigt sich zudem, dass der Wortschatz aus Tschirners Kategorie *Ausbildung* sich bei Lübke über die thematischen Kategorien *Schulwesen* (51,4 %), *Beruf* (13,2 %), *Denken* (10,3 %) und *Sprache* (5,1 %) verteilt. Das Potenzial einer solchen Heatmap liegt auch darin, sichtbar zu machen, welche Kategorien besonders stark streuen. So zeigt sich etwa, dass Tschirners Kategorie *Allgemeine Begriffe* mit der gleichnamigen Kategorie bei Lübke nur wenige Überschneidungen aufweist, der Wortschatz hier vielmehr sehr unterschiedlichen Kategorien zugewiesen wird, insbesondere *Zeit, Raum, Menge*. Die Wörter aus Lübkes Kategorie *Allgemeine Begriffe* hingegen streuen bei Tschirner noch stärker, soweit sie überhaupt vorkommen. Die größte Dispersion weisen die Lexeme in Lübkes Themen *Wille* und *Denken* über die thematischen Kategorien bei Tschirner auf.

Der frequenzorientierte Ansatz erlaubt jedoch nicht nur Rückschlüsse aus aggregierten Daten, sondern ermöglicht es auch, für jedes einzelne Wort die Zuordnung zu einzelnen thematischen Klassen in den Blick zu nehmen. Tabelle 3 zeigt einen Ausschnitt aus einer Liste, die für jedes Lemma angibt, ob ein Wort in einem der untersuchten Grundwortschätze vorkommt und wenn ja in welchen thematischen Kategorien.

Tab. 3: Zuordnung einzelner Lemmata zu thematischen Kategorien in sieben Grundwortschätzen

	Baldegger Kontakt-schwelle	*Hiratsuka: 4000 Wörter*	*Langen-scheidt: Basic German Vocabulary*	*Lübke: Lern-wortschatz*	*Reimann / Dinsel: Großer Lern-wortschatz*	*Tschirner: Grund- und Aufbauwort-schatz*	*Schaum's Outline of German Vocabulary*
Herz	Gesundheit und Hygiene	Der menschliche Körper	Der Mensch	Körper	Gesundheit und Krank-heit / Der Mensch	Körper und Gesundheit	Arzt
Heu	--	Dorf und Feldarbeit	--	--	Stadt und Land	--	--
Hilfe	Aktualität; Themen von Allgemeinem	--	Mensch und Gesellschaft	Beruf / Seelischer und geistiger Bereich,	Der Mensch	Öffentliche und private Dienst-	--

	Interesse			Gefühle		leistungen	
Himbeere	--	Garten	--	--	--	--	Restaurant
Himmel	Umwelt	Religion / Himmel und Gestirne	Umwelt / Öffentliches Leben	Natur / Ethik, Religion	Erde und Weltraum / Kulturelles Leben	Umwelt	--
Hintergrund	--	--	Räumliche Begriffe	--	Kulturelles Leben	Politik und Gesellschaft	--
Hitze	--	Wetter	Umwelt	Umwelt	Erde und Weltraum	Umwelt	Küche
Hobby	Freizeit und Unterhaltung	--	Kunst und Interessen	Freizeit / Verkehr	Feste und Freizeit	Personalien, Informationen zur Person	--
Hochwasser	--	Unfälle und Katastrophen	Umwelt	--	Erde und Weltraum	--	--
Hochzeit	--	Familie	Mensch und Gesellschaft	Privatleben	Feste und Freizeit / Die Familie	Freizeit und Unterhaltung	--
Hof	Wohnen	Wohnung und Möbel	Umwelt / Alltagswelt / Mensch und Gesellschaft	Wohnen	zu Hause	Wohnen	im Hotel
Holz	Wohnen	Feldblumen und wilde Pflanzen	Technik und Materialen	Tiere, Pflanzen	Stadt und Land	Umwelt	--

Die Tabelle 3 zeigt, dass sich die Wortschätze grundsätzlich danach unterscheiden lassen, ob ein Lemma in mehreren Themenkategorien vorkommen kann oder nicht, und damit auch, wie differenziert die Bedeutungsbeschreibung in den jeweiligen Grundwortschätzen ist. Ein Lemma wie „Himmel“ kann demnach entweder ausschließlich den Kategorien *Umwelt/Natur/Erde und Weltraum* zugeschrieben oder zusätzlich den Kategorien *Ethik/Ethik, Religion/Kulturelles Leben*. Zusammenfassend können wir festhalten, dass hinsichtlich des Lehrbuchtyps Grundwortschatz der frequenzorientierte Ansatz eine dreifache diagnostische Funktion haben kann:

1. die Kohärenz der Wortschätze bzw. Selektivität eines einzelnen Grundwortschatzes zu analysieren, um die Passgenauigkeit des Wortschatzes für bestimmte Zielgruppen zu bestimmen,
2. die Kohärenz der thematischen Gliederung zu überprüfen: kultur- oder zielgruppenspezifische Gliederungen sichtbar machen, offensichtliche Restkategorien identifizieren,
3. die Differenziertheit der semantischen Beschreibung zu messen, indem das Vorkommen der Lemmata in unterschiedlichen thematischen Kategorien untersucht wird.

Gleichzeitig bietet der lehrbuchanalytisch orientierte frequenzorientierte Ansatz auch bislang zu wenig genutzte Möglichkeiten, die zur Erstellung bzw. Verbesserung von Grundwortschätzen beitragen könnten:

1. die Möglichkeit, durch Schnittmengenberechnung im Sinne des lexikographischen Ansatzes (Haderlein 2008; Schnörch 2002) einen Beitrag zur Bestimmung des zentralen Wortschatzes zu leisten,
2. durch *topic modelling* (Steyvers/Griffiths) datengeleitet thematische Gliederungen für das für die Zielsprache als repräsentativ angesehene Korpus zu berechnen.

Analyse von Lehrbüchern

Im Folgenden wollen wir zeigen, welchen Beitrag eine frequenzorientierte Herangehensweise für die Analyse allgemeiner Lehrwerke für Deutsch als Fremdsprache leisten kann. Dabei werden wir die Wortschatzselektion, den Wortschatzaufbau und die Kommunikationsbereichsspezifik des Wortschatzes in Lehrbüchern in den Blick nehmen. Vorher jedoch müssen wir die Daten, auf denen unsere Analysen beruhen, genauer beschreiben.

Lehrbuchdaten

Die Basis für die vorliegende Untersuchung bilden mehrere, vor allem japanische DaF-Lehrwerke, die digitalisiert vorliegen. Die Erarbeitung dieser Daten geht auf ein Vorgängerprojekt zurück, bei dem die Bücher gescannt und mit einer OCR-Software in digitalen Text konvertiert wurden (Bubenhofer u. a. XX). Es handelt sich um folgende Bücher:

1. *Ein Sommer in Deutschland*. Herausgegeben von Kurahei Ogino, Andrea Raab. 4. Aufl., Asahi, Tokyo 2009. (Im Folgenden: *Sommer.*)
2. *Farbkasten Deutsch neu 1*. Herausgegeben von Mayumi Itayama, Ursula Shioji, Yuko Motokawa, Takako Yoshimitsu. 26. Aufl., Sanshusha, Tokyo 2007. (Im Folgenden: *Farbkasten.*)
3. *Hallo München. Neu*. Herausgegeben von Ichiro Sekiguchi. Hakusuisha, Tokyo 2008. (Im Folgenden: *München*.)
4. *Meine Deutschstunde*. Herausgegeben von Tomoaki Seino. 4. Aufl., Asahi, Tokyo 2008. (Im Folgenden: *Deutschstunde.*)
5. *Modelle neu 1*. Herausgegeben von Andreas Riessland u. a. 6. Aufl., Sanshusha, Tokyo 2009. (Im Folgenden: *Modelle.*)
6. *Szenen 1*. Herausgegeben von Shuko Sato u. a. 13. Aufl., Sanshusha, Tokyo 2009. (Im Folgenden: *Szenen 1.*)
7. *Szenen 2*. Herausgegeben von Shuko Sato u. a. 13. Aufl., Sanshusha, Tokyo 2009. (Im Folgenden: *Szenen 2.*)
8. *Themen 1 neu. Kursbuch*. Herausgegeben von Hartmut Aufderstraße u. a. Hueber, Ismaning 2003. (Im Folgenden: *Themen*.)
9. *em neu. Hauptkurs*. Herausgegeben von Michaela Perlmann-Balme, Susanne Schwalb. Hueber, Ismaning 2008. (Im Folgenden: *em*.)

Die Texte wurden mit Metadaten (Autorinnen und Autoren, Erscheinungsjahr, Zielgruppe) ausgezeichnet, japanischer Text ausgesondert und die Lektionen und weitere Untergliederungen (Lesetext, Übung, Grammatik, Wortschatz etc.) markiert. Zudem wurden die Texte maschinell mit Wortarten und Lemmata annotiert unter Verwendung des TreeTaggers (vgl. Schmid 1994). Dabei sollte man sich bewusst sein, dass an mehreren Stellen der Korpusaufbereitung Fehler entstehen können: Bereits beim OCR-Prozess können Erkennungsfehler auftreten, ebenso bei der maschinellen Wortartenerkennung und der Lemmatisierung.
Mit diesen Schritten werden Lehrwerke für vielfältige Analysen erschlossen. Unsere Untersuchungen beschränken sich zwar auf den verwendeten Wortschatz, doch wären aufgrund des POS-Taggings (Wortartenannotation, die teilweise auch syntaktische Funktionen beschreibt) ebenso syntaktische Analysen durchführbar. Fragen des ‚Weltbildes' in einem Lehrbuch – und dies trifft nicht nur auf Lehrwerke für Deutsch als Fremdsprache zu – könnte man beispielsweise über N-Gramm-Analysen (vgl. Scharloth/Bubenhofer 2012) empirisch weiter ergründen.
Im Verlauf der datengeleiteten Korpusanalysen haben wir Wortschatzlisten nach

unterschiedlichen Kriterien, z. B. Frequenzlisten für ganze Lehrbücher, Wortschatzlisten nach Kapiteln oder nach grammatischen Kategorien (POS), generiert. Die Tabelle 4 gibt eine schematische Übersicht über die Schritte bei der Korpuserstellung, -aufbereitung und -analyse:

Tab. 4: Workflow beim Pre-Processing der Lehrwerke

SCHRITT		MITTEL	ERGEBNIS	KOMMENTAR
Korpuserstellung				
1	Scan	ScanSnap S510	Bilddatei	Nicht durchsuchbar
2	OCR	Omnipage 16	Textdatei	Nach Text durchsuchbar
3	Annotation Textstruktur	Manuell	XML-Datei	Nach Tags und Text durchsuchbar
4	Annotation Tokenebene: Lemmatisierung, POS-Annotation	TreeTagger	XML-Datei	Nach Lemmata und POS durchsuchbar
Korpusauswertung				
5	Wortlisten-Generierung	Perl-Skript	csv-Datei	Grundsätzlich sind Listen auf der Basis jeder annotierten Metainformations-Kategorie für jede Kategorie auf der Tokenebene möglich
6	Wortlisten-Vergleich	Perl-Skript	csv-Datei	

Grundwortschatzdaten

Um einen Maßstab für die Distribution von Wortschatz in den Lehrwerken zu haben, wurde der von einem Teil der Autoren dieses Beitrags datengeleitet errechnete Kernwortschatz des Deutschen herangezogen (vgl. Okamura/Lange/Scharloth 2012). Das von der Japanese Society for the Promotion of Science finanzierte Forschungsprojekt *Basic German Vocabulary for Foreign Language Learners: A data-driven Approach*[3] (im Folgenden auch *Basic German*) hat das Ziel, den Kernwortschatz des Deutschen nicht nur auf der Basis der Häufigkeit von Lexemen zu berechnen, sondern auch ihre thematische und temporale Stabilität sowie ihre Produktivität in die Berechnung des Kernwortschatzes einfließen zu lassen. Zum zentralen Wortschatz zählen demnach jene Lexeme, die (1) häufig vorkommen, die (2a) über einen längeren Zeitraum gleichmäßig häufig auftreten (also keine Modewörter sind), (2b) in Texten unterschiedlicher thematischer Prägung gleichmäßig distribuiert sind, die (3a) als lexikalische Morpheme in vielen Ableitungen und Zusammensetzungen auftreten, die (3b) als Lexeme selbst häufig sind und (3c) die als lexikalische Morpheme häufiger als Zweitglied in Komposita verwendet werden. Dabei wurden die folgenden Werte normalisiert (teilweise logarithmiert) und gewichtet berechnet und mit Hilfe eines Vektordistanzmodells[4] ein Ranking der Lexeme erstellt:

3 Vgl. die Internetpräsenz: http://www.basic-german.com [Stand: 10.09.2014].

4 Für eine detaillierte Beschreibung der zugrunde liegenden statistischen Modelle vgl. Lange/Okamura/Scharloth (2015 i. E.).

- *Frequenz:* Häufigkeitsklasse eines Lexems
- *Temporale Stabilität:* das Dispersionsmaß Gries' DP (vgl. Gries 2008) jahresweise
- *Thematische Stabilität:* Gries' DP über Rubriken und Teilforen
- *Produktivität:* Anzahl der Komposita-Lemmata, Frequenz der Komposita, Häufigkeitsklassenverteilung der Komposita (Entropie), Verteilung Erst-/Zweitglied

Bei der Zusammenstellung des Korpus, auf dessen Basis der Kernwortschatz berechnet wurde, gingen die Autoren von zwei kommunikativen Grundkonstellationen aus: Einerseits nahmen sie mehrfachadressierte und konzeptionell schriftliche Texte in das Korpus auf, andererseits aber auch Texte, die persönlich adressiert und konzeptionell mündlich sind. So setzte sich das insgesamt rund 850 Millionen Wörter umfassende Korpus aus Zeitungstexten der Jahre 1990 bis 2012 (370 Millionen Wortformen, Tab. 6) sowie aus Online-Diskussionsforen aus den Jahren 1998 bis 2012 (rund 480 Millionen Wortformen, Tab. 6) zusammen.

Tab. 5: 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
politikforen.net	3.260.363	263.866.105
Gesamt Foren:	*5.985.086*	*474.067.581*

Tab. 6: 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*

Der Kernwortschatz wurde sowohl für das gesamte Korpus als auch für die beiden Kommunikationsbereiche getrennt berechnet. Für die folgende Lehrwerksanalyse wurden die Daten analog zu Häufigkeitsklassen auf der Basis eines konstanten Vektordistanzintervalls in Klassen eingeteilt (im Folgenden: Vektordistanzklassen/Vks). Angehörige der gleichen Vektordistanzklasse haben in der Zusammenschau der berechneten Merkmale ähnliche Frequenz-, Stabilitäts- und Produktivitätswerte, ihre Distanz zum optimalen Vektor (höchste Häufigkeit, optimale Stabilität, höchste Produktivitätswerte) ist daher ähnlich. Weil die Differenz der Vektordistanzen zwischen zwei Lexemen im Ranking immer mehr abnimmt, nimmt die Anzahl der Vertreter je Klasse bei gleichbleibendem Vektordistanzintervall zu. Tabelle 7 gibt einen Überblick über die Distribution von Lexemen in den Vektordistanzklassen.

Tab. 7: Übersicht über die Vektordistanzklassen

Klasse	*Anzahl Lexeme*	*Summe Lexeme*
0	3	3
1	2	5
2	14	19
3	23	42
4	48	90
5	123	213
6	206	419
7	412	831
8	704	1535
9	1066	2601
10	1509	4110
11	2135	6245
12	2894	9139
13	4284	13423
14	5734	19157
15	7979	27136
16	10044	37180

Wortschatzselektion in den untersuchten Lehrwerken

Zunächst ist die Frage von Interesse, welche Vektordistanzklassen in den Lehrbüchern vorkommen. Abbildung 2 zeigt zum einen die Anzahl unterschiedlicher Lexeme, die in einem Lehrbuch auftreten, zum anderen auch, auf welche Vektordistanzklassen sich diese Lexeme verteilen.

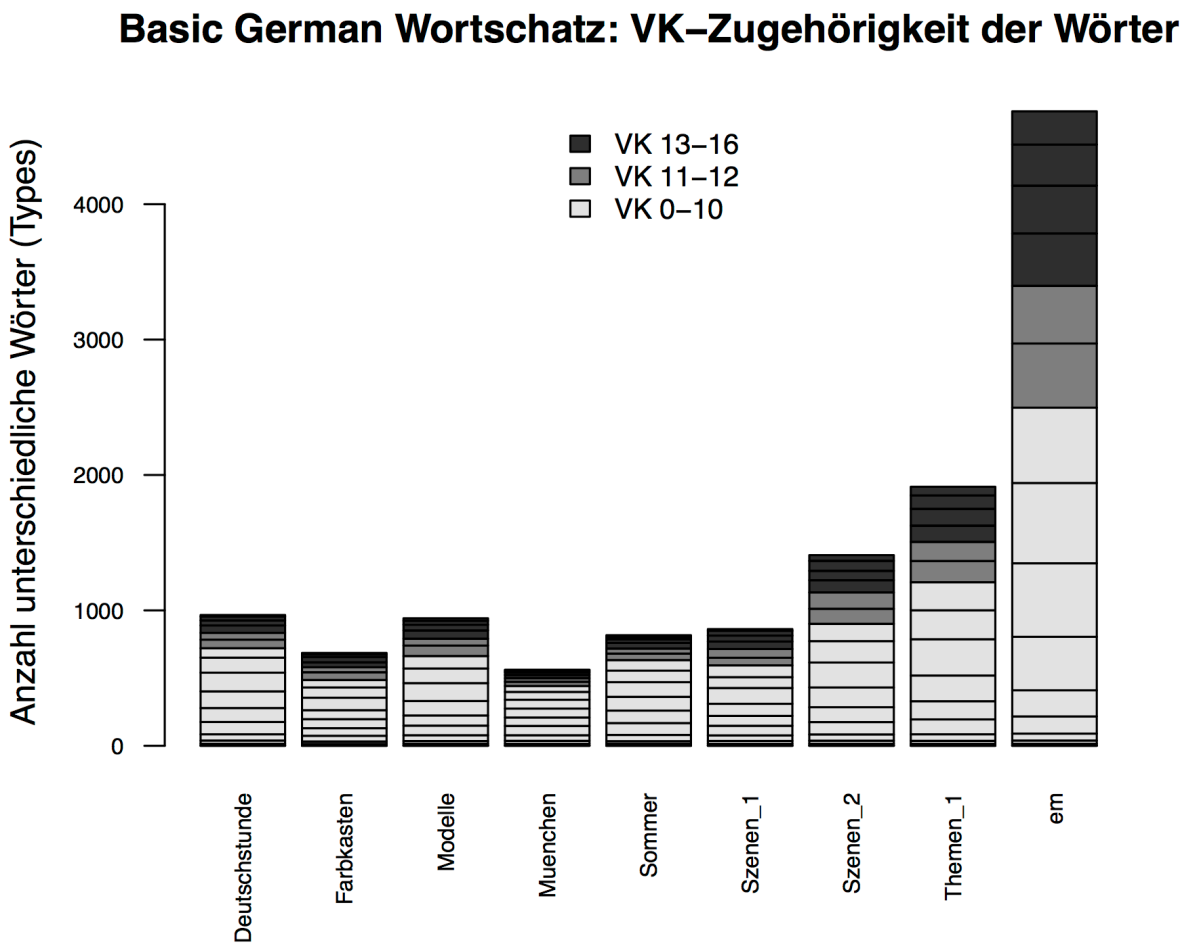

Abb. 2: Verteilung des Vokabulars über die VKs gemäß Grundwortschatz

Das Ergebnis ist für die meisten Lehrbücher positiv: Sie berücksichtigen zu einem sehr hohen Prozentsatz Lexeme aus den Vektordistanzklassen 1 bis 10; diese umfassen die rund 4000 häufigsten, stabilsten und produktivsten Lexeme. Je umfangreicher die Lehrbücher werden, desto mehr Wortschatz aus höheren Vektordistanzklassen wird verwendet. Dies ist nicht weiter überraschend, dennoch erscheint die Zahl der Lexeme aus den Vektordistanzklassen 13 bis 16 bei *Themen 1* und *em* unnötig hoch. Aus Lernersicht wäre es besser, diesen Wortschatzbereich zunächst zugunsten der Vektordistanzklassen 0 bis 12 zu vermeiden, also der rund 10.000 frequentesten, stabilsten und produktivsten Lexeme.

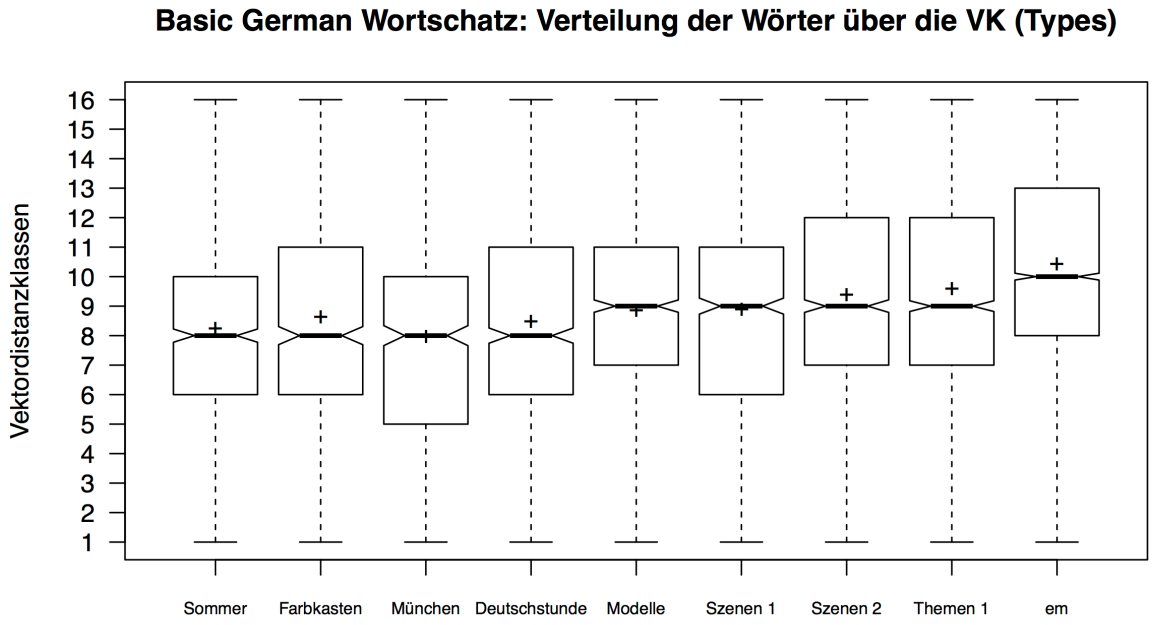

Abb. 3: Mediane (dicker waagrechter Balken), Mittelwerte (+) und Begrenzung des oberen und unteren Quartils (Box) der Vektordistanzklassen in den Lehrbüchern; 50% der Daten liegen innerhalb der Box; unterschiedliche Wörter (Types)

In Abbildung 3 ist ersichtlich, dass sich die Lehrwerke *Sommer*, *Farbkasten*, *München* und *Deutschstunde* nicht signifikant in der Wortschatzselektion unterscheiden. Eine zweite Gruppe sind die Bücher *Modelle*, *Szenen 1* und *Szenen 2* sowie *Themen 1*, die gegenüber der ersten Gruppe ein schwierigeres Vokabular verwenden. *em* wiederum unterscheidet sich diesbezüglich von allen anderen und verwendet das schwierigste Vokabular innerhalb der untersuchten Lehrwerke. Natürlich sind die Ergebnisse nur im Kontext des Umfangs des jeweils im Lehrwerk eingeführten Vokabulars aussagekräftig. Aber im Vergleich etwa zu *Deutschstunde* ist die Wortschatzauswahl von *Modelle* offensichtlich weniger gelungen, zumindest wenn man die Daten aus dem Forschungsprojekt *Basic German Vocabulary for Foreign Language Learners: A data-driven Approach* zugrunde legt und voraussetzt, dass Lehrwerke sich bei der Wortschatzauswahl am zentralen Wortschatz orientieren sollten.
Die Analysen zeigen also deutliche Unterschiede in den Lehrwerken, die auch Rückschlüsse auf die Qualität der Wortschatzselektion zulassen. Der frequenzorientierte Ansatz ist damit geeignet, die Selektion des Wortschatzes in Lehrbüchern mit intersubjektiv nachvollziehbaren Kriterien zu analysieren, zu kritisieren oder gar anzuleiten.

Wortschatzaufbau

Der frequenzorientierte Ansatz ermöglicht es auch, den Wortschatzaufbau in Lehrbüchern in den Blick zu nehmen. In unserer Analyse waren wir von der Annahme ausgegangen, dass Wortschatz aus höheren Vektordistanzklassen sich häufiger in den späteren Lektionen der Lehrbücher findet. Dies ist jedoch nicht der Fall, vielmehr zeigt sich in fast allen Lehrbüchern ein stabiles Distributionsmuster: Rund 40% des Vokabulars stammt aus den VKs 0 bis 5, weitere 40% aus den VKs 6 bis 10 und rund 20% aus den VKs 11 bis 16 (vgl. exemplarisch Abb. 4).

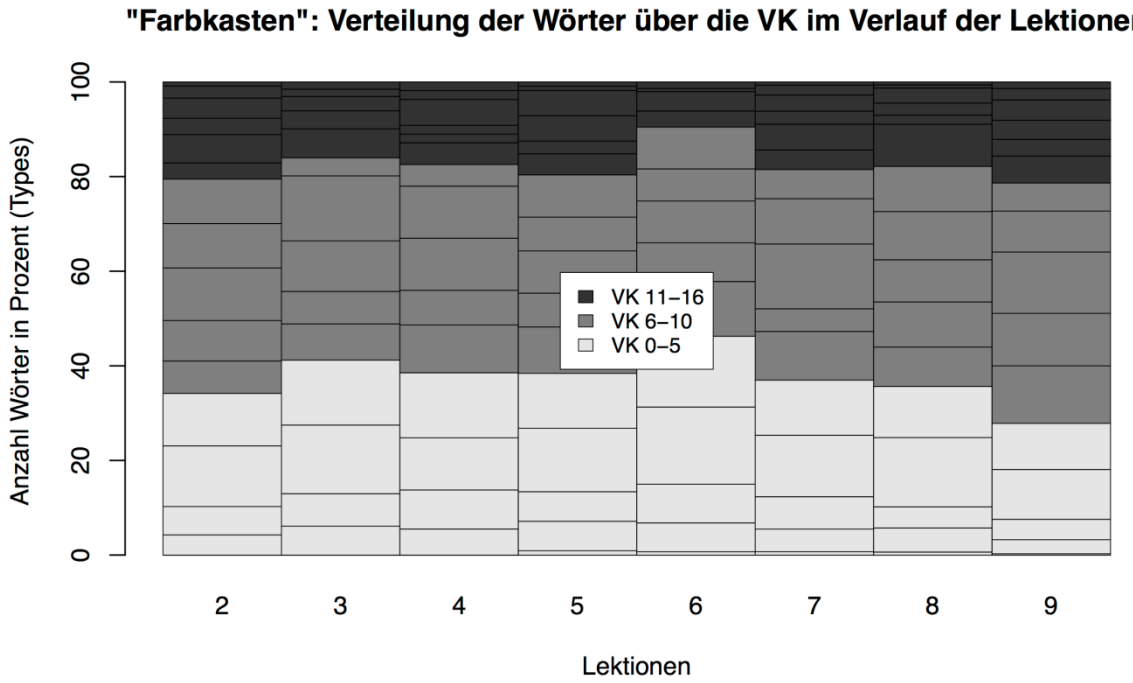

Abb. 4: Lektionenweise Verteilung des Vokabulars über Vektordistanzklassen im Lehrwerk *Farbkasten*

Eine Ausnahme bildet wiederum das Lehrwerk *em*. In ihm wird von Anfang an sehr viel mehr Wortschatz aus den Vektordistanzklassen 6 bis 10 und 11 bis 16 eingeführt (Abb. 5).

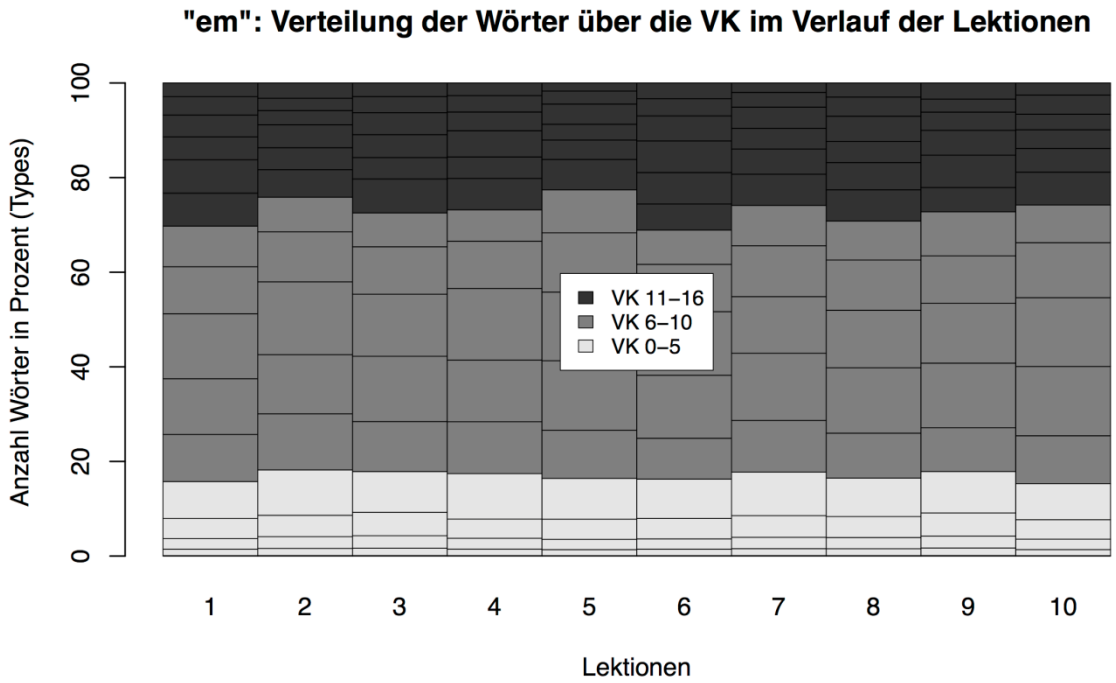

Abb.5: Lektionenweise Verteilung des Vokabulars über Vektordistanzklassen im Lehrwerk *em*

Wortschatz in Abhängigkeit von kommunikativen Grundkonstellationen

Wie bereits dargestellt, wurde im Rahmen des Projekts *Basic German Vocabulary for Foreign*

Language Learners: A data-driven Approach der Kernwortschatz auch getrennt nach den kommunikativen Grundkonstellationen mehrfachadressierend und konzeptionell schriftlich einerseits und persönlich adressiert und konzeptionell mündlich andererseits berechnet. Dies bietet die Möglichkeit, die Lehrwerke auch daraufhin zu untersuchen, zu welcher der beiden Grundkonstellationen ihre Wortschatzselektion neigt.
Die Ergebnisse, die in den Abbildungen 6 und 7 visualisiert sind, widersprechen der Annahme, dass die Lehrbücher für einen kommunikativ-pragmatisch orientierten Unterricht konzipiert sind: Am Maßstab Grundwortschatz „Zeitungen" gemessen, bewegt sich bei den meisten Lehrbüchern das Vokabular hauptsächlich zwischen den Klassen 5 und 10/11, während sich, gemessen am Grundwortschatz „Foren", bei den meisten Lehrbüchern das Vokabular zwischen 7 und 11/12 bewegt. Die Lehrwerke orientieren sich demnach eher an der mehrfachadressierenden konzeptionell schriftlichen kommunikativen Grundkonstellation als an der persönlich adressierenden, konzeptionell mündlichen – ein Befund, der für Lehrwerke, die für einen am kommunikativ-pragmatischen Paradigma orientierten Unterricht konzipiert sind/erstellt wurden/…, überrascht.

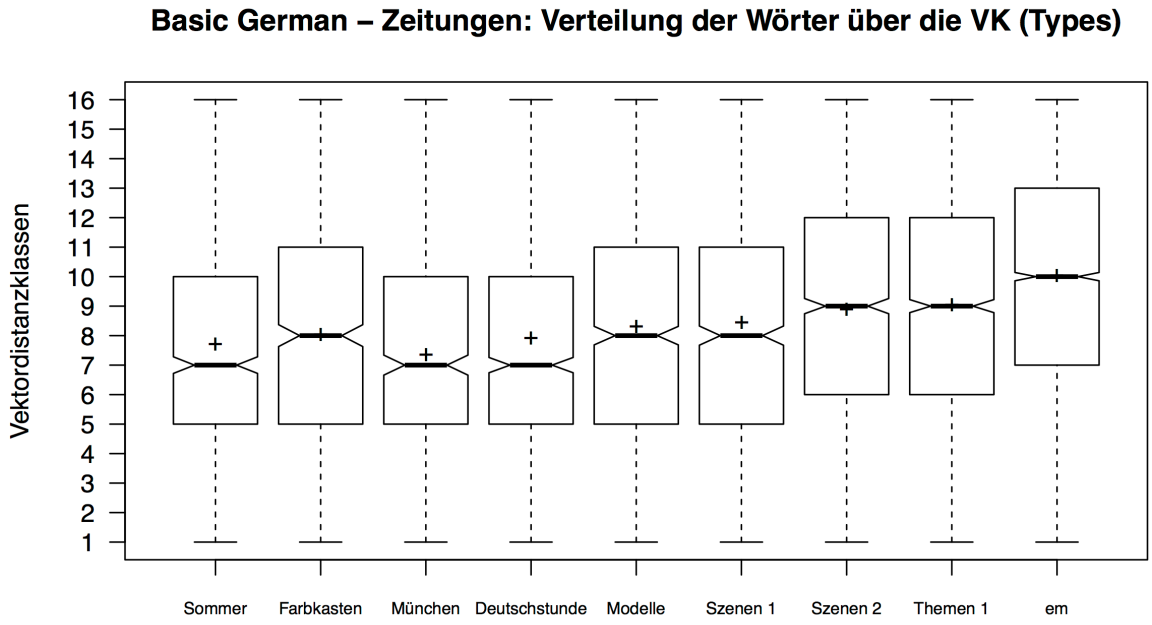

Abb.6: Zeitungskorpus – Mediane (dicker waagrechter Balken), Mittelwerte (+) und Begrenzung des oberen und unteren Quartils (Box) der Vektordistanzklassen in den Lehrbüchern

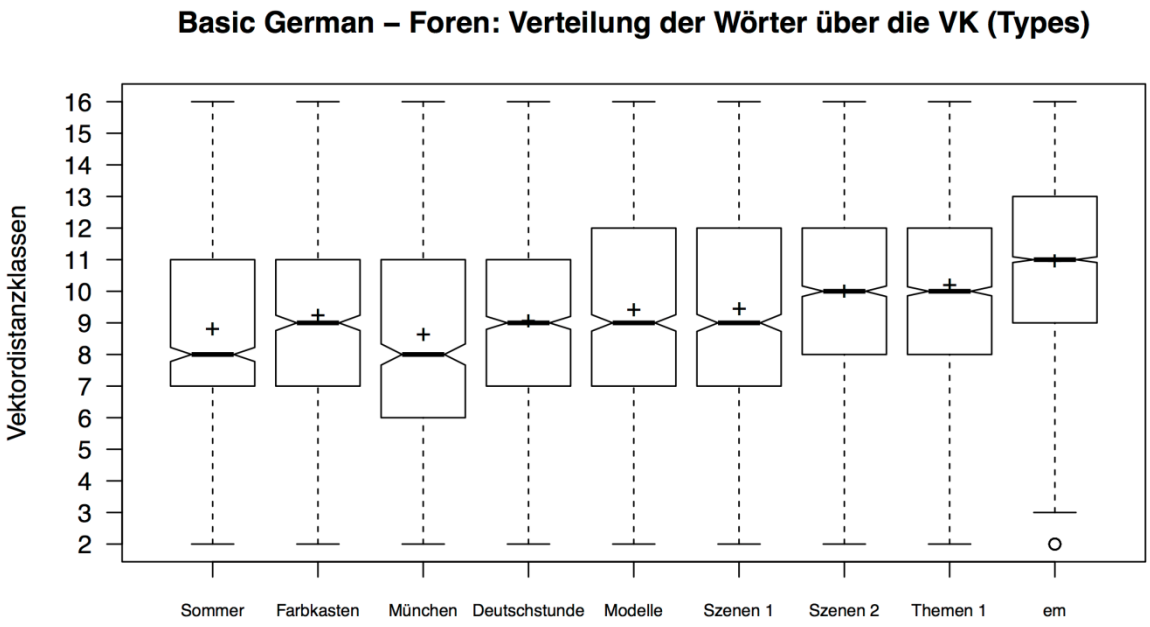

Abb. 7: Forenkorpus – Mediane (dicker waagrechter Balken), Mittelwerte (+) und Begrenzung des oberen und unteren Quartils (Box) der Vektordistanzklassen in den Lehrbüchern

Welches Vokabular, das in den Lehrbüchern Verwendung findet, wird über den Foren-Grundwortschatz besser abgedeckt als über den Zeitungs-Grundwortschatz? Tabelle 8 zeigt Lexeme in grober thematischer Ordnung, die gemäß Foren-Grundwortschatz in eine Vektordistanzklasse von 0 bis 10 fielen (und damit zu den rund 4000 häufigsten, stabilsten und produktivsten Lemmata gehören), gemäß Zeitungs-Grundwortschatz jedoch mit einer Vektordistanzklasse ab 13 ausgewiesen worden sind (und damit für dieses Korpus erst ab einem Rang höher als 13.000 zu finden sind).

Tab. 8: Lexeme aus den Lehrwerken, die bei der Grundwortschatzberechnung auf Basis des Forenkorpus eine niedrige Vektordistanzklasse haben, jedoch eine hohe im Zeitungskorpus

Eigenschaften/Gefühle	*Personen*	*Maße*
Assoziation	Migranten	cm
Dummheit	Partnerin	km
Eifersucht	Staatsbürgerschaft	
aufregen	alleinerziehend	*Anderes*
mitbekommen		Advent
sadistisch	*Essen/Körper*	Hexe
	Erkältung	Ecke
Phatische Kommunikation	Gurke	Grammatik
Glückwunsch	Senf	irgendwas
Hallo	Vitamin	nochmal
Gratulation	Yoga	siehe
Willkommen		welche
		tendenziell
		öfters

Offensichtlich sind Ausdrücke des Begrüßens und Glückwünschens, also Elemente ritueller Kommunikation, die in Zeitungstexten selten vorkommen. Lexeme dieser Art wären in einem kommunikativ-pragmatisch orientierten Grundwortschatz zu erwarten, da sie Alltagssituationen abbilden, die im Unterricht gelernt werden sollen. In eine ähnliche Kategorie gehören Lexeme, die Personen beschreiben oder Gefühle ausdrücken, sowie Themen des

täglichen Lebens wie Gesundheit und Essen (*Erkältung*, *Gurke* etc.).
Umgekehrt stellt sich die Frage, welche Lexeme gemäß Zeitungssprache relativ gebräuchlich sind, in der Forensprache jedoch selten vorkommen. Tabelle 9 zeigt Lexeme in den untersuchten Lehrbüchern, die gemäß Zeitungs-Grundwortschatz eine Vektordistanzklasse von 0 bis 10 aufweisen, nach dem Foren-Grundwortschatz jedoch eine Klasse ab 13.

Tab. 9: Lexeme aus den Lehrwerken, die bei der Grundwortschatzberechnung auf Basis des Zeitungskorpus eine niedrige Vektordistanzklasse haben, jedoch eine hohe im Forenkorpus

Personen	*Verkehr*	*Anderes*
Biografie	Laster	Campus
Porträt	Lieferant	Kabine
	Lkw	Öffnung
Identität	Route	Übersicht
Freiburger		Schale
Mainzer	*Eigenschaften (zuordnen)*	Schirm
koreanisch	bescheinigen	Statue
kroatisch	beschwören	Stift
portugiesisch	entzünden	Strebe
	prophezeien	Ticket
Geschäftsleben	hoffnungsvoll	Vorsprung
Geschäftsführer	innovativ	Wächter
Designer	lässig	Zeichnung
Filiale	rar	
Kanzlei	renommiert	bilanzieren
Mandant		blitzen
Redakteur	*Bewegung*	hüllen
Slogan	besichtigen	proben
Kostüm	besteigen	rangieren
Kundschaft	erkunden	verkleiden
Vertrieb	gleiten	verzögern
Zentrale	kreuzen	überreichen
Zuwachs	münden	
	pendeln	
	pilgern	
	eilig	
	forsch	
	hektisch	

Es handelt sich um einige Lexeme des Geschäfts- und Berufslebens, Themen, die offensichtlich in Webforen seltener, weniger regelmäßig und dann mit nicht sehr differenziertem Wortschatz diskutiert werden. Auffällig sind auch Adjektive und Verben, die Bewegung ausdrücken und die beispielsweise in Erzählungen verwendet werden. Weiter gehören dazu Verben und Adjektive, die Zustände oder Eigenschaften beschreiben und eher zum gehobenen Sprachstil gehören (*entzünden*, *prophezeien*, *beschwören*).
Es zeigen sich deutlich die Unterschiede in der Verteilung des funktionalen Wortschatzes zwischen Zeitungs- und Forensprache: Diese liegen einerseits in den Themen, andererseits in den vorkommenden kommunikativen Situationen. Vor allem letztere sind für den kommunikativ-pragmatisch orientierten DaF-Unterricht von großer Bedeutung. Dabei muss bedacht werden, dass der oben gemachte Vergleich auf der Basis der untersuchten Lehrwerke entstanden ist: Natürlich existieren noch weit mehr Unterschiede zwischen den beiden Wortschätzen – die gezeigten Unterschiede sind aber solche, die für die Praxis der Lehrbücher eine Rolle spielen.

Fazit

Die Analysen sollten das Potenzial frequenzorientierter Ansätze für die Analyse von Lehrwerken, exemplarisch für Wortschatzfragen in Lehrbüchern für Deutsch als Fremdsprache, aufzeigen. *Frequenzorientiert* wurde dabei in einem weiteren Sinn als ‚die Verteilung von Lexemen betreffend' verstanden. Folgende Merkmalstypen kamen dabei zum Einsatz:

- Häufigkeit von Lexemen in Lehrwerken und Lehrwerkteilen/einzelnen Lektionen (z. B. Schnittmengenberechnung in Grundwortschätzen)
- Distribution von Lexemklassen in Lehrwerken und Lehrwerkteilen (z. B. Vektordistanzklassen)
- Zuordnung von Lexemen zu Lexemklassen in unterschiedlichen Lehrwerken (z. B. thematische Gliederung in Grundwortschätzen)

Mit Hilfe dieser Merkmalstypen konnten im Wesentlichen zwei Fragen erörtert werden:

- Wird der Gegenstandsbereich durch die Lehrwerke kohärent konstruiert? – Im Fall der Grundwortschätze kann man konstatieren, dass die Frage, was Teil des Grundwortschatzes sein und wie dieser Grundwortschatz gegliedert werden soll, sehr unterschiedlich beantwortet wird.
- Werden Lehrwerke ihren Ansprüchen gerecht? – Hinsichtlich der Selektion des Wortschatzes wählen die meisten Lehrwerke für Anfängerinnen und Anfänger vorwiegend häufig gebrauchte, stabil vorkommende und produktive Lexeme. Allerdings orientieren sie sich dabei, obwohl sie sich selbst im kommunikativ-pragmatischen Paradigma verorten, noch zu stark an der Zeitungssprache.

Trotz dieser Ergebnisse dürfen jedoch die mit dem Ansatz verbundenen methodischen Probleme nicht aus dem Blick geraten. Der frequenzorientierte Ansatz ist an der sprachlichen Oberfläche orientiert und ebnet semantische Differenzierungen des Wortschatzes ein. Er ist zwar grundsätzlich auch geeignet, grammatikalische Aspekte zu operationalisieren, pragmatische Aspekte sind allerdings außerhalb seiner Reichweite. Darüber hinaus sagen distributive Analysen natürlich nichts über die Qualität der Didaktisierung aus und können nur in ihrem Kontext gedeutet werden. Dennoch haben frequenzorientierte Ansätze das Potenzial, einzelne Aspekte der Lehrbucherstellung stärker zu objektivieren und zum Gegenstand der Reflexion zu machen.

Literaturverzeichnis

Schulbücher

Ein Sommer in Deutschland. Herausgegeben von Kurahei Ogino, Andrea Raab. 4. Aufl., Asahi, Tokyo 2009.

em neu. Hauptkurs. Herausgegeben von Michaela Perlmann-Balme, Susanne Schwalb. Hueber, Ismaning 2008.

Farbkasten Deutsch neu 1. Herausgegeben von Mayumi Itayama, Ursula Shioji, Yuko Motokawa, Takako Yoshimitsu. 26. Aufl., Sanshusha, Tokyo 2007.

Hallo München. Neu. Herausgegeben von Ichiro Sekiguchi. Hakusuisha, Tokyo 2008.

Meine Deutschstunde. Herausgegeben von Tomoaki Seino. 4. Aufl., Asahi, Tokyo 2008.

Modelle neu 1. Herausgegeben von Andreas Riessland u. a. 6. Aufl., Sanshusha, Tokyo 2009.

Szenen 1. Herausgegeben von Shuko Sato u. a. 13. Aufl., Sanshusha, Tokyo 2009.

Szenen 2. Herausgegeben von Shuko Sato u. a. 13. Aufl., Sanshusha, Tokyo 2009.

Themen 1 neu. Kursbuch. Herausgegeben von Hartmut Aufderstraße u. a. Hueber, Ismaning 2003.

Literatur

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

Bubenhofer, Noah u. a. (Jahr): Welcher Wortschatz? Korpuslinguistische Untersuchungen zur Wortschatzselektion japanischer Deutschlehrbücher für Anfänger. In: *Doitsugo Kyoiku – Deutschunterricht in Japan* 16, S. 43-60.

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

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

Gries, Stefan Thomas (2008): Dispersions and adjusted frequencies in corpora. In: *International Journal of Corpus Linguistics* 13. Heft 4/2008, S. 403-437.

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

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

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

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

Lange, Willi/Okamura, Saburo/Scharloth, Joachim (i. E.): Grundwortschatz Deutsch als Fremdsprache: Ein datengeleiteter Ansatz. In: Jörg Kilian/Jan Eckhoff (Hgg.): *Deutscher Wortschatz – beschreiben, lernen, lehren. Beiträge zur Wortschatzarbeit in Wissenschaft, Sprachunterricht, Gesellschaft.* Frankfurt a. M. u. a.

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

Okamura, Saburo/Lange, Willi/Scharloth, Joachim (2012): Methoden der Bestimmung des Kernwortschatzes Deutsch. In: Saburo Okamura/Willi Lange/Joachim Scharloth (Hgg.): Grundwortschatz Deutsch: Lexikografische und fremdsprachendidaktische Perspektiven. Tokyo, S. 29-44. (= Studienreihe der Japanischen Gesellschaft fuer Germanistik 088)

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

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

Rosengren, Inger (1970-1977): *Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt.*

Süddeutsche Zeitung. 2 Bde. Lund.

Scharloth, Joachim/Noah Bubenhofer (2012): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Ekkehard Felder/Marcus Müller/Friedemann Vogel (Hgg.): *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen.* Berlin/New York, S. 195-230.

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees.* Working Paper. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf [Stand: 15.09.2014].

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

Steyvers, Mark/Griffiths, Tom (2007): Probabilistic Topic Models. In: Thomas K. Landauer u. a. (Hgg.): *Handbook of Latent Semantic Analysis.* London, S. 424-440.

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