GERMANICA WRATISLAVIENSIA 129
Acta Universitatis Wratislaviensis No 3163
Wrocław 2009

Erwin Tschirner
Leipzig

Korpuslinguistik und Fremdsprachenunterricht

In der Wissenschaft gibt es zwei grundsätzlich unterschiedliche Ansätze, um Erkenntnisse zu gewinnen, den Rationalismus und den Empirismus (Lemnitzer/Zinsmeister 2006). Rationalisten gehen davon aus, dass die Vernunft alleine, ohne auf die Sinne angewiesen zu sein, Erkenntnis von der Welt gewinnen kann. Erkenntnis wird mittels deduktiver Verfahren aus ersten Prinzipien abgeleitet, wobei die Wahrheit dieser Prinzipien nicht bezweifelt wird (Specht 1979). Eine grundsätzlich rationalistische Wissenschaft ist die Mathematik, in der Sprachwissenschaft ist es z. B. die generative Grammatik. Empiristen postulieren, dass Erkenntnis von der Wirklichkeit nur durch Erfahrung möglich ist. Erkenntnis basiert auf dem sinnlich Wahrgenommenen, auf unmittelbar gegebenen und dadurch unbezweifelbaren Elementen (Gawlick 1980). Grundsätzlich empirische Wissenschaften sind die Naturwissenschaften, in der Sprachwissenschaft gehört dazu die Korpuslinguistik.

Vergleicht man die generative Grammatik mit der Korpuslinguistik, treten deutliche Unterschiede zu Tage. Die generative Grammatik postuliert als oberstes Prinzip für Spracherwerb und Sprachbenutzung eine angeborene Wissensgrundlage in Form eines kognitiven Moduls – die *Language Faculty* oder das Sprachmodul. Ziel der Forschung ist es, die linguistische Kompetenz des Menschen zu beschreiben und zu erklären. Grammatische Prinzipien werden durch Introspektion abgeleitet. Als Grundlage dienen neben der rationalistischen Deduktion Grammatikalitätsurteile auf der Basis meist selbstkonstruierter, kontextloser Sätze. Die Korpuslinguistik befasst sich mit beobachtbaren Daten authentischer Sprachverwendung. Durch die meist rechnergesteuerte Analyse sehr großer Mengen von authentischen Texten werden typische Muster erfasst – grammatischer, semantischer und lexikalischer Art, insbesondere usuelle Wortverbindungen. Ziel ist eine performanzorientierte Sprachbeschreibung.

Die generative Grammatik geht von einer strengen Trennung von Grammatik und Lexik aus und stellt die Kreativität der Sprache in den Mittelpunkt, die freie Generierung von Sätzen und damit das von Sinclair (1991) so genannte *Open-*

Choice Prinzip. Die Korpuslinguistik geht von einem fließenden Übergang zwischen Grammatik und Lexik aus und postuliert: „Syntax is driven by lexis". „Syntactic structures and lexical items are co-selected. Particular syntactic structures tend to co-occur with particular lexical items and lexical items seem to occur in a limited range of structures (Francis 1993). Im Mittelpunkt steht die Usualität und damit das *Idiom* Prinzip (Sinclair 1991), die Formelhaftigkeit der Sprache. Während sich die generative Grammatik vordringlich mit der Sprachfähigkeit des Menschen befasst, mit abstrakten grammatischen Prinzipien, die für alle natürlichen Sprachen gelten, befasst sich die Korpuslinguistik mit Einzelsprachen bzw. dem Vergleich zweier oder auch mehrerer Einzelsprachen.

Ich werde im Folgenden argumentieren, dass korpuslinguistische Ansätze für die Beschreibung ebenso wie für die Didaktik und Methodik des Deutschen als Fremdsprache ein größeres Potenzial haben als rationalistische Ansätze wie z. B. der generative Ansatz. Ich werde versuchen, dies an drei übergreifenden Themen festzumachen: die Formelhaftigkeit der Sprache, Häufigkeitseffekte und die Frage nach der Autonomie der Syntax. Zuvor möchte ich allerdings aus einer fremdsprachendidaktischen Perspektive die Vorteile einer korpusbasierten Beschreibung des Deutschen kurz zusammenfassen (s. dazu auch: Fandrych/Tschirner 2007).

Korpusbasierte Ansätze aus der Sicht der Fremdsprachendidaktik

Die Korpuslinguistik befasst sich mit der Erforschung möglichst großer, repräsentativer Mengen von Sprachdaten. Ziel ist die Beschreibung einer Einzelsprache. Der Fokus liegt auf Performanzdaten, wobei quantitative Informationen eine zentrale Rolle spielen. Diese Performanzperspektive ist die große Stärke korpuslinguistischer Untersuchungen. Anstelle selbstkonstruierter isolierter Sätze verwendet sie echte Beispiele in ihren jeweiligen Kontexten und fördert damit die Authentizität und die Natürlichkeit der Beispiele.

Die Performanzperspektive bewirkt weiterhin eine Textfokussierung, eine Fokussierung auf die Bedeutung und Verwendung von Wörtern im Kontext. Dies ist wichtig, da Wörter, besonders häufige Wörter, komplexe Bedeutungsstrukturen haben, die sich ohne Kontext nicht oder kaum erfassen lassen. Des Weiteren führen korpusbasierte Untersuchungen oft zu anderen Darstellungen grammatischer Phänomene als sie in Grammatikbüchern angeboten werden, vor allem was ihre Verwendung, Bedeutung und Gewichtung angeht. Schließlich lassen sich über Korpusuntersuchungen Register- und Textsortenunterschiede herausarbeiten, die für die Natürlichkeit und Idiomatizität fremdsprachlicher Produktionen wichtig sind.

Ein wichtiges Kriterium der Korpuslinguistik ist die erschöpfende Untersuchung der vorhandenen Datengrundlage. Dabei werden Aussagen über die Häufig-

keit lexikalischer Einheiten und grammatischer Strukturen und ihre Kombinationsmöglichkeiten getroffen, die von einem Muttersprachler, rein auf der Basis seiner Intuition, so nicht getroffen werden können.

> The human being, contrary to popular belief, is not well organized for isolating consciously what is central and typical in the language; anything unusual is sharply perceived, but the humdrum everyday events are appreciated subliminally (Sinclair/Renouf 1988).

Performanzperspektive, Häufigkeit, das Zentrale und Typische, das Authentische und Natürliche, Textsorten- und Registerunterschiede: Das sind wichtige Stichwörter sowohl für die Korpuslinguistik wie auch für die Fremdsprachendidaktik und die Darstellung einer Sprache aus der Fremdperspektive. Im Folgenden möchte ich nun auf eine der zentralen Beobachtung eingehen, die sich in den letzten 20 Jahren aus der Erforschung von Lernersprachen und der Beschäftigung mit der Lernerperspektive ergeben haben, nämlich die Formelhaftigkeit der Sprache.

Formelhaftigkeit der Sprache

Bereits Pawley und Syder (1983) postulierten, dass Muttersprachler beim Sprechen nicht jede Äußerung komplett neu generieren, sondern auf eine Vielzahl von Versatzstücken zurückgreifen, auf Hunderttausende von lexikalisierten Phrasen, Sätzen und Teilsätzen, die sie im Laufe von vielen Jahren als Ganzes gespeichert haben und als Ganzes abrufen können. Dies ermöglicht es ihnen, fließend zu sprechen, und dies ist die Erklärung dafür, dass Muttersprachler idiomatisch sprechen sowie erkennen, wenn dies jemand tut bzw. nicht tut.

Formelhafte Sprache ist ein Oberbegriff für Phänomene, die von Kookkurrenzen über Kollokationen und Phraseologismen bis hin zu Sprichwörtern und Redensarten reichen. Eine formelhafte Sequenz ist nach Wray/Perkins (2000: 1)

> a sequence, continuous or discontinuous of words or meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

Formelhaftigkeit ist demnach daran ersichtlich, dass keine Regeln angewendet werden. Die – im Normalfall – mündliche Produktion erfolgt schnell und phonologisch kohärent, d. h. ohne Pausen und ohne Zögern, was darauf hindeutet, dass die formelhafte Wortverbindung als ganze Einheit aus dem mentalen Lexikon abgerufen wird.

Weitere Merkmale formelhafter Sequenzen sind ihre Invarianz, ihr Grad formaler Fixiertheit, und ein deutlich häufigeres Auftreten.

> Usuelle Wortverbindungen sind über das Einzelwort hinausgehende sprachliche Erscheinungen, die als komplexere Einheiten reproduziert werden können und deren Elemente einen höheren Wahrscheinlichkeitsgrad des Miteinandervorkommens besitzen, als das bei okkasionellen Wortverbindungen der Fall ist (Steyer 2000).

Hinzu kommt, vor allem bei idiomatischen Redewendungen, ein gewisser Grad an semantischer Undurchsichtigkeit.

Die Verwendung formelhafter Sprache hat vor allem zwei Funktionen: eine psycholinguistische und eine soziolinguistische. Zum einen erleichtert Formelhaftigkeit das fließende Sprechen und die Möglichkeit der Konzentration auf Inhalte und soziale Aspekte und nicht auf die grammatische Regelsuche und Regelanwendung. Für den muttersprachlichen Hörer hat dies den Vorteil, dass sich die Vorhersagbarkeit des Gehörten erhöht. Zum anderen werden über formelhafte Wendungen soziale Informationen wie Gruppenzugehörigkeiten weitergegeben und wahrgenommen.

Auch in Lernersprachen treten formelhafte Sequenzen auf, als Routinierung häufiger Verbindungen ohne Reflexion der inneren Struktur. Im Gegensatz zum muttersprachlichen Erwerb, in dem zuerst vor allem formelhaft, dann vor allem analytisch, dann wiederum holistisch gelernt wird, bis im Alter von ca. 8 Jahren ein Gleichgewichtszustand zwischen analytischer und holistischer Sprachverwendung erreicht wird, wird im Fremdsprachenerwerb formelhafte Sprache eher selten als Lernstrategie genutzt. „There is very little evidence that adult learners naturally extrapolate grammatical or lexical information from larger strings“ (Wray 2000).

Die fehlende Formelhaftigkeit wird von Yorio (1989) als nicht-phonologischer Akzent bezeichnet. Er wirkt sich u. a. dadurch aus, dass Lerner seltener als Muttersprachler von typischen Wortverbindungen Gebrauch machen, dass, wenn sie auftreten, es sich um andere handelt als die, die Muttersprachler verwenden würden, dass sie nicht mit derselben Häufigkeit wie bei Muttersprachlern verwendet werden, dass sie in anderen syntaktischen Strukturen auftreten als bei Muttersprachlern und dass sie unterschiedliche pragmatische Funktionen haben.

Altenberg/Granger (2001) verglichen zum Gebrauch des englischen Verbs *to make* Daten aus Lernerkorpora mit denen von Muttersprachlern und stellten fest, dass in Lernertexten Phraseologismen wie *make a decision*, *make a claim*, *make an effort* viel seltener verwendet wurden als in muttersprachlichen Texten. Aufgrund zahlreicher Studien, die den unzureichenden Gebrauch formelhafter Sprache bei Fremdsprachenlernern feststellen, fordern Altenberg/Granger (2001) und andere, dass Lerner gezielt auf formelhafte Verwendungsweisen, insbesondere häufiger Wörter, hingewiesen werden müssen.

Als Beispiel dafür, wieviel Muttersprachler über die Verwendungsweise ihrer Wörter wissen – und wieviel Nichtmuttersprachler nicht wissen, wenn man sie nicht darauf hingewiesen hat – beziehe ich mich auf eine Studie, die sich mit den als Synonyma bezeichneten Adjektiven *ewig* und *unendlich* befasst (Meißner 2006; 2008).

Die zwei wichtigsten Forschungsfragen waren: 1. Werden die Hauptbedeutungsvarianten der Adjektive *ewig* und *unendlich* in der Reihenfolge ihrer Häufig-

keit von den wichtigsten DaF-Großwörterbüchern und anderen wichtigen Wörterbüchern erfasst? 2. Welche kontextsensitiven Bedeutungsvarianten weisen diese Adjektive auf und wie hoch ist der Anteil solcher Varianten? Dazu wurden alle Belegfälle im Herder/BYU-Korpus und im DWDS-Korpus ermittelt und ausgewertet. Das Herder/BYU-Korpus (Tschirner/Jones 2005) entstand aus einer Zusammenarbeit zwischen dem Herder-Institut der Universität Leipzig und der Brigham Young University in Provo, Utah. Es ist eine repräsentative Textsammlung der gesprochenen und geschriebenen Sprache der deutschsprachigen Länder seit 1990 und umfasst ca. 4,2 Mio. laufende Wörter. Das DWDS-Korpus ist das Kernkorpus des Digitalen Wörterbuchs der Deutschen Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften. Der Teil, der für diese Studie verwendet wurde, umfasst 10 Mio. laufende Wörter, nämlich die Texte, die in den Jahren 1990–2000 verfasst wurden.

Tabelle 1. Häufigkeit der Bedeutungsvarianten von *ewig* und *unendlich* im Herder/BYU- und im DWDS-Korpus (14,2 Mio. Tokens)

		ewig			*unendlich*
	Belege	550		***Belege***	276
			1	räumlich ohne Grenze	6%
1	zeitlich ohne Ende/ Grenze	61%	2	zeitlich ohne Ende/ Grenze	7%
2	(emot.) sehr lange	9%			
3	sich immer wiederholend	1%	3	sehr	57%
			4	mathematisch	20%
5	feste Verbindungen	24%	5	feste Verbindungen	4%
			6	hohe Anzahl	5%
			7	fotografisch	1%

Tabelle 1 zeigt in der zweiten Zeile die Anzahl der Belege aus beiden Korpora: *ewig* kommt mit 550 Belegen zirka doppelt so häufig vor wie *unendlich* mit 276 Belegen. In den Spalten 1 und 2 sieht man die Reihenfolge der Bedeutungsvarianten von *ewig* in den Wörterbüchern und Spalte 3 zeigt die Häufigkeitsverteilung der Belegfälle. Die erstgenannte Bedeutung in den Wörterbüchern, *zeitlich ohne Ende/Grenze* ist mit 61% auch die häufigste unter den Belegfällen. An zweiter Stelle der Häufigkeitsskala mit 24% steht allerdings die Verwendung von *ewig* in festen Verbindungen z. B.: *das ewige Leben*, *das ewige Licht*, *die ewige Ruhe*, *ewig und drei Tage, immer und ewig*. Auf solche festen Verbindungen wird nur in zwei der acht untersuchten Wörterbücher hingewiesen. Die in den Wörterbüchern an zweiter Stelle genannte Bedeutung *sehr lange* ist mit 9% vergleichsweise selten, die an dritter Stelle genannte Bedeutung *sich immer wiederholend* ist mit 1% sehr selten.

Die Spalten 4 und 5 zeigen die Reihenfolge der Bedeutungsvarianten von *unendlich* in den Wörterbüchern und die Spalte 6 zeigt die Häufigkeitsverteilung unter den Belegfällen. Hier wird ein eklatanter Unterschied zwischen den Wörterbüchern und der aktuellen Verwendung von *unendlich* sichtbar. Während die in Wörterbüchern zuerst genannten Bedeutungen *räumlich ohne Grenze* und *zeitlich ohne Ende/Grenze* verhältnismäßig selten sind, 6% bzw. 7%, ist die dritte Bedeutung, *sehr*, mit 57% für mehr als die Hälfte der Belegfälle verantwortlich. Selbst die Bedeutung, die in Wörterbüchern an vierter und letzter Stelle steht, nämlich die mathematische Bedeutung, die bei de Gruyter, Duden und Wahrig sogar fehlt, ist mit 20% deutlich häufiger als die erstgenannten Bedeutungen. Dazu gibt es in den Korpora Belegfälle, die nicht wesentlich seltener sind als die Standardwörterbuchbedeutungen, nämlich 5% und 1%, die aber in keinem Wörterbuch auftauchen. Dazu gehört die Bedeutung *hohe Anzahl*, z. B. *Die Wunder, die Kunststoff ermöglichte, schienen unendlich zu sein*, und es gehört dazu die fotographische Bedeutung, z. B. *Der Schärfebereich ist in vier Stufen wählbar: 0,8 m, 1,5 m, 3 m und unendlich.* In festen Verbindungen tritt *unendlich* eher selten auf. Die Forschungsfrage 1: Entspricht die Reihenfolge der Bedeutungen in Wörterbüchern ihrer Auftretenshäufigkeit im wirklichen Leben? sollte in diesem Fall wohl eher mit *nein* beantwortet werden.

Die zweite Frage nach der Kontextsensitivität soll an den häufigsten Varianten beider Adjektive erläutert werden, nämlich der Bedeutung *zeitlich ohne Ende* von *ewig* und der Bedeutung *sehr* von *unendlich*.

Tabelle 2. Wörtliche und kontextsensitive Bedeutungen von *ewig* in der Bedeutung von *zeitlich ohne Ende* im Herder/BYU- und im DWDS-Korpus

ewig	*wörtlich*	*ironisch*	darunter: *nicht ewig*	*insgesamt*
Belege	76	256	36	332
Prozent	23%	77%	11%	100%

Wie Tabelle 2 zeigt, nimmt die wörtliche Bedeutung knapp ein Viertel der Belegfälle ein, 77% der Verwendung von *ewig* ist ironisch gemeint, z. B. *ewiger Zweitligist.* In der Verneinung *nicht ewig* wird das Adjektiv fast nur ironisch verwendet und erhält die Bedeutung *eher kurz*, z. B. *Trödel nicht so herum! Unser Sauerstoffvorrat reicht schließlich nicht ewig.*

Eine andere Art semantischer Selektionspräferenzen zeigt die Verwendung des Adjektivs *unendlich* in der Bedeutung *sehr*, die mit 57% die mit Abstand häufigste Verwendung dieses Adjektivs ist.

Tabelle 3. Semantische Selektionspräferenzen des Adjektivs *unendlich* in der Bedeutung *sehr* im Herder/BYU- und DWDS-Korpus

unendlich	Beispiele	Tokens	Lexeme	Tok %	Lex %
menschl. Eigenschaft	müde, einsam, reich	50	42	34	47
Abstrakta	schwierig	25	23	17	26
Raum	groß, lang, weit	26	12	18	13
Zeit	langsam	12	6	8	7
Menge	viel	35	6	24	7
Insgesamt		*148*	*89*	*100*	*100*

Betrachtet man jedes Lexem einzeln (letzte Spalte der Tabelle 3), so verstärkt das Adjektiv *unendlich* in 47 Prozent der Belegfälle eine menschliche Eigenschaft wie *müde, einsam* oder *reich*. Die andere Hälfte der Belegfälle unterteilt sich in einige wenige weitere Kategorien, nämlich Abstrakta wie *schwierig*, ein weiteres Viertel, und Adjektive oder deadjektivische Substantivableitungen des Raumes wie *groß, lang, weit, Größe, Weite*, der Zeit wie *langsam*, und der Menge wie *viel*.

Diese Daten zu den oft als synonym gehandelten Adjektiven *ewig* und *unendlich* sollten nicht nur an einem Beispiel die Formelhaftigkeit der Sprache zeigen, sondern auch Hinweise auf das Wissen geben, das Muttersprachler über die Wörter ihrer Sprache angesammelt haben, das es ihnen ermöglicht, idiomatisch zu sprechen, und das nur über Korpusanalysen erfassbar ist. Im Folgenden wende ich mich nun meinem zweiten Thema zu, der Häufigkeit von Wörtern und ihr Einfluss insbesondere auf das Leseverständnis.

Häufigkeit

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

Tabelle 4. Textdeckung der häufigsten 1000 bis 5000 Lexeme im amerikanischen Englischen und im Deutschen

	1000	2000	3000	4000	5000
Englisch	72%	79,7%	84%	86,7%	88,6%
Deutsch	72,8%	78,6%	81,7%	83,6%	85%

Wie Tabelle 4 zeigt, erfassen im amerikanischen Englischen die häufigsten 1000 Lexeme ca. 72% der laufenden Wörter eines Textes, die häufigsten 2000 ca. 79,7%, die häufigsten 3000 ca. 84%, die häufigsten 4000 ca. 86,7% und die häufigsten 5000 ca. 88,6% (Kučera/Francis 1967). Nach einer Auszählung basierend auf dem Herder/BYU-Korpus (Tschirner/Jones 2005) erfassen im Deutschen die häufigsten 1000 Wörter ca. 72,8% der laufenden Wörter eines Textes, die häufigsten 2000 ca. 78,6%, die häufigsten 3000 ca. 81,7%, die häufigsten 4000 ca. 83,6% und die häufigsten 5000 ca. 85%.

Entgegen den Annahmen der Lesedidaktik der 70-er und 80-er Jahre des letzten Jahrhunderts haben empirische Untersuchungen ergeben, dass nicht nur 60–70 Prozent der laufenden Wörter eines Textes bekannt sein müssen, um ihn zu verstehen, sondern 95% und mehr (Nation 2001). Um Wörter aus dem Kontext erraten zu können und um unbekanntes Vokabular auf implizite Art und Weise zu lernen – wichtige Ziele der fremdsprachlichen Lesedidaktik – müssen sogar mindestens 97% der laufenden Wörter eines Textes verstanden werden (Nation 2001). Nach Laufer (1997) benötigt man für das Englische einen Wortschatz, der mindestens die 5000 häufigsten Lexeme umfasst, um einen durchschnittlichen Zeitungstext oder Fachaufsatz relativ zügig lesen und verstehen zu können, eine Studie aus den Niederlanden postuliert für akademische Texte sogar einen Mindestwortschatz von 10.000 Lexemen (Hazenberg/Hulstijn 1996).

Wie groß muss der Wortschatz im Deutschen sein, um für unterschiedliche Textsorten die nötige Textdeckung zu erreichen? Um diese Frage zu beantworten, wurden ca. 10% der Texte des Herder/BYU-Korpus getrennt nach Textsorten analysiert. Für jeden Einzeltext wurde ermittelt, welcher Prozentsatz seiner laufenden Wörter (Tokens) über die häufigsten 1000, 2000, 3000 und 4000 Lexeme des Deutschen abgedeckt wird.

Tabelle 5. Subkorpus aus Herder/BYU-Korpus zur Erfassung der Textdeckung der häufigsten 4000 Wörter des Deutschen

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

Tabelle 5 zeigt die ausgewählten Textsorten, die Anzahl der Texte, die Anzahl der Tokens in diesen Texten und die Größe der Subkorpora im Vergleich zum Gesamtkorpus der jeweiligen Textsorte im Herder/BYU-Korpus in Prozent. Für die

gesprochene Sprache wurden ca. 8% der Texte des Herder/BYU-Korpus der gesprochenen Sprache gewählt, für die Zeitungssprache ca. 5% des Zeitungskorpus, allerdings nur aus den Bereichen Politik und Kommentar der Frankfurter Rundschau, der Süddeutschen Zeitung und der Welt. Für die literarische Sprache wurden ca. 20% des Literaturkorpus untersucht und für die Fachsprache ca. 10% des Fachtextkorpus, und zwar Fachzeitschriften und Uni-Einführungen, also die anspruchsvolleren Texte des fachsprachlichen Korpus.

Tabelle 6. Textdeckung ausgewählter Textsorten des Herder/BYU-Korpus durch die häufigsten 4000 Wörter des Deutschen.

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

Tabelle 6 zeigt die Ergebnisse der Textdeckungsstudie zu ausgewählten Textsorten des Herder/BYU-Korpus. Die häufigsten 2000 Wörter des Deutschen decken ca. 90% der Tokens in Texten der gesprochenen Sprache ab. Die häufigsten 4000 Wörter decken knapp 85% der Tokens belletristischer Texte ab, mit den darin enthaltenen Namen kommt man auch auf knapp 90%. Für Zeitungstexte und Fachtexte genügen die häufigsten 4000 Wörter jedoch nicht. Bei Zeitungen kommt man damit auf knapp 80%, mit Namen auf ca. 87%, und bei Fachtexten kommt man selbst mit Namen nur auf ca. 83%.

Erinnern wir uns: Studien zum Englischen haben gezeigt, dass zügiges, verständnisvolles Lesen erst möglich ist, wenn mindestens 95% der Tokens bekannt sind. Fügt man den häufigsten 4000 Wörtern des Deutschen die im Durchschnitt in den untersuchten Texten vorhandenen Eigennamen hinzu, so kommt man in der gesprochenen Sprache schon relativ nahe an diese 95% heran. Bei belletristischen Texten und bei Zeitungstexten genügen die häufigsten 4000 Wörter nicht ganz und bei Fachtexten fehlt noch Einiges. Diese Zahlen deuten darauf hin, dass Wortschatzgröße eine eminent wichtige Rolle beim Lesen spielt.

Haben fortgeschrittene Lerner den zum Lesen nötigen Wortschatz? In einer Studie an der Universität Leipzig wurde ermittelt, wie groß der Wortschatz von Anglistikstudierenden zu Beginn ihres Studiums und nach acht Jahren Englisch-

unterricht in der Schule ist. Das Resultat: 78% hatten einen Lesewortschatz von 2000 Wörtern, 28% von 3000 Wörtern und 21% von 5000 Wörtern (Tschirner 2004). Damit hatte nur jeder Vierte/jede Vierte den für ein Studium minimal notwendigen Lesewortschatz. Diese Ergebnisse weisen darauf hin, dass wir nicht nur vieles nicht wissen, was Häufigkeitseffekte beim Fremdsprachenlernen angeht, sondern auch, dass dieses Wissen viel Potenzial für die Fremdsprachendidaktik enthält. Im Folgenden greife ich nun meinen dritten und letzten Punkt meines Beitrags auf, nämlich die Frage nach der Autonomie der Syntax.

Autonomie der Syntax

Im Gegensatz zur generativen Grammatik wird in der Korpuslinguistik davon ausgegangen, dass zwischen Inhalts- und Funktionswörtern kein qualitativer Unterschied besteht, da beide durch typische Muster, in denen sie auftreten, charakterisiert werden und in diesen Mustern grammatische und lexikalische Elemente untrennbar vereint sind. Wie wir gesehen haben, weisen Lexeme Selektionspräferenzen und -beschränkungen auf. Dies müsste der Grammatik zugeordnet werden, wenn man nach Morris (1938) die Grammatik als die Wissenschaft bezeichnet, die sich mit dem Verhältnis der Zeichen zueinander, mit ihrer Form und ihrer Reihenfolge befasst. Wie wir im Folgenden sehen werden, sind die grammatischen Möglichkeiten von Wörtern ungleich verteilt, so dass man auch von der Lexik in der Grammatik sprechen sollte. Dass dies lange Zeit nicht erkannt wurde, lässt sich vor allem darauf zurückführen, dass Muttersprachlerintuition nur bedingt mit der tatsächlichen Sprachverwendung übereinstimmt, und dass in Grammatikdarstellungen immer wieder die gleichen auffälligen klassischen Beispiele und Belege verwendet werden, die in einem kollektiven Beispielgedächtnis von Forschern und Lexikographen enthalten sind. Diese Armut an Beispielen wird vor allem in Einführungen zur Generativen Linguistik deutlich, in denen es die Autoren oft noch nicht einmal notwendig finden, anstelle der englischen Beispielsätze deutsche zu verwenden.

Eine Zusammenstellung der häufigsten Einheiten bildenden Kollokationen des Herder/BYU-Korpus ergibt eine hohe Anzahl von Präpositionalphrasen. Die häufigste Kollokation *zum Beispiel* würde mit ihrer Häufigkeit von 769 pro eine Million Tokens auf Platz 160 der häufigsten Wörter des Deutschen kommen. *In der Regel* käme mit 97 Okkurrenzen auf Platz 881, *auf jeden Fall* auf Platz 1317 und *in der Stadt* auf Platz 1803. Insgesamt kommen in den Präpositionalphrasen, die unter die häufigsten 4000 Wörter kommen würden, zehn unterschiedliche Präpositionen vor. Die Präposition *in* hat dabei mit 58% der Mehrwortausdrücke den Löwenanteil, die Präposition *auf* kommt auf 18%, die Präpositionen *an* und *mit* auf jeweils 5%.

Interessant ist die Verwendung des Kasus bei den drei Wechselpräpositionen *an*, *auf* und *in*. Die Präposition *an* regiert in allen vier Fällen den Dativ: *am Ende* (Häufigkeit: 98 pro 1 Mio. laufende Wörter), *an der Uni/versität* (51), *an dieser Stelle* (20) und *am nächsten Tag* (19). Die Präposition *auf* regiert in knapp der Hälfte der Fälle den Dativ, immer mit einer lokalen Bedeutung – *auf dem Boden*, *auf dem Tisch*, *auf dem Weg*, *auf der Straße*, *auf der Bühne*, *auf der anderen Seite* – während sie, wenn sie den Akkusativ regiert, meist eine nicht-lokale, übertragene Bedeutung hat – *auf den ersten Blick*, *auf jeden Fall*, *auf keinen Fall*, *auf die Frage*, *auf diese Weise*. Nur in zwei Fällen wird eine grammatische Wahl sichtbar, wobei die Häufigkeiten bei *Weg* in Richtung Dativ tendieren und bei *Tisch* in Richtung Akkusativ. Die Präposition *in* wiederum tendiert in häufigen Mehrworteinheiten eindeutig in Richtung Dativ, nämlich in 93% aller Fälle, wobei die drei Akkusativeinheiten eine Richtungsperspektive beinhalten – *in die Hand*, *in die Stadt*, *in die Schule* – während die Dativeinheiten neben der lokalen Perspektive häufig eine temporale oder übertragene Bedeutung haben. Bei den zwei Substantiven, die sowohl mit dem Dativ wie mit dem Akkusativ auftreten – *Hand*, *Stadt* – überwiegt der Dativ bei *Hand* im Verhältnis von ca. 2 : 1 und bei *Stadt* von ca. 3 : 1.

Zu ähnlichen Ergebnissen kommen Folsom (1984) und Jones (2000). Folsom analysierte 8727 Vorkommen der Wechselpräpositionen im schriftlichen LIMAS-Korpus und Jones untersuchte die Wechselpräpositionen in einem 700.000 laufende Wörter umfassenden mündlichen Korpus.

Tabelle 8. Häufigkeit der Dativrektion in Prozent bei Wechselpräpositionen in schriftlichen (Folsom 1984) und mündlichen (Jones 2000) Korpora.

	in	an	hinter	unter	vor	zwischen	neben	über
Folsom	86,4%	77,9%	89,4%	90,5%	92,5%	96,8%	97,3%	9%
Jones	90%	62,9%	84%	88%	98,5%	99,5%	97%	0,1%

Wie Tabelle 8 zeigt, wird der Dativ bei Folsom und Jones nach *in* in 86,4% bzw. 90% aller Fälle verwendet und nach *an* in 77,9% bzw. 62,9% aller Fälle. Eine noch größere Dativlastigkeit weisen *hinter* (89,4% bzw. 84%), *unter* (90,5% bzw. 88%), *vor* (92,5% bzw. 98,5%), *zwischen* (96,8% bzw. 99,5%) und *neben* (97,3% bzw. 97%) auf, während *über* vor allem eine Adjektivpräposition zu sein scheint. Mit dem Dativ kommt sie in schriftlichen und mündlichen Korpora nur in 9% bzw. 0,1% der Fälle vor.

Sinclair/Renouf (1988: 148) stellen die folgenden Kriterien für einen Wortschatzlehrplan auf: „the commonest word forms in the language, their central patterns of usage, and the combinations which they typically form". Würde man diese Kriterien auf die erwähnten Wechselpräpositionen übertragen und würde man unter „häufigste Wortformen" und „zentrale Verwendungsmuster" die gewählte Häufigkeit von 4000 akzeptieren, müsste man *in* und *an* als Dativpräpositionen

einführen, die ähnlich wie andere Präpositionen auch mal einen anderen Kasus regieren können, bei denen es aber nur in seltenen Fällen eine echte Wahl gibt.

Wenn man beim Spracherwerb von einem Einheitenlernen ausgeht, also einem Speichern von Wortketten, die bei einer genügend hohen Anzahl gleicher oder ähnlicher Wortketten Anlass zur Grammatikkonstruktion geben, d. h. zum Aufbau intuitiver, mentaler, prozeduraler Grammatikregeln, dann könnte man solche im natürlichen Sprachgebrauch häufig vorkommenden Einheiten nutzen, um durch ihren Einsatz (und ggf. ihre Analyse) im Unterricht diesen Grammatikerwerb zu fördern bzw. zu beschleunigen.

Ich habe in meinem Beitrag versucht zu zeigen, wie wichtig Häufigkeitseffekte für das Lehren und Lernen von Fremdsprachen sind und dass viele dieser Effekte erst jetzt untersucht werden können, da wir die dafür notwendigen Datenmengen in digitaler Form und die dafür notwendigen Softwareprogramme haben. Ich habe versucht, dies an drei wichtigen Themen der Korpuslinguistik zu zeigen: der Formelhaftigkeit der Sprache, der Wichtigkeit von Häufigkeitsverteilungen und der nicht vorhandenen Trennbarkeit von Grammatik und Lexik. In einem kurzen Ausblick möchte ich nun zuletzt noch auf ein großes Problem in der Wortschatzdidaktik des Deutschen hinweisen, nämlich der fehlenden wissenschaftlichen Untermauerung der Auswahl der Inhalte.

Ausblick: Grund- und Aufbauwortschätze des Deutschen als Fremdsprache

Ich habe im Vorhergehenden argumentiert, dass ein Wortschatz von mindestens 4000 Wörtern von großer Bedeutung für das Lesen in der Fremdsprache ist und darüber hinaus für den weiterführenden impliziten Wortschatzerwerb, der vor allem über das Lesen stattfindet. Wichtig dabei ist, dass es sich bei diesen Wörtern um die häufigsten 4000 Wörter des Deutschen handelt und nicht um x-beliebige, z. B. die Wörter von 16.000–20.000, die zusammen weniger als 1 Prozent zur Textdeckung beitragen. Grund- und Aufbauwortschätze des Deutschen enthalten meist 4000 Wörter, so z. B. der Grund- und Aufbauwortschatz Deutsch von Langenscheidt (in der aktuellsten Version *Basic German Vocabulary* von 1991). Ein Vergleich dieses Wortschatzbüchleins mit dem *Frequency Dictionary of German* von Routledge (Jones/Tschirner 2006) ergibt gravierende Unterschiede, die insbesondere darauf zurückzuführen sind, dass *Basic German Vocabulary* nicht auf einer empirisch erfassten Häufigkeitsliste beruht, sondern auf einer Kombination und intuitiven Weiterentwicklung der Listen von Kaeding (1897), Meier (1967), Ortmann (1975) u. a. (Langenscheidt 1991, S. VIII).

Ungefähr ein Drittel (32%) der Wörter in *Basic German Vocabulary* gehört nicht zu den häufigsten 4000 Wörtern des Deutschen.[1] Tabelle 9 stellt anhand von drei willkürlich gewählten Bereichen dar, in welche Tausendergruppe nach Jones/Tschirner die ersten 15 Wörter, die von *Basic German Vocabulary* jeweils zur Gruppe 1–2000 der jeweiligen Liste gezählt werden, gehören.

Tabelle 9. Vergleich *Basic German Vocabulary* 1–2000 mit Jones/Tschirner (2006)

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

Insgesamt sind in diesen drei Bereichen nur ein Drittel der Wörter (25 von 75) der *Basic German Vocabulary* Liste 1–2000 unter den häufigsten 2000 Wörtern des Deutschen, ein weiteres Drittel (25) fällt unter die häufigsten 4000 Wörter und das letzte Drittel (25) gehört nicht zu den häufigsten 4000 Wörtern des Deutschen. Man kann sicherlich argumentieren, dass alle Wörter, die in *Basic German Vocabulary* enthalten sind, nützliche Wörter sind, die gelernt werden sollten. Dies ist wahrscheinlich auch der Fall. Das Problematische ist aber der Umkehreffekt, nämlich wie viele der häufigsten 4000 Wörter nicht aufgenommen werden konnten, weil weniger häufige Wörter aufgenommen wurden. Dabei handelt es sich um fast 40 Prozent der häufigsten 4000 Wörter des Deutschen. Tabelle 10 zeigt eine Auswahl der häufigsten 1000 Wörter des Deutschen, die nicht in *Basic German Vocabulary* enthalten sind.

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

Tabelle 10. Auswahl der häufigsten tausend Wörter, die nicht in *Basic German Vocabulary* enthalten sind

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

Eine Untersuchung, die sich mit dem *Grundwortschatz Deutsch* von Klett (1966) befasste, kam zu ähnlichen Ergebnissen (Lipinski/Ebert/Horn 2005). Die Schnittmenge zwischen den 2060 Lexemen des Grundwortschatzes mit den häufigsten 2056 Wörtern des Häufigkeitswörterbuchs von Routledge ist knapp 56%, d. h. 44% der Wörter des *Grundwortschatz Deutsch* sind nicht unter den häufigsten 2056 Wörtern des Deutschen und 44% der häufigsten 2056 Wörter des Deutschen sind nicht im *Grundwortschatz Deutsch* enthalten.

Auch ein Vergleich mit *Profile Deutsch* kommt zu keinem anderen Ergebnis. Die Schnittmenge zwischen den 2089 Wörtern der Niveaustufen A1–B1 (produktiv) und dem *Frequency Dictionary* von Routledge ist 60%, d.h. 40% dieser Wörter gehören nicht zu den häufigsten 4000 Wörtern des Deutschen. Das wirklich Problematische ist hier allerdings wieder der Umkehrschluss. 55% der häufigsten 1000 Wörter ist nicht in den *Profile Deutsch* Listen für A1–B1 (produktiv) aufgenommen. Dazu gehören Wörter wie: *Idee*, *Künstler*, *Vorstellung*, *Bewegung*, *Form*, *Werk*, *Medien*, *Rahmen*; *auftreten*, *beteiligen*, *betreiben*, *klingen*, *stimmen*, *beschäftigen*, *umfassen*; *rund*, *breit*, *eng*, *unmittelbar*, *bewusst*, *gering*, *deutlich*, *notwendig*, und *plötzlich*. Betrachtet man die häufigsten 2000 Wörter des Deutschen, so sind 61% dieser Wörter nicht in *Profile Deutsch* A1–B1 (produktiv) enthalten.

Wenn man nun davon ausgeht – und die Wirklichkeit des Lehrwerkschreibens sieht in der Tat so aus –, dass die Wortschatzvorgaben von *Profile Deutsch* von aktuellen Lehrwerken des DaF übernommen werden, dann wird noch einmal sehr deutlich, wie wichtig korpuslinguistische Untersuchungen für das Lehren und Lernen von Deutsch als Fremdsprache geworden sind.

Literatur

Altenberg, Bengt/Granger, Sylviane (2001): *The grammatical and lexical patterning of ‚make' in native and non-native student writing*. In: Applied Linguistics 22, 173–195.

Balhar, Susanne u. a. (Hrsg.) (2004): *Pons-Großwörterbuch Deutsch als Fremdsprache*. Stuttgart: Klett.

Basic German Vocabulary (1991). Berlin: Langenscheidt.
Biber, Douglas/Conrad, Susan/Reppen, Randi (1994): *Corpus-based approaches to issues in Applied Linguistics*. In: Applied Linguistics 15, 169–189.
Carver, Ronald P. (1994): *Percentages of unknown words in a text as a function of the relative difficulty of the text: Implications for instruction*. In: Journal of Reading Behavior 26, 413–437.
Drosdowski, Günther u.a. (Hrsg.) (1995): Duden – *Das große Wörterbuch der deutschen Sprache*. Mannheim: Duden Verlag.
Eickhoff, Birgit u.a. (Hrsg.) (2002): Duden. *Standardwörterbuch Deutsch als Fremdsprache*. Mannheim: Duden Verlag.
Ellis, Nick (2002): *Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition*. In: Studies in Second Language Acquisition 24, 143–188.
Fandrych, Christian/Tschirner, Erwin (2007): *Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel*. In: Deutsch als Fremdsprache 44, 195–204.
Fenk-Oczlon, Gertraud (2001): *Familiarity, information flow, and linguistic form*. In: Bybee, Joan/Hopper, Paul (Hrsg.): *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins, 431–449.
Folsom, Marvin (1984): *Prepositions with the dative or accusative in written and spoken German*. In: Pfeffer, Alan (Hrsg.): *Studies in descriptive German grammar*. Heidelberg: Groos, 19–32.
Francis, Gill (1993): *A corpus-driven approach to grammar.* In: Baker, Mona/Francis, Gill/Togniniı-Bonelli, Elena (Hrsg.): *Text and Technology. In Honour of John Sinclair*. Amsterdam: Benjamins, 137–156.
Gawlick, Günter (1980): *Geschichte der Philosophie in Text und Darstellung.* Bd. 4: *Empirismus*. Stuttgart: Reclam.
Glaboniat, Manuela/Müller, Martin/Rusch, Paul (2002): *Profile Deutsch*. Berlin: Langenscheidt.
Götz, Dieter/Haensch, Günter/Wellmann, Hans (Hrsg.) (2003): *Langenscheidts Großwörterbuch Deutsch als Fremdsprache*. Berlin: Langenscheidt.
Grundwortschatz Deutsch (1966). Stuttgart: Klett.
Howarth, Peter (1989): *Phraseology and second language proficiency*. In: Applied Linguistics 19, 24–44.
Hu, Hsueh-Chao Marcella/Nation, Paul (2000): *Unknown vocabulary density and reading comprehension*. In: Reading in a Foreign Language 13, 403–430.
Jones, Randall (1997): *Creating and Using a Corpus of Spoken German*. In: Wichmann, Anne et al. (Hrsg.): *Teaching and Language Corpora*. London: Longman, 146–156.
Jones, Randall (2000): *A corpus-based study of German accusative/dative prepositions*. In: Dodd, Bill (Hrsg.): *Working with German corpora*. Birmingham: University of Birmingham Press, 116–142.
Jones, Randall/Tschirner, Erwin (2006): *A frequency dictionary of German. Core vocabulary for learners*. London: Routledge.
Hazenberg, Suzanne, Hulstijn, Jan (1996): *Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation*. In: Applied Linguistics 17, 145–63.
Kaeding, Friedrich Wilhelm (1897/98): *Häufigkeitswörterbuch der deutschen Sprache*. Band 1, 2. Steglitz bei Berlin.
Kempcke, Günther/Seelig, Barbara/Wolf, Birgit (Hrsg.) (2000): *Wörterbuch Deutsch als Fremdsprache*. Berlin: de Gruyter.
Klappenbach, Ruth/Steinitz, Wolfgang (Hrsg.) (1964): *Wörterbuch der deutschen Gegenwartssprache*. Berlin: Akademie-Verlag.
Kučera, Henry/Francis, W. Nelson (1967): *A computational analysis of present-day American English. Providence*, RI: Brown University Press.

Laufer, Batia (1997): *The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess*. In: Coady, James/Huckin, Thomas (Hrsg.): *Second language vocabulary acquisition*. Cambridge: Cambridge University Press, 20–34.

Lemnitzer, Lother/Zinsmeister, Heike (2006): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Lipinski, Silke/Ebert, Alexandra/Horn, Franziska (2005): *Grundwortschatz Deutsch*. Vortrag im Seminar Korpuslinguistik, Herder-Institut, Universität Leipzig am 27.1. 2005.

Meier, Helmut (1967): *Deutsche Sprachstatistik*. Hildesheim: Olms.

Meißner, Cordula (2006): *Die Kontroverse um eine gebrauchsorientierte Beschreibung des Sprachsystems mit Hilfe der Korpuslinguistik – unterschiedliche Perspektiven aus Theorie und Praxis dargestellt am Beispiel der Synonyme „ewig" und „unendlich"*. Abschlussarbeit im Aufbaustudiengang Deutsch als Fremdsprache. Universität Leipzig.

Meißner, Cordula (2008): *Eine gebrauchsorientierte Beschreibung des Sprachsystems mit Hilfe der Korpuslinguistik – das Beispiel der Synonyme* ewig *und* unendlich. In: Deutsch als Fremdsprache 45, 8–13.

Morris, Charles (1938). *Foundations of the theory of signs*. Chicago: The University of Chicago.

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

Ortmann, Wolf Dieter (1975): *Hochfrequente deutsche Wortformen*. München: Goethe Institut.

Qian, David (2002): *Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective*. In: Language Learning 52, 513–536.

Pawley, Andrew/Syder, Frances (1983): *Two puzzles for linguistic theory: Native-like selection and native-like fluency*. In: Richards, Jack / Schmidt, Richard (eds.): *Language and communication*. London: Longman, 191–226.

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

Sinclair, John/Renouf, A. (1988): *A lexical syllabus for language learning*. In: Carter, Ronald/McCarthy, Michael (Hrsg.): *Vocabulary and language teaching*. London: Longman, 140–160.

Specht, Rainer (1979): *Geschichte der Philosophie in Text und Darstellung*. Bd. 5: *Rationalismus*. Stuttgart: Reclam.

Steyer, Kathrin (2000): *Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikographische Möglichkeiten*. In: Deutsche Sprache 28, 101–125.

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

Tschirner, Erwin (2004): *Der Wortschatzstand von Studierenden zu Beginn ihres Anglistikstudiums*. In: Fremdsprachen Lehren und Lernen 33, 114–127.

Tschirner, Erwin/Jones, Randall (2005): *Das Herder/BYU-Korpus der deutschen Gegenwartssprache im deutschen Sprachraum* [Intranet]. Leipzig: Herder-Institut.

Wahrig, Gerhard/Wahrig-Burfeind, Renate (Hrsg.) (1997): *Deutsches Wörterbuch*. Gütersloh: Bertelsmann.

Wray, Alison (1999): *Formulaic language in learners and native speakers*. In: Language Teaching 32, 213–231.

Wray, Alison (2000): *Formulaic sequences in second language teaching: Principle and practice*. In: Applied Linguistics 21, 463–489.

Wray, Alison/Perkins, Mick (2000): *The functions of formulaic language: An integrated model*. In: Language and Communication 20, 1–28.

Yorio, Carlos (1989): *Idiomaticity as an indicator of second language proficiency*. In: Hyltenstam, Kenneth/Obler, Loraine (Hrsg.): *Bilingualism across the lifespan*. Cambridge: Cambridge University Press, 55–72.