Preprint von: Bubenhofer, Noah / Willi Lange / Saburo Okamura / Joachim Scharloth (2011): Welcher Wortschatz? Korpuslinguistische Untersuchungen zur Wortschatzselektion japanischer Deutschlehrbücher für Anfänger. In: Doitsugo Kyoiku - Deutschunterricht in Japan, 16, S. 43-60.

Bubenhofer, Noah/ Lange, Willi /Okamura, Saburo/ Scharloth, Joachim:

Welcher Wortschatz? Korpuslinguistische Untersuchungen zur Wortschatzselektion japanischer Deutschlehrbücher für Anfänger.

1. Einleitung

Für Lernende, Lehrende, Lehrwerkautoren und Curriculumplaner ist die „Wortschatzfrage“ eine sehr große Herausforderung. Wie viele Wörter sollen Deutschlernerinnen und Deutschlerner mit welcher Geschwindigkeit lernen? Und welche Wörter sollen zuerst gelernt werden, damit die Lernerinnen und Lerner möglichst schnell in der Lage sind, einfache Texte zu lesen oder in der Alltagskommunikation zu bestehen? In diesem Aufsatz untersuchen wir, welche Antworten japanische Lehrbuchautorinnen und -autoren geben.
Konkret untersuchen wir folgende Fragen:

- Mit welchen Wortschatzmengen werden japanische Deutscherlernende in unterschiedlichen Anfängerlehrbüchern tatsächlich konfrontiert?
- Welche Auftretenshäufigkeit haben die eingeführten Wörter in den Lehrbüchern?
- Welche Überschneidungen gibt es zwischen den Lehrbüchern?
- Lassen sich Lehrbücher im Hinblick auf den in ihnen vorkommenden Wortschatz gruppieren?

Diese kleine Studie soll der Anfang einer intensiveren Beschäftigung mit dem zentralen Wortschatz des Deutschen und seiner Vermittlung im Unterricht für japanische Studierende sein. Die zu Anfang skizzierten normativen Fragen werden nämlich nur selten methodisch abgesichert auf der Basis empirischer Daten entschieden. Unser Aufsatz möchte daher einen Beitrag dazu leisten, dass sich empirisch orientierte Verfahren in der Forschung Deutsch als Fremdsprache in Japan noch stärker etablieren. Außerdem hoffen wir, dass mit dieser Arbeit die DaF-Diskussion im Bereich Lehrwerkanalyse und Lehrwerkkritik wieder belebt wird[1].

[1] „Der Versuch, Lehrwerke einer quantitativen Analyse zu unterziehen und damit eine Verbindung zwischen einer hermeneutischen und einer empirischen Lehrwerkanalyse

Der korpuslinguistische Ansatz bestimmt unsere Vorgehensweise. Einzelne Lehrwerke werden als Subkorpora eines umfangreichen Lehrwerkskorpus betrachtet. Diese Subkorpora vergleichen wir in einem datengeleiteten Verfahren im Hinblick auf die in ihnen vorkommenden Lemmata. In der Forschung im Fach Deutsch als Fremdsprache werden zunehmend vorhandene Korpora der deutschen Sprache genutzt[2]. Dabei interessieren oft Eigenschaften der Sprachverwendung im Standarddeutschen, die im Unterricht Deutsch als Fremdsprache für relevant erachtet werden. Teilweise werden auch eigene Korpora zusammengestellt. Dann geht es zum Beispiel um häufigen Wortschatz im Deutschen oder darum, an großen Mengen von Lernertexten typische Abweichungen von den Regularitäten des Deutschen zu identifizieren[3]. Unser Lehrwerkskorpus stellt insofern eine Novität im Bereich Deutsch als Fremdsprache dar.

Wissenschaftliches Arbeiten ist immer mit einem gewissen Maß an Abstraktion verbunden. In den folgenden Untersuchungen gibt es eine doppelte Verkürzung des Gegenstandes. Erstens werden Lernende nicht nur in Lehrbüchern mit Wortschatz konfrontiert. Weitere Quellen sind andere Unterrichtsmedien (Audio, Video, Internet), das Unterrichtsgespräch und die Beschäftigung/der Kontakt mit der Zielsprache außerhalb des Unterrichts. Zweitens hat gesteuerter Sprachunterricht natürlich nicht allein die Aufgabe, Wortschatz zu vermitteln.
Dennoch lassen sich aus unseren Analysen eindeutige Tendenzen hinsichtlich des Umgangs mit Wortschatzfragen ausmachen.

---

zu schaffen, hat sich bislang als Sackgasse erwiesen.“ Krumm/ Ohms-Duszenko (2001:1035).

[2] Die erste wegweisende Übersicht für DaF findet sich bei Fandrych/ Tschirner (2007); ausführlicher und aktueller Lüdeling/ Walter (2010). Eine umfassende Einführung in die Korpuslinguistik bietet Bubenhofer (2006-2010).

[3] Beispiele sind das frequenzbasierte Wörterbuch Deutsch als Fremdsprache von Tschirner/ Randall (2006) und das fehlerannotierte Lernerkorpus des Deutschen FALCO. (http://korpling.german.hu-berlin.de/falko/index.jsp)

2. Korpusaufbau

2.1 Datengrundlage, Datenaufbereitung und Datenauswertung

Für die Untersuchungen wurde ein Korpus der folgenden sieben japanischen Anfängerlehrbücher Deutsch als Fremdsprache erstellt:

Itayama, Mayumi/ Shioji, Ursula/Motokawa, Yuko/ Yoshimitsu, Takako: Farbkasten Deutsch neu 1. 26. Auflage, Tokyo: Sanshusha 2007. (Abgekürzt als *Farbkasten*)

Ogino, Kurahei/ Raab, Andrea: Ein Sommer in Deutschland. 4. Auflage, Tokyo: Asahi 2009. (Abgekürzt als *Ein Sommer*)

Riessland, Andreas/ Waragai, Ikumi/ Kimura, Goro Christoph/ Hirataka, Fumiya/ Raindl, Marko/ Ohta, Tatsuya: Modelle neu 1. 6. Auflage, Toyko: Sanshusha 2009. (Abgekürzt als *Modelle*)

Rikkyo Universität Institut für Deutsche Sprache (Hrsg.): Straße neu. 3. Auflage, Tokyo: Asahi 2008. (Abgekürzt als *Straße*)

Sato, Shuko/ Shimoda, Kyoko/ Papentin, Heike/ Oldehaver, Gesa: Szenen 1. 13. Auflage, Tokyo: Sanshusha 2009. (Abgekürzt als *Szenen*)

Seino, Tomoaki: Meine Deutschstunde. 4. Auflage, Tokyo: Asahi 2008. (Abgekürzt als *Deutschstunde*)

Sekiguchi, Ichiro: Hallo München. Neu. Tokyo: Hakusuisha 2008. (Abgekürzt als *Hallo München*)

Bei der Auswahl war entscheidend, dass es sich um allgemeine Deutschlehrwerke für Null-Anfänger handelt und nicht um spezifische Einführungen (z.B. in die deutsche Grammatik). Außerdem sollten verschiedene Verlage berücksichtigt werden. Darüber hinaus gab es kein Selektionskriterium und die Auswahl ist somit bis zu einem gewissen Grad zufällig. Im Rahmen der Untersuchungen wurde auch ein Korpus von in Deutschland erschienenen Lehrwerken erstellt. Für Vergleichszwecke greifen wir hier aber nur auf ein Lehrwerk zurück:

Aufderstraße, Hartmut/ Bock, Heiko/ Gerdes, Mechthild/ Müller, Jutta/ Müller, Helmut: Themen 1 neu. Kursbuch. Ismaning: Hueber 2003. (Abgekürzt als *Themen*)

Alle Lehrwerke wurden gescannt und dann mit einer Texterkennungssoftware digital lesbar gemacht. Metadaten (Titel, Kapitel, Seite, Text/Grammatik/Übung) wurden anschließend manuell annotiert. Mittels einer selbst erstellten Software wurden die relevanten Lehrwerksteile identifiziert und dem TreeTagger[4] für eine automatische Annotation mit Lemmata und Wortarten (parts-of-speech / POS[5]) übergeben.

Mit diesen Schritten werden Lehrwerke für vielfältige Analysen erschlossen. Unsere Untersuchungen beschränken sich zwar auf den verwendeten Wortschatz, doch wären aufgrund des POS-Taggings ebenso syntaktische Analysen durchführbar. Fragen des „Weltbildes“ in einem Lehrbuch könnte man beispielsweise über N-Gramm-Analysen[6] 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. Beispiele hierfür finden sich in Abschnitt 3 und 4.

Die folgende Tabelle gibt eine schematische Übersicht über die Schritte bei der Korpuserstellung und -analyse:

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	TreeTagger	XML-Datei	Nach Lemmata und POS

[4] Vgl. Schmid (1994).
[5] Zum dabei verwendeten Stuttgart-Tübingen-Tagset (STTS-Tagset) vgl. Schiller et al. (1995).
[6] Mit der Berechnung von N-Grammen (Wortkombinationen/Kollokationen/ Kookurrenzen) kann man typische oder auch untypische Sprachverwendung sichtbar machen. Näheres hierzu bei Bubenhofer (2006-2010) unter dem Kapitel „Eigenes Korpus“ – Daten analysieren.

	Tokenebene: Lemmatisierung, POS-Annotation			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	

2.2 Datenqualität und Aussagekraft

Bevor wir erste Ergebnisse vorstellen, noch ein Wort zur Datenqualität und zur Verlässlichkeit der Aussagen.

Die Lehrbuchtexte liegen uns leider nicht im Textformat vor. Deshalb mussten sie gescannt und automatisch in digitale Textdateien überführt werden. Da die OCR-Erkennung nicht fehlerfrei funktioniert, sind die Schritte 1 und 2 eine Quelle für Verunreinigungen in den Daten. Die neuesten OCR-Programme arbeiten bei reinen Textvorlagen sehr zuverlässig. Lehrbücher einhalten aber auch Grafiken und Bilddarstellungen, die Probleme bereiten. Bei Lückenübungen kommt es ebenfalls zu Fehlerkennungen.

Japanische Deutschlehrbücher enthalten sowohl deutschen als auch japanischen Text. Omnipage 16 professional erkennt entweder den japanischen oder den deutschen Text. Den japanischen Text haben wir für die vorliegende Studie vernachlässigt. Die OCR-Erkennung ergibt dann für die japanischen Wörter kryptische Zeichen, die vom Tagger als nicht-standardsprachlich identifiziert werden und in den Schritten 5 und 6 von uns ausgesondert wurden. Auf die sehr aufwändige manuelle Bereinigung der OCR-Erkennung wurde verzichtet. Eine Stichprobe ergab, dass die Fehlerquote im oberen einstelligen Bereich liegt.

In Schritt 3, der manuellen Annotation, wurden innerhalb der dokumentenspezifischen

Metadaten zunächst Autor und Titel annotiert. Zusätzlich haben wir Layout- und Textstrukturmerkmale annotiert: Kapitel, Seite und Textsorte (Lesetext, Darstellung der Grammatik, Übung, Wortschatzliste). Je tiefer und genauer die manuelle Annotation ist, desto feinmaschiger kann später der Zugriff auf den Wortschatz erfolgen. Doch die manuelle Annotation ist sehr arbeitsintensiv.

In Schritt 4 wird unter anderem der TreeTagger eingesetzt. Sowohl bei der automatischen Zuweisung von Wortkategorien als auch bei der anschließenden Lemmatisierung kommt es gelegentlich zu Fehlern. Bei diesem Schritt muss man mit einer Fehlerquote im mittleren einstelligen Bereich rechnen.

Die dargestellten Fehlerquellen führen dazu, dass unsere Analysen eine *Annäherung* an den Wortschatz der Lehrwerke bieten, jedoch keine exakte Abbildung. Im Folgenden wollen wir das Potenzial der so aufbereiteten Daten anhand eines Wortschatzprofils eines Lehrbuchs illustrieren (Kapitel 3) und im Anschluss die Ergebnisse eines Vergleichs der untersuchten Lehrbücher vorstellen (Kapitel 4).

3. Wortschatzprofil eines Lehrbuchs

3. 1 Allgemeine Gesichtspunkte

Zur Illustration der Analyseprinzipien und der Analysemöglichkeiten greifen wir auf den „Klassiker“ *Hallo München* zurück. Der Wortlistengenerator (Schritt 5) erstellt Frequenzlisten getrennt für die einzelnen Kapitel des Lehrbuchs oder auch den kumulierten Wortschatz. Bei letzterer Liste hat man also in der Lektion 1 nur den Wortschatz von Lektion 1, in Lektion 2 den Wortschatz von Lektion 1 und 2 und in der letzten Lektion den Wortschatz des gesamten Lehrbuchs. Die folgende Tabelle zeigt die oberen Ränge der nach Frequenzen geordneten kumulierten Listen für Lektion 1 und Lektion 12, die letzte Lektion von *Hallo München*:

Hallo München Ausschnitte aus den kumulierten Listen

Lektion 1		Lektion 12	
34	ich	181	sein
20	sein	177	ich
19	Sie|sie	128	ein
18	wohnen	122	Sie|sie

17	in	114	in
11	gut	106	haben
8	und	69	er
8	kommen	64	du
8	gern	63	und
7	Tag	58	Szene
6	wir	57	wir
6	Szene	54	fahren
5	morgen	52	dies
5	aus	51	kommen
4	woher	46	werden

In der linken Spalte sieht man die absolute Frequenz und in der rechten Spalte das Lemma. Die Sortierung erfolgt standardmäßig nach der Frequenz. In einer Tabellenkalkulations-Software wie Excel ist aber eine alphabetische Sortierung ohne Probleme zu bewerkstelligen.

Einige weitere Erläuterungen: Das häufigste Lemma im Deutschen, der bestimmte Artikel mit seinen Formen, wird aus technischen Gründen in keine der Listen aufgenommen. Außerdem gibt es in unseren Wortschatzlisten keine Eigennamen. Der Name *Klaudia* ist im Lehrwerk *Hallo München* hochfrequent ebenso wie der Stadtname *Berlin*. Für Wortschatzvergleiche sind Eigennamen störend. Folglich haben wir sie herausgefiltert. Eine solche Filterung ist möglich, weil der Lehrbuchtext getaggt wurde und im Tagset Eigennamen mit dem Tag NE versehen sind. In einem weiteren Sinne erschließen die Tags den Wortschatz nach grammatischen Kategorien und die Wortschatzlisten lassen sich auch so ausgeben. Hierzu ein Beispiel aus der Lektion 1:

Hallo München Ausschnitt POS Liste

Lektion 1

34	PPER_ich	4	ADV_jetzt
19	VAFIN_sein	3	VVFIN_trinken
19	PPER_Sie|sie	3	VVFIN_studieren
18	VVFIN_wohnen	3	VVFIN_spielen
17	APPR_in	3	VVFIN_hören
9	ADJA_gut	3	VVFIN_heißen
8	VVFIN_kommen	3	PTKNEG_nicht
8	KON_und	3	PTKANT_ja

8	ADV_gern	3	PTKANT_danke
7	NN_Tag	3	PPER_sie
6	PPER_wir	3	PPER_es
6	NN_Szene	3	NN_Wiedersehen
5	ADV_morgen	3	NN_Tennis
4	PWAV_woher	3	NN_Herr
4	APPR_aus	3	NN_Bier

Das Lehrbuch *Hallo München* umfasst in unserer Auswertung 4392 Wortformen, die 741 Lemmata zugeordnet sind. Diese absoluten Zahlen haben für sich genommen schon eine Aussagekraft hinsichtlich des Umfangs des im Lehrwerk vorkommenden Wortschatzes. Differenziertere Aussagen sind aber möglich, wenn man das Auftreten eines Wortes mit dem Auftreten anderer Wörter in Beziehung setzt. Wir wollen hier drei Messwerte, die unserer Meinung nach allgemeine Wortschatzeigenschaften eines Lehrwerkes gut erfassen, einführen:

a) **Wiederholungsrate**: diese Zahl ergibt sich, wenn man die Zahl der Wortformen durch die Zahl der Lemmata teilt. Im Prinzip ist das nichts anderes als die type-token-ratio, nur eben bezogen auf Lemmata. Für *Hallo München* erhalten wir 5,9. Mit anderen Worten kommt jedes Lemma durchschnittlich etwa 6 Mal vor.
b) **Aktivrate**: diese Prozentzahl misst den Anteil der Lemmata am gesamten Wortschatz, die die Frequenz 5 und höher haben. Hier ergibt sich für *Hallo München* ein Wert von 24,3 Prozent.
c) **Solitärrate**: diese Zahl ist der prozentuale Anteil an Lemmata, die nur ein Mal im ganzen Lehrwerk vorkommen. *Hallo München* hat einen Wert von 37,0 Prozent.

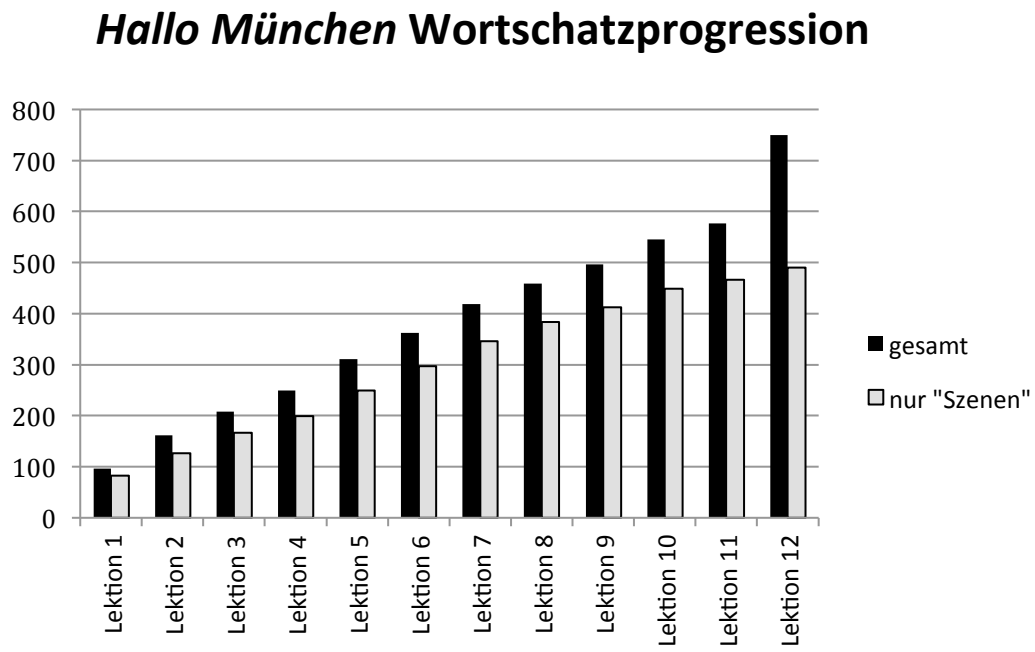
Wenn man die Aktivrate und die Solitärrate kennt, weiß man auch, wie viel Prozent der Lemmata mit einer Frequenz von 2-4 im Lehrwerk vorkommen.
Mit der „Aktivrate“ wollen wir den Bereich definieren, in dem ein Lemma durch sein häufigeres Auftreten prominent wird. Es gibt viele Gesichtspunkte, die bei der Aneignung von Wortschatz eine Rolle spielen[7]. Ein Faktor ist dabei auch die Kontakthäufigkeit. Je häufiger der Kontakt ist, desto deutlicher wird es für den Lerner,

[7] Wir können hier nicht auf diese wichtigen Fragen eingehen. Eine Einführung gibt Nation (2001). Eine aktuelle Anleitung zu eigener Forschung bietet Schmitt (2010).

dass es sich um ein relevantes Wort handelt[8]. Auf der anderen Seite bringt es der inkrementelle Charakter des Wortschatzlernens mit sich, dass ein Lerner mehrere Kontakte mit dem Wort braucht, um es sich anzueignen[9].
Eine hohe Solitärrate ist für Lerner im Anfängerbereich unter anderem deshalb ein Problem, weil eine Fokussierung beim Wortschatzlernen erschwert wird.
Die vorgestellten Messwerte werden im Abschnitt 4 beim Lehrwerkvergleich wieder aufgegriffen.

3.2 Spezielle Gesichtspunkte eines Lehrwerks

Über die Ausgabe des Wortschatzes nach Kapiteln kann man auch die Wortschatzprogression darstellen. Da in der Annotation auch zwischen Textsorten unterschieden wurden, ist es auch möglich, die Progression der Kerntexte, der „Szenen“, allein zu betrachten.



Die Wortschatzprogression ist kontinuierlich. Es gibt keine Plateau-Lektionen, in denen nur wiederholt und kein neuer Wortschatz vermittelt wird.
Bei der Gesamtprogression gibt es in der ersten Lektion fast 100 neue Lemmata. Dann

[8] Vgl. Nation (2001: 396).
[9] Vgl. Schmitt (2010: 20).

schwankt der Wortschatzzuwachs zwischen 65 und 31 neuen Lemmata pro Kapitel. Die Lektion 12 fällt mit einem Zuwachs von über 150 Lemmata aus dem Rahmen, weil die „Kaffeepausen", optionale Lektionsteile, angehängt sind.
Die Wortschatzprogression der „Szenen" allein verläuft flacher. Das bedeutet aber andererseits, dass etwa 50 Prozent zusätzlicher Wortschatz in die Grammatikteile, die Übungen und die Kaffeepausen hineingepackt ist.

Über das POS-Tagging erschließen sich die grammatischen Kategorien des Wortschatzes und so kann man die Darstellung von einzelnen Kategorien im Lehrbuch genauer greifen. Als kleines Beispiel seien zunächst die Adjektive in *Hallo München* angeführt. Die attributive (ADJA) und prädikative (ADJD) Verwendung von Adjektiven ist insgesamt sehr ausgeglichen. Hier die Rangliste der fünf am häufigsten verwendeten Adjektive:

23	ADJA	gut	22	ADJD	gut
10	ADJA	rot	11	ADJD	schön
6	ADJA	schön	7	ADJD	wirklich
5	ADJA	neu	5	ADJD	groß
5	ADJA	klein	4	ADJD	müde

Es gibt eine ganze Reihe von polaren Adjektiven: alt-neu/modern, teuer-billig, fit-müde, groß-klein, künstlich-natürlich, schwarz-weiß. Allerdings wird auf eine mnemotechnisch hilfreiche Zusammenstellung der polaren Adjektive verzichtet. Die beiden am häufigsten im Buch vorkommenden Adjektive *gut* und *schön* haben im ganzen Buch kein polares Pendant.

In Lehrbüchern gibt es kleinere oder größere Wortschatzlisten, die den Lernern wichtigen Wortschatz nahelegen. So auch bei *Hallo München*, wo man im Anhang eine Liste mit 41 „besonders wichtigen unregelmäßigen Verben" und die dazugehörigen Stammformen findet. Im Lehrwerk wurde also eine Auswahl aus den ca. 200 deutschen unregelmäßigen Verben[10] vorgenommen. Wir kennen die Auswahlkriterien nicht und

[10] Die Zählungen schwanken, je nach dem, wie stark die Darstellung historisch orientiert ist und ob die „gemischten Verben" wie *denken* oder *bringen* in eine eigene Gruppe kommen oder nicht.

können nur in das Lehrbuch sehen. Der größere Teil der Verben ist auch im Lehrbuch mittel- oder hochfrequent. Aber ein Teil findet sich außer in der Liste nur noch ein oder zwei Mal im ganzen Lehrbuch (*helfen*, *tragen, singen*, *waschen*, *werfen*). Schließlich gibt es eine Reihe von Verben, die ausschließlich in der wichtigen Liste vorkommen (*halten*, *liegen*, *nennen*, *steigen*, *sterben*, *treffen*, *wachsen*, *ziehen*).

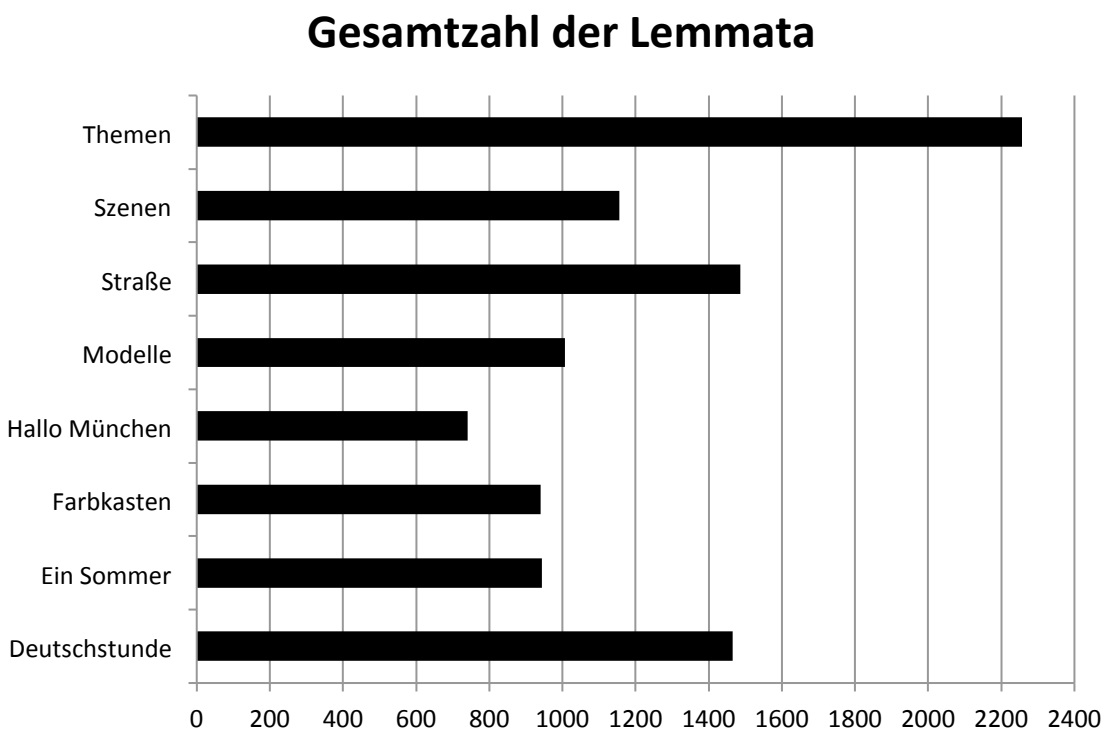
Wir haben in diesem Kapitel das Potenzial einer lehrwerkimmanenten Analyse anhand einiger zentraler Kategorien illustriert. Eine weitere Differenzierung wäre selbstverständlich möglich und wünschenswert. Im folgenden Kapitel richtet sich der Fokus auf den Vergleich der Basisprofile der untersuchten Lehrbücher.

4. Lehrbuchvergleiche

Auch wenn die japanischen Lehrwerke normalerweise das gleiche Zielpublikum haben, nämlich Studienanfänger, die Deutsch als 2. Fremdsprache gewählt haben, so gibt es doch unterschiedliche Vorstellungen darüber, was auf welche Weise vermittelt werden soll. Dies gilt auch für Umfang und Art des Wortschatzes. Wir halten diese Diskussion für zentral, wollen jedoch hier statt der Ausarbeitung eines weiteren normativen Vorschlags die in japanischen Lehrbüchern enthaltenen impliziten Normen anhand unserer Analyse rekonstruieren.

4. 1 Wortschatzumfang und allgemeine Frequenzen

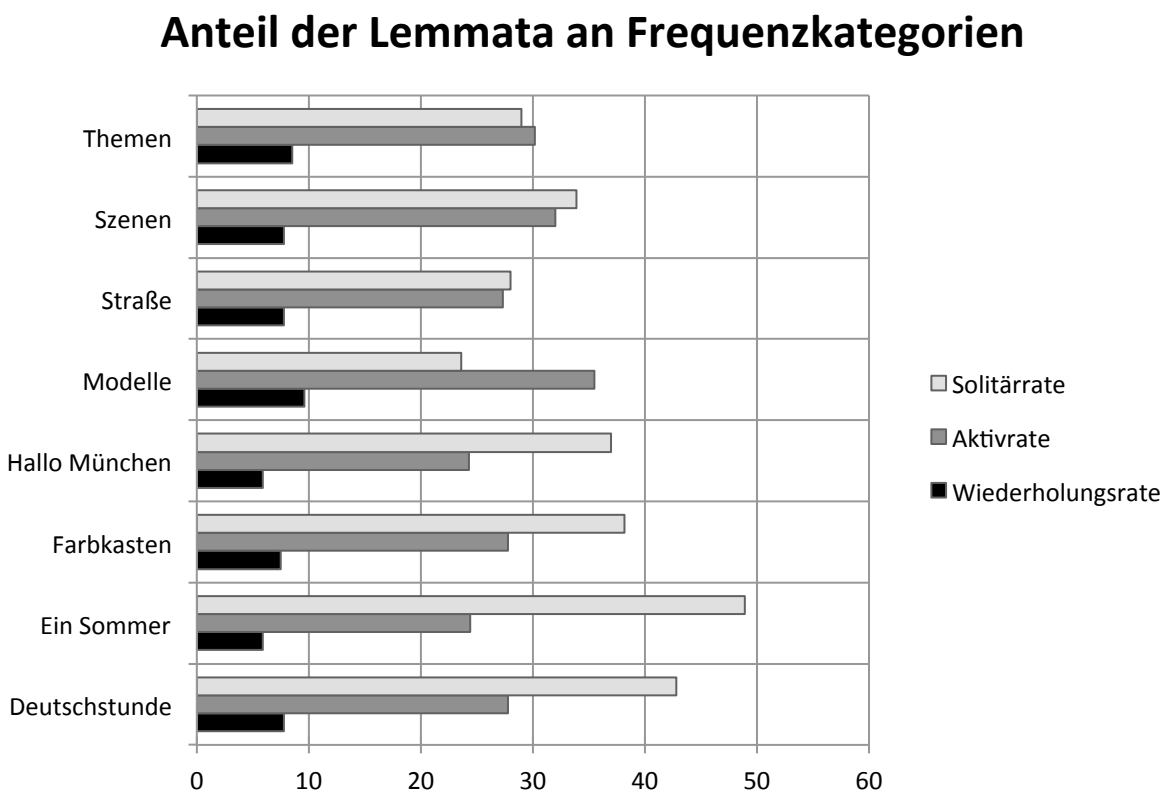
Zunächst vergleichen wir die Gesamtzahl der Lemmata, die in den ausgewählten japanischen Lehrwerken auftreten. Um einen weiteren Anhaltspunkt zu geben, haben wir auch *Themen* aufgenommen.



In den untersuchten japanischen Lehrwerken werden die Lernenden mit durchschnittlich etwa 1100 Lemmata konfrontiert. Dabei liegt das Maximum bei fast 1500 (*Straße* und *Deutschstunde*) und das Minimum bei etwa bei 750 (*Hallo München*).
In *Themen* finden sich doppelt so viele Lemmata wie im durchschnittlichen japanischen Lehrwerk.

Im Abschnitt 3 haben wir drei Messwerte vorgestellt, die unserer Meinung nach hilfreich bei der Profilierung des Wortschatzes sind: Wiederholungsrate, Aktivrate und Solitärrate.

Preprint von: Bubenhofer, Noah / Willi Lange / Saburo Okamura / Joachim Scharloth (2011): Welcher Wortschatz? Korpuslinguistische Untersuchungen zur Wortschatzselektion japanischer Deutschlehrbücher für Anfänger. In: Doitsugo Kyoiku - Deutschunterricht in Japan, 16, S. 43-60.



Die Wiederholungsrate bewegt sich bei den japanischen Lehrwerken zwischen 5,9 (*Hallo München* und *Deutschstunde*) und 9,6 (*Modelle*). Der Mittelwert liegt bei 7,5. Das sind deutliche Unterschiede, die noch klarer werden, wenn man die anderen beiden Messwerte betrachtet. Die Aktivrate liegt bei den japanischen Lehrwerken bei 28,4. Das bedeutet, dass 28,4% der Lemmata 5 Mal und öfter vorkommen. Bei *Modelle* sind es 35,5% und bei *Ein Sommer* 24,4%. bzw. *Hallo München* 24,3%. Die durchschnittliche Solitärrate beträgt 36,1 – bei sehr starken Abweichungen nach oben und unten. *Deutschstunde* ist ebenso wie *Ein Sommer* durch eine sehr hohe Anzahl an Lemmata gekennzeichnet, die nur 1 Mal im Lehrwerk vorkommen (42,8% bzw. 48,9%). Auf der anderen Seite steht *Modelle* mit auffallend niedrigen 23,6%.

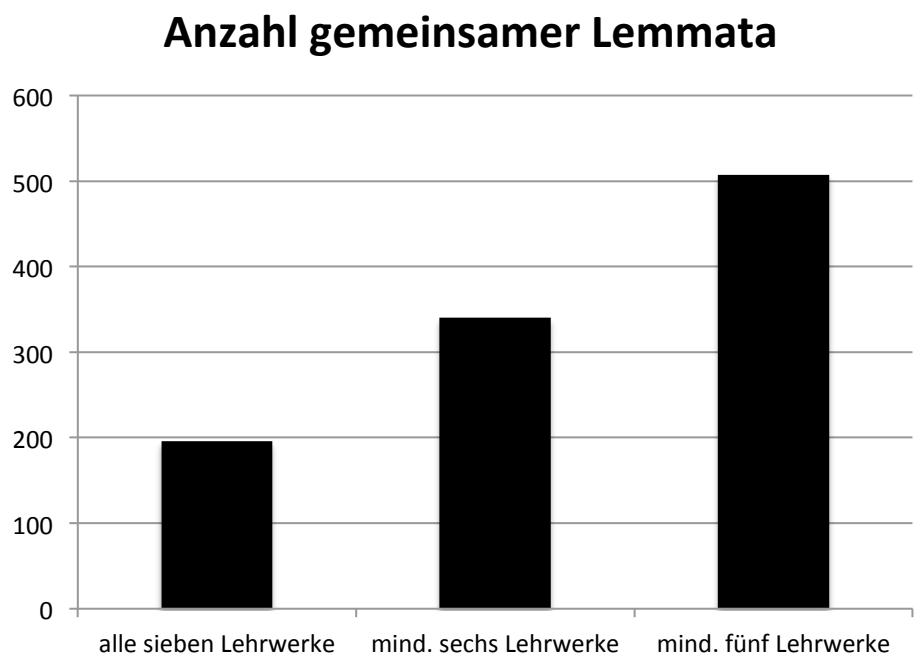
4. 2 Wortschatzüberschneidungen

Welcher Wortschatz ist den japanischen Lehrwerken gemeinsam? Wo liegen deutliche Unterschiede? Um diese Fragen zu beantworten, haben wir zunächst eine Datenmatrix erstellt. Das Vorkommen eines Lemmas in einem Lehrbuch wird mit dem Wert 1 bezeichnet, das Fehlen mit einer 0. Hier ein ganz kleiner Ausschnitt:

	Farb-kasten	Hallo München	Deutsch-stunde	Modelle	Ein Sommer	Straße	Szenen	**Summe**
Essen	1	1	1	1	1	1	1	**7**
gleich	1	1	1	1	1	1	1	**7**
Viertel	1	1	1	1	1	1	1	**7**
Buch	1	1	1	1	1	1	1	**7**
sprechen	1	1	1	1	1	1	1	**7**
Jahr	1	1	1	1	1	1	1	**7**
studieren	1	1	1	1	1	1	1	**7**
sie	1	1	1	1	1	1	1	**7**
fahren	1	1	1	1	1	1	1	**7**
Fußball	1	1	1	1	1	1	1	**7**
denken	0	1	1	1	1	1	1	**6**
Tisch	1	0	1	1	1	1	1	**6**
Party	1	1	1	1	0	1	1	**6**
frisch	1	1	1	1	1	1	0	**6**

In der ersten Spalte sind sämtliche im Gesamtkorpus vorkommenden Lemmata erfasst. In den weiteren Spalten ist das Vorkommen/Nichtvorkommen in den einzelnen Lehrwerken markiert. Die Frequenzen der Lemmata innerhalb der Lehrwerke haben wir bei unserem Vergleich nicht berücksichtigt. In der Spalte ganz rechts findet sich die Summe der Lehrwerke, in denen das jeweilige Lemma vorkommt. Der gewählte Ausschnitt aus der Datenmatrix zeigt einen Wortschatzbereich, in dem es zwischen den japanischen Lehrwerken sehr starke Überschneidungen gibt. Die Lemmata werden in allen japanischen Lehrwerken verwendet (Summe = 7) oder sind in sechs der sieben Lehrwerken vertreten.

Die folgende Grafik gibt eine Übersicht über die Anzahl der Lemmata, die in mehr als vier Lehrbüchern vorkommen. Es zeigt sich, dass die Zahl der Lemmata, die in allen Lehrwerken vorkommen, mit 196 relativ gering ist. 340 Wörter kommen in sechs Lehrbüchern vor, 507 in fünf. Allerdings muss hier aufgrund der in 2.2 dargestellten Probleme bei der Korpuserstellung und -annotierung von einem gewissen Fehleranteil ausgegangen werden.



4. 3 Gruppierung der Lehrwerke auf der Basis des vorkommenden Wortschatzes

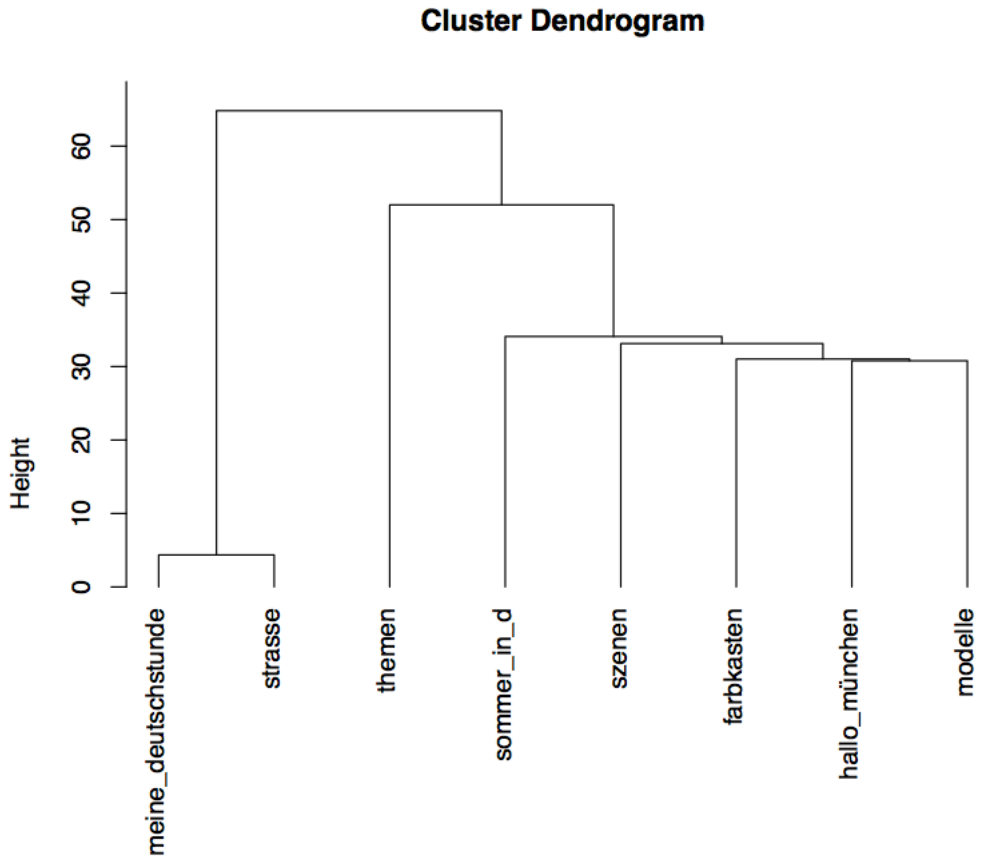
In einem letzten Analyseschritt sind wir der Frage nachgegangen, welche Lehrwerke im Hinblick auf die Distribution der Lemmata eine große Ähnlichkeit aufweisen. Auf der Basis der im vorangegangenen Abschnitt vorgestellten Datenmatrix haben wir eine Clusteranalyse durchgeführt. Die Clusteranalyse gehört zu den strukturentdeckenden Verfahren der multivariaten Statistik. Sie dient dazu, die Objekte eines unklassifizierten Merkmalsdatensatzes in Gruppen einzuteilen, die in sich möglichst homogen sind und sich von den anderen Gruppen möglichst stark unterscheiden[11].

Wir haben auf der Basis der Datenmatrix eine hierarchische Clusteranalyse (Ward) mit dem Statistikprogramm R[12] durchgeführt. Die Ergebnisse von Clusteranalysen können auf unterschiedliche Weise visualisiert werden. Die am häufigsten verwendete Methode ist das Dendrogramm. Dendrogramme sind Baumgraphen, die die Zerlegung einer Datenmenge in immer kleinere Teilmengen darstellen. Die äußersten Enden des Graphen ("Blätter") bestehen aus Clustern, die je ein einzelnes Objekt der Datenmenge umfassen; die "Wurzel" aus einem Cluster, das sämtliche Objekte der Datenmenge

[11] Vgl. Runkler (2010: 105).
[12] Vgl. Ihaka/ Gentleman (1996).

umfasst. Ein Knoten eines Dendrogramms repräsentiert die Vereinigung aller seiner Kindknoten. Die Länge der Kanten zwischen zwei Knoten bildet die Distanz zwischen den beiden Mengen ab[13].

Das folgende Dendrogramm zeigt im Hinblick auf den verwendeten Wortschatz große Unterschiede, aber auch große Ähnlichkeiten zwischen den japanischen Lehrbüchern. Während *Deutschstunde* und *Straße* ein Cluster bilden, das sich maximal von den anderen Lehrwerken unterscheidet, sind sich die restlichen japanischen Lehrwerke aus Sicht des in ihnen verwendeten Wortschatzes sehr ähnlich.



5. Desiderata

Wir haben uns in dieser Untersuchung darauf beschränkt, aufzuzeigen, wie der Wortschatz von japanischen Deutschlehrbüchern für Anfänger analysiert und die

[13] Zu den Layout-Prinzipien von hierarchischen Graphen vgl. Brockenauer/ Cornelsen (2001: 212).

Lehrbücher untereinander verglichen werden können. Dabei zeigten sich große Unterschiede im Hinblick auf die Wortschatzmenge, das gewählte Vokabular und die Verwendungshäufigkeit des einmal eingeführten Vokabulars.

Auf der Ebene der Lehrwerkproduktion lassen sich daraus unterschiedliche Schlussfolgerungen ziehen. Die Lehrwerke, die wir untersucht haben, bieten wenig oder keine Aussagen darüber, welchen Wortschatz sie vermitteln wollen und was über das Wortschatzlernen angenommen wird. Das ist in vielerlei Hinsicht sowohl für die Lernenden als auch auf die Lehrenden ein Problem. Aus unseren Ergebnissen lässt sich jedoch ableiten, dass den beiden Lehrbuchclustern unterschiedliche Annahmen über Umfang und Qualität des Lernerwortschatzes zugrunde liegen. Ein deutlich höheres Maß an Transparenz scheint uns hier wünschenswert.
Auf der Forschungsebene wäre es wichtig, den Wortschatz der Lehrwerke mit Referenzwortschätzen zu vergleichen[14]. Haben zum Beispiel Sammlungen wie Profile (Glaboniat/ u.a.(2005))oder Randall/ Tschirner (2006) eine normative Wirkung auf die japanischen Lehrbuchautorinnen und –autoren? Nicht zuletzt sollte überprüft werden, ob die Referenzwortschätze selbst, die oft einen autoritativen Anspruch haben oder zugewiesen bekommen, methodisch transparent und gut abgesichert sind.

Bibliographie

Brockenauer, Ralf/ Cornelsen, Sabine (2001): Drawing Clusters and Hierarchies. In: Michael Kaufmann/ Dorothea Wagner (Hrsg.): Drawing Graphs. Methods and Models. Berlin/ Heidelberg/ New York et al.: Springer, S. 193-227.
Bubenhofer, Noah (2006-2010): Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. (http://www.bubenhofer.com/korpuslinguistik/)
Fandrych, Christian/ Tschirner, Erwin (2007): Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. In: Deutsch als Fremdsprache, Heft 4, S. 195–204.

[14] In diese Richtung gehen auch Lymperakakis / Sapiridou (2010).

Glaboniat, Manuela/ Müller, Martin/ Rusch, Paul/ Schmitz, Helen/ Wertenschlag, Lukas (2005): Profile deutsch. Niveaustufen A1-C2. Version 2.0. Berlin/ München/ Wien/ Zürich/ New York: Langenscheidt

Krumm, Hans-Jürgen/ Ohms-Duszenko, Maren (2001): Lehrwerkproduktion, Lehrwerkanalyse, Lehrwerkkritik. In: Helbig, Gerhard/ Götze, Lutz/ Henrici, Gerd/ Krumm, Hans-Jürgen: Handbuch Deutsch als Fremd- und Zweitsprache. HSK 19,2 . Berlin/ New York: de Gruyter, S.1029–1041.

Lüdeling, Anke/ Walter, Maik (2010): Korpuslinguistik. In: Krumm, Hans-Jürgen/Fandrych, Christian/ Hufeisen, Britta/ Riemer, Claudia (Hrsg.): Handbuch Deutsch als Fremd- und Zweitsprache. (Neubearbeitung). HSK 35. Berlin: Mouton de Gruyter, S. 315-322.

Lymperakakis, Panagiotis/ Sapiridou, Andromachi (2010): Korpusbasierte Worthäufigkeitslisten und Wortschatz – eine quantitative und qualitative Analyse am Beispiel des Fremdsprachenlehrwerkes „Deutsch – ein Hit! 1“. In: Info DaF, Heft 4, S. 368-382.

Nation, Paul I. S. (2001): Learning Vocabulary in Another Language. Cambridge: University Press.

Ihaka, Ross/ Gentleman, Robert (1996): R: A Language for Data Analysis and Graphics. In: Journal of Computational and Graphical Statistics, 5,3, S. 299-314. (R statisticspackage: www.r-project.org/)

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

Runkler, Thomas A. (2010): Data Mining. Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden: GWV Fachverlage.

Schiller, Anne/ Teufel, Simone/ Thielen, Christine (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart: [Working Paper] Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, Tübingen: Seminar für Sprachwissenschaft.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. Stuttgart: [Working paper] Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Schmitt, Norbert (2010): Researching Vocabulary: A Vocabulary Research Manual. London: Palgrave Press.