ANKE LÜDELING & MAIK WALTER

Korpuslinguistik für Deutsch als Fremdsprache

Sprachvermittlung und Spracherwerbsforschung[1]

1 Einleitung

Die Korpuslinguistik beschäftigt sich mit dem Aufbau, der Auszeichnung und der Auswertung von Korpora, wobei Korpora für einen bestimmten Zweck zusammengestellte (elektronische) Textsammlungen sind (EAGLES 1996). Korpuslinguistische Verfahren können dazu genutzt werden, in Korpora lexikalische Einheiten und formal beschreibbare Strukturen wie beispielsweise Wortklassen (z. B. kausale Konnektoren oder Adjektive mit dem Suffix *-lich*) oder grammatische Muster (z. B. die Verwendung von Hilfsverben, Infinitivgruppen oder Wechselpräpositionen) zu untersuchen.

Dieser Artikel gibt einen Überblick über die Verwendung von Korpora in der Fremdsprachvermittlung und der -erwerbsforschung, denn in beiden Gebieten spielen Korpusdaten und Korpora zunehmend eine wichtige Rolle.[2]

In gewisser Weise werden in der Spracherwerbsforschung und für die Sprachvermittlung natürlich schon seit langem mehr oder weniger systematische Sammlungen von authentischen Sprachdaten eingesetzt (siehe unter vielen anderen z. B. Jordens 1983 zum Kasuserwerb, Edelhoff 1985 zu authentischen Texten in der Vermittlung oder Meyer 2008 für eine Beschäftigung mit vorelektronischen Korpora in der Grammatikschreibung). Viele

---

[1] Der vorliegende Artikel ist eine erweiterte Version des von uns verfassten Artikels *Korpuslinguistik* im derzeit aktualisierten HSK 19 *Deutsch als Fremdsprache*. Für wichtige inhaltliche Hinweise danken wir Amir Zeldes und Hagen Hirschmann. Wir freuen uns über weitere Hinweise und Anregungen. Kontakt: anke.luedeling@rz.hu-berlin.de sowie maik@zedat.fu-berlin.de.

[2] Für die Zwecke dieses Aufsatzes ist die Unterscheidung zwischen Fremd- und Zweitspracherwerb nicht relevant; wir werden daher immer neutral von Sprachvermittlung und Spracherwerbsforschung sprechen.

korpuslinguistische Verfahren (z. B. das Auswerten von Konkordanzen oder Häufigkeitszählungen) sind auch nicht neu oder abhängig von elektronischen Korpora, sondern wurden für vorelektronische Korpora schon genauso verwendet. Mit der weiteren Verfügbarkeit von elektronisch nutzbaren Korpora sind diese Verfahren allerdings einfacher geworden und gerade in quantitativer Hinsicht haben sich in den letzten Jahren für Lehrende und Lernende neue Möglichkeiten der Datennutzung ergeben. Im Folgenden beziehen wir uns daher nur auf elektronisch verfügbare Korpora.

Korpora bestehen meistens aus ganzen Texten oder längeren Textabschnitten – d. h. Korpusdaten sind mit (sprachlichem und außersprachlichem) **Kontext** verfügbar, so dass Kontextfaktoren systematisch ausgewertet werden können. Außerdem sind sie **strukturiert durchsuchbar** – dadurch werden Ergebnisse allgemein nachvollziehbar und reproduzierbar. Korpusdaten, beispielsweise eine digitalisierte Sammlung von Zeitungstexten, können mit zusätzlichen Informationen, wie beispielsweise Wortarten und grammatischen Funktionen aber auch Textsorte oder Erscheinungsdatum angereichert werden. Dieser Prozess, die **Annotation**, erweitert die Suchmöglichkeiten erheblich (vgl. Lemnitzer und Zinsmeister 2006: 60-100). Beispielsweise können – nach passender Annotation - Adverbien, indirekte Objekte oder die im November verfassten Wetterberichte gesucht werden. Außerdem erlauben Korpusdaten neben einer systematischen qualitativen Auswertung auch **quantitative** Untersuchungen. Die Möglichkeiten sind groß, die Fallstricke aber auch: Wie jede Methodik erfordert auch die Auswertung von Korpusdaten bestimmte Vorkenntnisse. Daher spricht man manchmal auch von ‚corpus literacy' (Mukherjee 2002: 179-180). Als eine Form der Medienkompetenz ist sie im modernen Fremdsprachenunterricht zu vermitteln. Wir konzentrieren uns in diesem Artikel auf solche Aspekte der Verwendung von Korpora, die für die Spracherwerbsforschung und Sprachvermittlung wesentlich sind. Für allgemeinere Einführungen in die Korpuslinguistik siehe z. B. Mukherjee (2002, 2008), Lemnitzer und Zinsmeister (2006), Scherer (2006), McEnery, Xiao und Tono (2006); für einen Überblick siehe die Artikel in Kallmeyer und Zifonun (2007), Lüdeling und Kytö (2008, 2009).

Der Artikel ist folgendermaßen gegliedert: Bevor wir in Abschnitt 4 auf Korpusdesign und Korpusvorverarbeitung eingehen, wollen wir motivieren, wie Korpora in der Sprachvermittlung (Abschnitt 2) und in der Spracherwerbsforschung (Abschnitt 3) eingesetzt werden können.

http://zope.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiterinnen/anke/anke_veroeffentlichungen, 04.05.2009

2 Korpora in der Sprachvermittlung

In der Sprachvermittlung werden in den letzten Jahren vermehrt Korpora eingesetzt. Die meisten Verfahren und Szenarios sind bisher für das Englische als Fremdsprache entwickelt worden (siehe z. B. Mukherjee 2002, 2008; Römer 2008, McEnery, Xiao und Tono 2006 oder die Proceedings der Konferenzen *Teaching and Language Corpora*). Prinzipiell sind diese Ansätze, wenn die Korpusressourcen zur Verfügung stehen (vgl. den Überblick über deutsche Korpora in Lemnitzer und Zinsmeister 2006), auch auf das Deutsche als Fremdsprache übertragbar (vgl. hierzu die Diskussion in der Folge von Fandrych und Tschirner 2007). Die Korpora, die hier eine Rolle spielen, sind L1-Korpora, die aus Texten bestehen, die von L1-Sprechern der zu erlernenden Sprache produziert worden sind. Sie können zum einen Informationen über die Häufigkeit linguistischer Einheiten (Wortbestandteile, Wörter, grammatische Strukturen) liefern und zum anderen über die Kontexte, in denen diese auftreten. Damit dienen sie der präzisen Beschreibung der zu erlernenden Sprache.

Im Folgenden werden wir zwei vermittlungsrelevante Anwendungen skizzieren: (1) qualitative Analysen, in denen Korpusdaten mit ihrem Kontext entweder einzeln als Beispiele oder systematisch als Konkordanzen verwendet werden, und (2) quantitative Analysen, in denen die Häufigkeiten von Lexemen, grammatischen Mustern oder anderen Kategorien gezählt und mit anderen Zählungen verglichen werden.

Diese Anwendungen dienen den Lehrenden, den Didaktikern (bspw. in der Erstellung von Lehrmaterialien) und den Lernenden. Diese drei Anwendungsbereiche werden jeweils thematisiert.

2.1 Qualitative Analysen

Bei der Präsentation der Suchergebnisse (der Treffer) einer Korpusabfrage werden Korpusbelege meistens als Konkordanzen ausgegeben, d. h., alle Belege, die auf den eingegebenen Suchausdruck passen, werden untereinander angeordnet (oft als kwic (= keyword in context)-Konkordanz bezeichnet). Der Suchausdruck kann sich dabei auf alles beziehen, was im Korpus kodiert ist. Abb. 1 zeigt eine Konkordanz für eine lexikalische Suche, Abb. 2 eine Suche, in der die Wortartannotation mitberücksichtigt wurde. Die Referenzen für alle hier angesprochenen Korpora finden sich in Abschnitt 6.1. In der Suche können je nach Suchwerkzeug Platzhalter, reguläre Ausdrücke etc. verwendet werden. Für einen Überblick über Suchstrategien und Konkordanzen siehe z. B. Wynne (2008).

Kontext	Treffer	Kontext
gerade	**wegen**	der Anfeindungen seiner deutschen Zeitgenossen
nur	**wegen**	ihres umfangreichen Illustrationszyklus, sondern
auch	**wegen**	ihrer gekürzten Textfassung des
Auslieferung	**wegen**	des ungeklärten Verhaltens des Protagonisten
Auswertungen	**wegen**	der engeren Zusammenhänge zwischen den
Bedenken	**wegen**	der aus der Sicht der
,	**wegen**	des Einflusses der Feminisierung von
als	**wegen**	des Embryos um seiner selbst
Täters	**wegen**	geminderter Tatschuld dann gemildert werden
Gläubigers	**wegen**	einer Leistungsstörung aufbauen. Ein

Abb. 1 KWIC-Konkordanz der ersten 10 Belege des Suchausdrucks wegen im Korpus Akademisches Deutsch 2006

Kontext	Treffer	Kontext
empirische	**Untersuchung von**	neueren (ab 2000)
einer	**Dienstleistungsorientierung gegenüber**	dem Bürger untersucht werden.
Die	**Auswertung von**	fMRI-Bildern ergab, dass das
eine	**Verknüpfung von**	visuellen Arealaktivitäten mit semantischem Wissen
Wedekinds	**Auseinandersetzung mit**	Wedekinds Friedrich Nietzsches Philosophie war ein
operative	**Behandlung von**	Mundhöhlenkarzinomen verursacht wurden, unter
die	**Konsonantenbeschreibung von**	Relevanz sind: Dauer,
der	**Differenzierung zwischen**	/s/ und /S/
die	**Veränderung von**	F1 und F2 sowie die
die	**Zuweisung von**	Genus im Deutschen wurden verschiedene

Abb. 2 KWIC-Konkordanz der ersten 10 Belege mit ung-Nominalisierungen und adjazenten Präpositionen im Korpus Akademisches Deutsch 2006. Für diese Suche wurden reguläre Ausdrücke (Mustersuchen) auf der String-Ebene und der Wortartannotationsebene verwendet.

Der Status der gefundenen Belege ist einfach: Mit dem Beleg in einem Korpus kann zunächst lediglich gezeigt werden, dass eine sprachliche Struktur verwendet wird. Aus dem Vorkommen folgt nichts über die Grammatikalität oder Akzeptabilität der Struktur – über die Grammatikalität kann nur ein vom Korpus unabhängiges Verfahren (Muttersprachler, Grammatik etc.) entscheiden. Aus dem Nichtvorkommen einer Struktur kann nichts über Verwendung und Grammatikalität abgeleitet werden, denn es darf in einer Korpusanalyse zunächst nur auf das jeweils untersuchte Korpus geschlossen werden (der Schluss auf ungesehene Daten ist nur unter bestimmten statistischen Voraussetzungen zulässig; man kann nicht ohne weiteres auf **die** Sprache oder **die** Varietät schließen), siehe Abschnitt 4.1. Etwas allgemeiner formuliert dient ein Korpus der Suche nach einem authentischen Beispiel für ein Wort oder eine Struktur. Ein Korpus der gesprochenen Sprache (hier Gespräche zwischen Müttern und ihren Töchtern) kann beispielsweise schnell Belege für die mit *weil* eingeleiteten V2-Sätze liefern (Beispiele (1) und (2)), die dann weiter untersucht werden können.

(1) *du hasch noch nie erlebt * daß ich irgendwas angezogen hab was dir net (gfällt) doch das was mir net gfällt des kommt oft vor weil über * also über gschmack läßt sich überhaupt net streite* **weil** *ich hab mein gschmack und du hasch dein gschmack und wenn dirs net gfällt dann * dann liegt des halt weil du=n andre gschmack hasch aber pf da kamma net drüber streite des geht net* (EK-Korpus, 1844)

(2) *nee* ***weil*** *des isch dein gschmack und ich ich hab mein gschmack* (EK-Korpus, 2028)

Lehrende können in dieser Weise das Korpus als Referenzquelle in zwei Situationen verwenden. Zum einen können gezielt Unterrichtsmaterialien für einen zu vertiefenden sprachlichen Gegenstand mit authentischen Daten erstellt werden:

(a) Übersichten zu grammatischen Strukturen wie dem Gebrauch der Modalverben in geschriebener und gesprochener Sprache und
(b) Arbeitsblätter zu scheinbaren Synonymen wie beispielsweise *ewig* und *unendlich* (Meißner 2008).

Zum anderen können Korpora als Orientierungspunkt für die Korrektur von Lernertexten fungieren. Problematisch sind in der Korrektur für die Lehrenden vor allem sprachliche Strukturen, die zwar grammatisch korrekt, aber nicht dem jeweiligen Kontext angemessen sind. Insbesondere nichtmuttersprachliche Lehrende einer Fremdsprache profitieren deshalb von Korpora. Auf Mukherjee (2002) geht in diesem Zusammenhang der Vorschlag einer abstrakten Korpusnorm zurück, die als Orientierungsgröße bei Korrekturen herangezogen wird und damit das rein intuitive Urteil des Korrektors (und damit dessen „Allmacht“) ablöst. Wegen des oben angesprochenen Belegstatus’ ist dieser Vorschlag allerdings mit Vorsicht zu betrachten.

Als didaktisches Konzept hat sich das datengesteuerte Lernen (data-driven learning) etabliert: Der Lerner erhält einen (sinnvoll ausgewählten und ggf. manipulierten) Input als Lernmaterial. Dieser besteht aus Konkordanzen (siehe die aus einigen der Beispiele aus Abb. 2 erstellten Lückentexte in Abb. 3, mit denen die Einsetzung von Präpositionen geübt wird), aus denen der Lerner versucht, selbständig Gebrauchsregeln abzuleiten, sie zu „entdecken“ (Johns und King 1991, Johns 2000 und Bernardini 2004).

Diese Arbeit ist eine empirische **Untersuchung ___** neueren (ab 2000) Verwaltungstexten.
Die **Auswertung ___** fMRI-Bildern ergab , dass das Konzept Farbe nicht in einer spezifisch eigenen Gehirnregion verarbeitet wird
Bei den Hochschullehrenden wird betrachtet , auf welche Weise sie **Anweisungen ___** Studierende formulieren , mit eigenen Wissensressourcen umgehen
Frank Wedekinds **Auseinandersetzung ___** Friedrich Nietzsches Philosophie war ein lebenslanger Prozeß

Abb. 3: Beispiel eines ad hoc aus Korpusbelegen erstellten Lückentextes für eine spezifische Aufgabe

Beispielbanken werden bereits seit langem in der Vermittlung von Fachsprachen eingesetzt, denn hier benötigt man – weil oft kein geeignetes Material vorhanden ist – sehr spezifische, möglichst aktuelle Spezialkorpora, die selbst kompiliert und anschließend durchsucht werden können (vgl. Bowker und Pearson 2002, O´Keeffe, McCarthy und Carter 2007).

In der **Didaktik** sollten diese in der Praxis entwickelten Ansätze einer kritischen Prüfung unterzogen werden. Für DaF wurden bislang keine entsprechenden empirischen Arbeiten vorgelegt. Ein ähnliches Bild liefern auch die Lehrmaterialien. Authentische Beispiele sind zwar in vielen aktuellen Lehrwerken zu finden, aber von einem systematischen Einsatz von Korpusdaten kann in DaF keine Rede sein. Inwieweit Korpusdaten sinnvoll eingesetzt werden können, ist sicher zu diskutieren. Während Römer (2006) zum Beispiel dafür plädiert, dass Lehrmaterialien die ‚Wirklichkeit' (d. h. die Verteilung in Referenzkorpora) abbilden, warnt Meunier (2002) vor einer Verabsolutierung und argumentiert, dass ein gut konstruiertes Beispiel häufig eine Struktur viel besser illustrieren kann als ein Beleg aus dem Korpus (siehe dazu auch die Diskussion von Widdowson und Sinclair in McEnery, Xiao und Tono 2006: 131-144). Trotzdem ist – insbesondere bei fortgeschrittenen Lernern – authentischer Input ein gutes Fundament einer reichhaltigen Lernumgebung. Gut (2007) zeigt, dass das auch für die gesprochene Sprache gilt.

Zusatzmaterialien wie Wörterbücher (vgl. den Überblick in Rothenhöfer, erscheint), die meisten Lernerwörterbücher des Englischen (z. B. des Cobuild-Projektes) und auch einige Lernergrammatiken (z. B. Rug und Tomaszewski 2001) nutzen authentische Beispiele. Ob deren Quelle ein Korpus oder aber das gute Gedächtnis der Lehrwerksautoren war, ist dabei nebensächlich. Die Lernerwörterbücher des Cobuild-Projektes haben Korpusdaten und deren Auswertungen systematisch eingebunden (siehe Abschnitt 2.2).

Auch **Lerner** können mit einer solchen Bank gewinnbringend arbeiten. Hierbei kommt es darauf an, Lerner für die unterschiedlichen Korpustypen zu sensibilisieren, d. h. ihnen bewusst zu machen, in welchen Korpora sie welche Sorte von Beispielen finden können und wie Suchabfragen aufgebaut werden. Daneben kann ein sprachliches Bewusstsein für das Verhältnis von Norm und „Varianz“ und die Unterschiede zwischen Varietäten gefördert werden. (wenn beispielsweise Sätze wie (1) und (2) als ‚gesprochen’ und nicht als ‚falsch’ eingeführt werden). Wenn Lerner wissen, wie sie gezielt suchen können und welche Rolle Kontexte und Varietäten für das Vorkommen einer Form spielen können, verwenden sie vielleicht weniger unkritisch problematische Suchen auf unbekannten Texten (mit bspw. Google; siehe Kilgarriff 2007). Gaskell und Cobb (2004) illustrieren, wie Lerner ihre produktiven Fertigkeiten in der geschriebenen Sprache mithilfe von Konkordanzen verbessern können.

Neben den einsprachigen L1-Korpora sind Parallelkorpora (ein Text/Korpus mit seinen Übersetzungen, im Idealfall satzweise aligniert; für andere Definitionen oder Verwendungen von Parallelkorpora siehe Johansson 1998, Olohan 2004) eine wichtige Quelle für den Unterricht mit fortgeschrittenen Lernern. Solche Korpora sind im Unterricht gut einzusetzen (allerdings besteht natürlich immer das Problem der Übersetzungseffekte oder des ‚Translationese’, siehe Gellerstam 1986, Baroni und Bernardini 2006 und andere). Frankenberg-Garcia (2005) z. B. zeigt, wie mit den entsprechenden Konkordanzen der Wortschatz systematisch erweitert werden kann.

Damit Korpora im Unterricht sinnvoll eingesetzt werden können, plädieren wir für ein spezifisches Korpustraining für die Lernenden. Ein solches Training soll eine *corpus literacy* aufbauen. Dies beginnt mit rein motivierenden Sequenzen, führt über die Erkundung von Möglichkeiten und Grenzen der Korpora hin zur Operationalisierung von komplexen sprachlichen Problemfällen. Ein Training sollte möglichst gebündelt erfolgen und bietet sich für die Projektarbeit an (möglichen Einwänden gegen ein solches Vorgehen begegnet

Conrad 2008). Ein ausformuliertes Training für DaF ist derzeit noch ein Desiderat und bleibt dringende Aufgabe der DaF-Didaktik.

2.2 Quantitative Analysen

Korpusdaten können gezählt werden (dass das Zählen natürlich kein Selbstzweck ist, sondern eine gut begründete Forschungsfrage voraussetzt, ist selbstverständlich). Es gibt Korpuswerkzeuge, mit denen Wortlisten mit Korpusfrequenzen erstellt werden können. In annotierten Korpora können auch Annotationskategorien gezählt werden. So können in einem Korpus, das mit Wortarten annotiert ist, z. B. die häufigsten Vollverben gesucht werden. Frequenzen sind immer nur im Vergleich zu anderen Frequenzen interpretierbar. Dabei können die Frequenzen von bestimmten Kategorien (Lemmata, Wortarten etc.) innerhalb eines Korpus oder Frequenzen einer Kategorie zwischen verschiedenen Korpora verglichen werden. So können z. B. Wörter oder Wortarten getrennt nach verschiedenen Registern, Domänen u. ä. aufgelistet werden. Abb. 4 zeigt die 9 häufigsten Vollverben in verschiedenen Domänen des Korpus Akademisches Deutsch 2006, im c’t-Korpus und im Bundestagsredenkorpus. Man sieht, dass die verschiedenen akademischen Register im Korpus Akademisches Deutsch, das Zusammenfassungen von wissenschaftlichen Dissertationen enthält, einander bei den häufigen Verben sehr ähnlich sind und sich von den anderen Registern unterscheiden. Registerunterschiede sind also schon bei den häufigen Wörtern zu finden. Fachsprachliches Vokabular findet man dann bei den ‚mittelhäufigen’ und seltenen Wörtern.

Akademisches Deutsch			c't	Bundestags-reden
Land-wirtschaft	**Medizin**	**Germanistik**		
zeigen	*Zeigen*	*zeigen*	*lassen*	*sagen*
untersuchen	*untersuchen*	*lassen*	*geben*	*geben*
entwickeln	*nachweisen*	*gehen*	*kommen*	*machen*
identifizieren	*Finden*	*stellen*	*bieten*	*kommen*
liegen	*kommen*	*versuchen*	*finden*	*gehen*
nachweisen	*vergleichen*	*beschreiben*	*sehen*	*tun*
finden	*ergeben*	*untersuchen*	*machen*	*wissen*
stellen	*liegen*	*finden*	*stehen*	*stellen*
ergeben	*bestimmen*	*Stehen*	*liegen*	*stehen*

Abb. 4 Die 9 häufigsten Vollverben im Korpus Akademisches Deutsch 2006 in den Domänen Landwirtschaft, Medizin, Germanistik (Literaturwissenschaft), in der Zeitschrift c't und in den Bundestagsreden

Lehrende können Frequenzlisten für curriculare Entscheidungen heranziehen. Dahinter steht die Annahme, dass häufige Strukturen vermittlungsrelevanter sind als seltene (Leech 2001). Aus dem Vergleich von Frequenzlisten von Fachsprachekorpora mit allgemeinsprachlichen oder gemischten Korpora kann Fachvokabular ermittelt werden. Coxhead (2000) hat zum Beispiel eine solche – sehr einflussreiche - korpusbasierte Wortliste der englischen Wissenschaftssprache erstellt, die auch in der Fremdsprachendidaktik diskutiert wurde (Nation 2001).

Die quantitative Auswertung und der vergleich zwischen verschiedenen Kategorien muss sich natürlich nicht nur auf Wörter beziehen. Alle im Korpus annotierten Kategorien und Kombinationen davon können gezählt und verglichen werden. Die dabei verwendeten statistischen Verfahren sind vielfältig; wir wollen hier nur zwei Bereiche erwähnen, die bei der Erstellung von Lehrmaterialien und für den Unterricht eine Rolle spielen, nämlich das Zusammenspiel von einzelnen lexikalischen Einheiten zu größeren Einheiten und die Ermittlung von grammatischen und stilistischen Eigenschaften von Varietäten.

Neben einzelnen Wörtern spielen typische Kombinationen von Wörtern (Kollokationen, s. z. B. Choueka 1988, Evert 2009; lexical bundles, s. Salem

http://zope.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/anke_veroeffentlichungen, 04.05.2009

1987, Altenberg und Eeg-Olofsson 1990, Biber, Conrad und Cortes 2004) und häufige grammatische Strukturen (Chunks, pre-fab units) für die nativelike fluency (und daher natürlich für die Vermittlung) eine Rolle (siehe dazu z. B. Aguado 2002, Handwerker und Madlener 2009, Artikel 25 im überarbeiteten HSK DaF). Solche Kombinationen können aus Korpusdaten berechnet werden. Lexical bundles sind einfach häufige feste Mehrwort-Kombinationen (i. d. R. ab 4 Wortformen). Kollokationen sind dagegen teilweise flexible Kombinationen, die ‚unwahrscheinlich' häufig zusammen (auch nah beieinander) auftreten. Ganz grob gesagt, wird dazu berechnet, welche Kombinationen von Wörtern (signifikant) häufiger auftreten als man es erwarten würde, wenn die Wörter unabhängig wären. Es reicht hier nicht aus, die bloße Häufigkeit der Kombinationen zu betrachten, da die Häufigkeit der einzelnen beteiligten Wörter für die ‚Unwahrscheinlichkeit' der Kombinationen eine Rolle spielt (für einen Überblick über die statistischen Verfahren siehe z. B. Evert 2009). Ähnlich könnte man dies auch für typische syntaktische Strukturen berechnen.

Gesprochene und geschriebene Varietäten unterscheiden sich nicht nur qualitativ, sondern hauptsächlich quantitativ. Quantitative Unterschiede sind schwierig zu vermitteln; häufig fehlt auch eine abgesicherte Datengrundlage. Zur Beschreibung von grammatischen und stilistischen Unterschieden zwischen Varietäten können multifaktorielle Verfahren herangezogen werden. In solchen Verfahren werden in zwei oder mehr Varietäten viele unterschiedliche Merkmale (Wortartverteilung, Satzlänge, Tempus, Verben in der zweiten Person etc.) gezählt. Jedes dieser Merkmale ist eine Dimension, in der sich die beiden Varietäten unterscheiden. Da eine solche multifaktorielle Situation für uns schwer zu verstehen ist, werden diese Faktoren dann auf einige wenige reduziert, indem Korrelationen zwischen den Merkmalen berechnet werden. Mit einer Hauptkomponentenanalyse (oder ähnlichen Verfahren) können dann sogenannte Faktoren berechnet werden, das sind solche Kombinationen von Korrelationen, die möglichst viele Merkmale abdecken. Nun folgt ein Interpretationsschritt: Es wird angenommen, dass die Korrelationen nicht zufällig sind, sondern auf eine gemeinsame Ursache zurückgeführt werden können. Mit einigen wenigen solcher Faktoren können dann die Unterschiede zwischen den Varietäten beschrieben werden (Biber, Conrad und Reppen 1998, Biber und Conrad 2009). Die so gewonnenen Erkenntnisse können für den Unterricht und die Erstellung von Lehrmaterialien genutzt werden (Biber et al. 2002).

In der **Didaktik** können Frequenzlisten für die empirische Absicherung eines Curriculums eingesetzt werden (Römer 2008: 114, Tschirner 2005, 2008). Die Frequenz von Wörtern, aber vor allem auch von lexikogrammatischen Mustern ist hierbei für die Behandlung im Unterricht eines der zentralen Kriterien (Leech 2001). Die Frequenz sprachlicher Strukturen bildet den wesentlichen Baustein korpusbasierter Lehrmaterialien. Pionier auf diesem Gebiet ist das Cobuild-Projekt unter der Leitung von John Sinclair. 1980 begannen dort die Arbeiten der späteren Bank of English, eines der größten englischsprachigen Korpora, aus denen neben Lernerwörterbüchern und -grammatiken auch zahlreiche Zusatzmaterialien hervorgingen (Mukherjee 2002: 37-38). Andere Projekte folgten diesem Vorbild. Darüber hinaus gibt es bereits vielfältige korpusbasierte multimediale Anwendungen, in denen Lerner selbstständig Konkordanzen aus Korpora erstellen können.

Neben der Lehrmaterialproduktion beginnt sich eine korpusbasierte Lehrwerksanalyse herauszubilden – hier wird das Lehrwerk selbst ein Korpus, das dann mit anderen Korpora verglichen werden kann. Römer (2005) hat am englischen Progressiv demonstriert, dass sich die Frequenzverteilung von einigen Kategorien in den Lehrwerken zum Teil drastisch von der Frequenzverteilung im ‚normalen' Englisch unterscheidet, was wiederum zu durchaus vermeidbaren negativen Effekten des Faktors *Instruktion* führen kann.

In den DaF-Lehrmaterialien finden sich bislang keine korpusbasierten Frequenzangaben (mit Ausnahme von Jones und Tschirner 2006, einem Frequenzwörterbuch, das sich auch an DaF-Lerner wendet).

Um **Lerner** selbst mit Häufigkeitslisten vertraut zu machen, ist es wichtig, dass sie einfache Auswertungstechniken kennen (beispielsweise den gravierenden Unterschied zwischen absoluten und relativen Häufigkeiten). Wir schlagen die folgende Progression vor, um dieses Grundwissen zu vermitteln:

(1) Nachvollzug von einfachen Korpusanalysen (z. B. Meißner 2008)
(2) Interpretation von kleineren Listen, die beispielsweise ausgewählte Strukturen vergleichen (vgl. Abb. 4)
(3) Fokussierung auf sprachliche Phänomene und relevante Kontexteigenschaften in größeren Listen (z. B. Auxiliarselektion beim Perfekt)
(4) Erstellung eigener Listen auf der Basis eines sprachlichen Unterrichtsgegenstandes (z. B. Modalpartikeln).
(5) Extraktion eigener Häufigkeitslisten in selbst zusammengestellten Korpora.

Diese fünf Schritte können Teil eines Korpustrainings (s.o.) sein. Es gibt eine ganze Reihe von weiteren Vorschlägen, wie Lerner an Korpora herangeführt werden und eigene Korpora bauen und auswerten können (vgl. Aston 2002). Insbesondere in der Vermittlung von Fachsprachen (bzw. in der Übersetzerausbildung) hat sich dieses Szenario bereits bewährt. Die Lerner können beispielsweise Musiktexte, juristische Texte oder Börsenmeldungen sammeln und mit frei verfügbaren Konkordanzwerkzeugen auswerten, unter anderem auch in Form von Häufigkeitslisten (zur kritischen Zusammenstellung von Konkordanzwerkzeugen vgl. z. B. Barlow 2004 oder Wiechmann und Fuhs 2006). Neben der zu erwerbenden Zielsprache können Lerner auch die eigene Lernersprache (Mukherjee und Rohrbach 2006) oder aber Lerneräußerungen im Allgemeinen betrachten. Eine wichtige Ressource sind hierbei so genannte Lernerkorpora, auf die wir im nächsten Abschnitt eingehen.

3 Korpora in der Spracherwerbsforschung

Um den Ablauf des Spracherwerbs (Interimssprache, interlanguage, Selinker 1972) zu erforschen, können nur von Lernern produzierte Daten betrachtet werden. Dies sind zum einen experimentell erhobene Daten und zum anderen authentische schriftliche oder mündliche Äußerungen von Lernern. Systematisch erstellte und gut dokumentierte Lernerkorpora (L2-Korpora) eignen sich für die Erwerbsforschung besser als bloße Fehlersammlungen, da (a) der Kontext einer verwendeten Äußerung mitbetrachtet werden kann, (b) sowohl ‚abweichende' und ‚zielsprachliche' Äußerungen desselben Lerners bei einer bestimmten Konstruktion untersucht werden können und (c) quantitative Vergleiche zu L1-Daten oder Lernerdaten anderer Lernergruppen (im Längsschnitt oder im Querschnitt) möglich sind. Daher werden immer häufiger Lernerkorpora als Datenbasis für die Erwerbsforschung eingesetzt (Granger 2002, 2008; Skiba 2008; Walter und Grommes 2008).

In den folgenden Abschnitten beschäftigen wir uns mit der Auswertung von Lernerkorpora in der Spracherwerbsforschung. Die Nutzung von Lernerkorpora in der Vermittlung, bspw. zur Bewusstmachung von Lernschwierigkeiten (vgl. Mukherjee und Rohrbach 2006) oder zur Erstellung von Lehrmaterialien (Tono und Aoki 1998) steht noch ganz am Anfang und wird hier daher nicht weiter behandelt.

3.1 Fehleranalyse

Lernerdaten sind oft auf Abweichungen zu einer Norm oder einer angenommen ‚korrekten' Form ausgewertet worden. An dieser Stelle können wir auf die Diskussion zur Definition von ‚Fehler' oder die Bedeutung von Lernerfehlern nicht eingehen (siehe dazu Corder 1981, Kleppin 1997, Ellis und Barkhuizen 2005: 51-72 sowie Artikel 120 im überarbeiteten HSK DaF), sondern konzentrieren uns auf die Verfahren zu einer systematischen Fehlerannotation.

Für eine systematische Fehleranalyse ist eine Fehlerannotation sinnvoll, da so die Analyse transparent und reproduzierbar wird (s. u.). Fehlerannotation bedeutet, dass jeder Fehler mit einem Fehlertag versehen wird. Dazu sind wie bei jeder Form von Kategorisierung mehrere Entscheidungen zu treffen. Zunächst muss ein Fehlertagset (eine Menge möglicher Fehlerkategorien) entwickelt werden, dann muss entschieden werden, an welcher Stelle ein Fehlertag ‚andockt' (der Fehlerexponent). Für Überlegungen dazu und vorhandene Fehlertagsets (zumeist zum Englischen als Fremdsprache) siehe z. B. Dagneaux et al. (1996), Izumi, Uchimoto und Isahara (2005), Granger (2002, 2008), Doolittle (2008).

Das Fehlertagset wird immer vom Untersuchungszweck abhängen – ein Tagset für Wortbildungsfehler wird anders aussehen als ein Tagset für Wortstellungsfehler. Bei der Erstellung von Tagsets muss immer zwischen Granularität und Handhabbarkeit abgewogen werden. Zu einem Tagset gehören allgemein verfügbare Vergaberichtlinien und idealerweise wird in einem kontrollierten Verfahren eine Urteilerübereinstimmung (Inter-Annotator Agreement, Carletta 1996) ermittelt.

Ein wesentliches Problem bei der Fehleranalyse ist die notwendige Annahme einer **Zielhypothese** – ein Fehler kann nur angenommen werden als eine Abweichung von einer (implizit angenommenen oder explizit angegebenen) ‚korrekten' Form. Da es oft viele verschiedene solche ‚korrekten' Formen geben kann, kann dieselbe Struktur unterschiedlich analysiert werden. Dies kann zu sehr unterschiedlichen Hypothesen über den Erwerb führen (siehe Lüdeling 2008). In folgender Lerneräußerung (3, aus dem Falko-Korpus, Lüdeling et al. 2008) kann man je nach Zielhypothese annehmen, dass – neben einem Interpunktionsfehler – ein *sich* zu viel verwendet wurde

(3', ein Argumentstrukturfehler) oder dass ein falsches Verb gewählt wurde (3'', ein lexikalischer Fehler).

(3) *Bei der akademischen Ausbildung lernt er weiterhin sich schnell neues Wissen zu erwerben* […] (Falko cbs012_2006_09)
(3') ZH: Bei der akademischen Ausbildung lernt er weiterhin, schnell neues Wissen zu erwerben […]
(3'') ZH: Bei der akademischen Ausbildung lernt er weiterhin, sich schnell neues Wissen anzueignen […]

Die Zahl der möglichen Zielhypothesen ist im Prinzip unbegrenzt und es gibt oft keine Möglichkeit, systematisch zwischen mehreren Zielhypothesen zu unterscheiden (das wird als einer der Gründe gegen Fehlerannotation angeführt, siehe dazu z. B. die Diskussion über die ‚comparative fallacy' von Bley-Vroman 1983 oder Year 2004 sowie Tenfjord, Hagen und Johansen 2006). Allerdings muss man sehen, dass auch Analysen, die keine explizite Fehlerannotation machen, auf einer impliziten Interpretation der Daten beruhen. Ohne die zugrundeliegenden Daten und die Interpretationsgrundlage sind die Ergebnisse von Fehlerstudien nicht einzuordnen. Daher ist es für die Nachvollziehbarkeit einer Analyse erforderlich, dass die Daten mit Zielhypothese und Annotation öffentlich zugänglich sind.

3.2 Kontrastive Analyse

Korpora sind auch die Datenquelle für kontrastive Untersuchungen. Dabei werden Frequenzen einer gegebenen Kategorie (Wörter, Wortarten, Wortartfolgen, Fehler einer bestimmten Art etc.) in einem Lernerkorpus mit Frequenzen derselben Kategorie in einem anderen Korpus verglichen. Das Vergleichskorpus kann ein L1-Korpus sein oder ein anderes L2-Korpus, das bspw. andere Textsorten enthält oder von Lernern mit einer anderen L1 oder einem anderen Sprachstand erstellt wurde.

Auch hier können wir nicht die Kontroverse um die kontrastive Analyse an sich aufrollen, wir können aber feststellen, dass eine neu konzipierte kontrastive Analyse, die die Interferenz nicht als alleinige Ursache für die (abweichende) Struktur von Lerneräußerungen ansieht (Contrastive Interlanguage Analysis, Granger 2002), in den letzten Jahren sehr viel und sehr sinnvoll

verwendet wird (Granger 2008 sowie die umfangreiche CECL-Bibliographie, s. u.).

Das Verfahren selbst – ein quantitativer Vergleich von zwei oder mehr Korpora im Bezug auf ein Merkmal - ist nicht lernerkorpusspezifisch – alle in Abschnitt 2.2 beschriebenen Verfahren können hier genauso verwendet werden. Dabei sind in allen Studien zwei Themen wesentlich, (a) die Vergleichbarkeit der Korpora und (b) die Interpretation der gefundenen Zahlen. Da Korpusfrequenzen von sehr vielen Faktoren abhängen (die meisten davon kennen wir wahrscheinlich noch gar nicht), müssen die Korpora, die betrachtet werden, in möglichst vielen Designparametern übereinstimmen (siehe dazu Abschnitt 4.1) – im Idealfall sollten sie sich *nur* in dem zu betrachtenden Parameter unterscheiden. Wenn die Korpora vergleichbar sind und die Frequenzen normalisiert sind, müssen die gefundenen Zahlen interpretiert werden. Die Interpretation der Ergebnisse ist natürlich abhängig von den gewählten Vorannahmen. Daher ist es auch hier wesentlich, dass die Rohdaten zur Verfügung stehen und alle Interpretationsschritte kenntlich gemacht werden.

Zu Illustration soll hier kurz die Methodologie für eine Mindergebrauchsstudie (underuse) vorgestellt werden (siehe Zeldes, Lüdeling und Hirschmann 2008), in der explorativ struktureller Mindergebrauch bei Lernern des Deutschen als Fremdsprache ermittelt werden soll. Die Daten stammen wieder aus dem Lernerkorpus Falko. In dieser Studie sollen solche Wörter und Wortartkategorien gefunden werden, die (a) häufig (also den Lernern bekannt) sind und (b) von Deutschlernern (unabhängig von deren L1) signifikant seltener gebraucht werden als von L1-Sprechern, da solche Strukturen auf strukturelle Lernschwierigkeiten hindeuten können. Abb. 5 und Abb. 6 zeigen einige der betrachteten Frequenzen für Wortformen und Wortartbigrammen (Folgen von zwei Wortartkategorien). Man kann sehen (Abb. 5), dass die Lerner bestimmte Wortformen wie etwa das Reflexivpronomen *sich* im Vergleich zu den L1-Sprechern zu selten gebrauchen (dunkle Farben in der Minus-Zeile zu diesem Wort), also vermeiden. Die Zeilen, in denen alle Spalten (also Lerner mit verschiedenen L1) einen Mindergebrauch anzeigen, sind potentiell interessant. Da das Korpus mit Wortarten getaggt ist, können auf dieselbe Art auch Wortartfolgen analysiert werden, siehe Abb. 6, wo Mindergebrauch von zwei aufeinanderfolgenden Adverbien deutlich wird. (Dies würde auch für jede andere Annotationsebene funktionieren.)

		Dänisch	**Englisch**	**Französisch**	**Polnisch**	**Russisch**
auch	+					
	-					
für	+					
	-					
sind	+					
	-					
sich	+					
	-					
Ich	+					
	-					

Abb. 5: Hier wird schematisch Über- und Mindergebrauch von verschiedenen Wörtern angezeigt. Dabei wird in der jeweiligen ‚+'-Zeile Übergebrauch (Lerner benutzen das Wort im Vgl. zu L1-Sprechern zu häufig) angezeigt und in der jeweiligen ‚-'-Zeile Mindergebrauch (Lerner benutzen das Wort im Vgl. zu Muttersprachlern zu selten). Nur statistisch signifikante Werte werden angezeigt. Die Intensität der Färbung zeigt den Grad des Über- oder Mindergebrauchs an (je dunkler, desto stärker). In den Spaltenüberschriften stehen die Muttersprachen der Lerner.

		Dänisch	Englisch	Französisch	Polnisch	Russisch
PPOSAT-NN	+					
	-					
ADV-ADV	+					
	-					
ADV-APPR	+					
	-					
PDAT-NN	+					
	-					

Abb. 6 Hier wird schematisch Über- und Mindergebrauch von Wortartbigrammen angezeigt. Die Wortarttags sind aus dem Stuttgart-Tübingen-Tagset Dabei wird in der jeweiligen ‚+'-Zeile Übergebrauch (Lerner benutzen das Wort im Vgl. zu L1-Sprechern zu häufig) angezeigt und in der jeweiligen ‚-'-Zeile Mindergebrauch (Lerner benutzen das Wort im Vgl. zu Muttersprachlern zu selten). Nur statistisch signifikante Werte werden angezeigt. Die Intensität der Färbung zeigt den Grad des Über- oder Mindergebrauchs an (je dunkler, desto stärker). In den Spaltenüberschriften stehen die Muttersprachen der Lerner.

Die Mindergebrauchsdaten, die automatisch aus den Korpora ermittelt werden, werden hier zunächst als Diagnostik für weiter zu untersuchende Kategorien verwendet; die Daten müssen dann natürlich qualitativ weiter analysiert werden. So konnten Zeldes, Lüdeling und Hirschmann (2008) bspw. zeigen, dass syntaktische Variabilität mit Mindergebrauch korreliert.

4 Methodik: Grundbegriffe der Korpuslinguistik

Nachdem wir einige Beispiele dafür gesehen haben, wie Korpora in der Sprachvermittlung und in der Spracherwerbsforschung eingesetzt werden können, möchten wir nun systematisch auf Korpusdesign und Korpusvorverarbeitung eingehen.

4.1 Korpusdesign

4.1.1 Parameter

Durch die jeweils vorliegende Forschungsfrage oder den Verwendungszweck werden die Parameter festgelegt, die die Textauswahl für ein Korpus beeinflussen. Dabei ist es nur sehr selten so, dass ein bestimmter Text verlangt wird; normalerweise wird durch die Parameter eine Menge an ‚repräsentativen' Texten bestimmt und die Annahme ist, dass die sprachlichen Eigenschaften vergleichbar sind. Wir kommen unten auf das Konzept ‚Repräsentativität' zurück.

Für Lernerkorpora bspw. sind unter anderem die Parameter in Abb. 7 relevant. Die Parameter und Werte sind nicht immer einfach zu bestimmen. So ist zum Beispiel die Unterscheidung zwischen gesprochener und geschriebener Sprache nur medial eindeutig; konzeptionell ist sie problematisch (so sind politische Reden oft geschrieben, um gesprochen zu werden, siehe dazu Koch und Österreicher 1985, Hunston 2008, Wichmann 2008 und viele andere). Die Werte für Register, Genre oder Texttyp/Textsorte/Register sind ebenso umstritten und schwer zu definieren (Biber und Conrad 2009).

Es gibt Korpora, die bezüglich der festgelegten Parameter relativ homogen sind (z. B. Lernerkorpora mit Texten einer Aufgabe von Lernern eines bestimmten Lernstands und einer gemeinsamen L1). Solche Korpora beantworten Fragen nach den sprachlichen Eigenschaften der betrachteten Varietät. Viele Forschungsfragen verlangen aber den Vergleich unterschiedlicher Varietäten (ein Beispiel haben wir in Abschnitt 3.2 gesehen). Idealerweise sollten Korpora, die zum Vergleich genutzt werden, sich nur in dem zu betrachtenden Parameter unterscheiden, so dass sprachliche Unterschiede direkt mit diesem Parameter korreliert werden können.

Auch gibt es Korpora, die unterschiedliche Varietäten in vorher festgelegten Anteilen enthalten (dazu gehören sogenannte Referenzkorpora wie DeReKo oder auch die Textbasis für das Digitale Wörterbuch der deutschen Sprache).

Parameter	Mögliche Werte und Beschreibung	Referenzen
Ursprungs-sprache der Text-produzenten	L1-Korpora, Lernerkorpora (mit einer oder mehreren L1)	Maden-Weinberger (2008) Walter und Schmidt (2008)
Medium	gesprochene Sprache vs. geschriebene Sprache	Möllering (2004) Gut (2008)
Genre, Texttyp, Register	Zeitungstexte, Romane, wissenschaft-liche Texte, Privatbriefe, Gespräche, …	Byrnes und Sinicrope (2008)
Erhebungs-zeitraum	‚synchron' (historisch, aktuell), für Querschnittsuntersuchungen ‚diachron', Longitudinalstudien	Skiba, Bressem und Dittmar (2008) Ahrenholz (2008), Doolittle (2008)
Korpussprache	eine L1, mehrere L1	
…	…	

Abb. 7: Mögliche Parameter für das Korpusdesign eines Lernerkorpus mit beispielhaften Referenzen zu Lernerkorpora für DaF/DaZ

4.1.2 Repräsentativität und Größe

Der Begriff ‚Repräsentativität' wird in der Literatur oft unklar (und falsch) verwendet. Der Begriff stammt aus der Statistik und beschreibt eine Strategie der Stichprobenerstellung. Dabei kann ein Korpus immer als eine Stichprobe einer größeren Grundgesamtheit verstanden werden; idealerweise sollen die in einem gegebenen Korpus beobachteten Eigenschaften auf die Grundgesamtheit generalisiert werden. Repräsentativität kann qualitativ so verstanden werden, dass ein Text repräsentativ für eine bestimmte Kombination von Parametern steht. Wenn man also z. B. die sprachlichen Eigenschaften von Sportnachrichten aus einer mittelgroßen deutschen Tageszeitung untersuchen möchte, sollte es egal sein, welche Artikel genau betrachtet werden. Dass das problematisch sein kann, ist offensichtlich (siehe dazu zum Beispiel Biber und Conrad 2009, wo gezeigt wird, dass selbst in scheinbar ‚homogenen' Regis-

http://zope.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/anke_veroeffentlichungen, 04.05.2009

tern wie wissenschaftlichen Aufsätzen in einem Fach signifikante Unterschiede zwischen Textabschnitten bestehen können, oder Evert 2006).

Noch deutlich problematischer ist es, wenn von der Stichprobe auf eine Grundgesamtheit geschlossen wird. So wird manchmal implizit angenommen, dass Referenzkorpora für ‚die Sprache' stehen (Biber 1993). Da aber die Zusammensetzung der Grundgesamtheit für eine sprachliche Varietät fast nie bekannt ist (Ausnahmen zu dieser Aussage sind nur abgeschlossene Grundgesamtheiten wie z. B. alle Romane von Goethe), kann keine im statistischen Sinne repräsentative Stichprobe gezogen werden. Verallgemeinerungen von Korpusbefunden sind daher immer schwierig.

Da sprachliche Phänomene unterschiedlich häufig und einige davon nicht gleich verteilt sind (siehe Zipf 1949, Baayen 2001, Evert 2006), gibt es keine ideale Korpusgröße. So können schon in relativ kleinen Korpora viele Vorkommen von definiten Artikeln gefunden werden, für bestimmte Wortbildungsmuster (z. B. das nichtmedizinische *–itis* wie in *Telefonitis*) hingegen reicht vielleicht ein Korpus mit 1 Milliarde Token nicht aus.

4.1.3 Authentizität und Verfügbarkeit

Wir möchten hier noch auf zwei Punkte eingehen, die oft als problematisch gesehen werden, die Authentizität und die Verfügbarkeit.

Das Thema Authentizität ist vor allem im Bezug auf Lernerkorpora viel diskutiert worden (aber auch allgemein Seidlhofer 2003, 77-123). Um eine gegebene Interimsgrammatik zu untersuchen, sollten möglichst ‚authentische' Daten der betrachteten Lerner erhoben werden. Aber was ist hier ‚authentisch'? Sprachliches Verhalten wird in der Regel bezogen auf einen außersprachlichen Kontext beschrieben. Lernende agieren im Unterricht permanent unter Beobachtung, gerade wenn es sich um zu bewertende sprachliche Leistungen handelt. Es ist bekannt, dass Menschen, die beobachtet werden, sich anders verhalten als Menschen, die sich unbeobachtet glauben (zum Beobachtereffekt vgl. z. B. Edmondson und House 2000: 36-37 sowie Ellis und Barkhuizen 2005: 11-50). In der Zweitspracherwerbsforschung konzentrierte man sich deshalb in den letzten Dekaden vorrangig auf den ungesteuerten Spracherwerb (vgl. Walter und Grommes 2008: 3-27). Authentizität in der Korpuslinguistik meint, dass die sozialen Rahmenbedingungen der Äußerungsproduktion identisch mit dem zu erfassenden Kontext sind, mit anderen Worten

es sich um einen natürlichen Kontext handelt. Dies impliziert insbesondere, dass die sprachlichen Äußerungen nicht für das Korpus „geschaffen“ wurden. Lernerdaten sind mit dem hier skizzierten Prinzip der Authentizität oft schwierig zu akquirieren. Das gilt auch für gesprochene Sprache und andere ’spezielle’ Korpora. Wie in Abschnitt 3.1 gezeigt, ist die Fehlerannotation – genau wie jede andere Annotation, s. u.- eine Interpretation der Daten, dies gilt für Korpora genauso wie für andere Daten. Daher sind Forschungsergebnisse ohne die dazugehörigen Daten und die Analysekategorien nicht nachvollziehbar. Wir plädieren deswegen dafür, dass Korpusdaten inklusive aller Annotationsebenen frei zur Verfügung gestellt werden, soweit dies rechtlich und ethisch möglich ist (siehe dazu z. B. Lehmberg et al. 2007).

4.2 Korpusvorverarbeitung

Korpusdaten werden oft mit Annotationen angereichert. In Abschnitt 3.1 haben wir bereits eine Sorte von Annotation, nämlich die Fehlerannotation, besprochen, in Abschnitt 3.2 haben wir gesehen, wie Wortartannotation in Mindergebrauchsstudien eingesetzt werden kann. In diesem Abschnitt betrachten wir Annotation systematischer. Annotationen sind für die Suche und Subkorpusbildung hilfreich. Man unterscheidet zwischen Metainformationen über die betrachteten Texte (Headerinformationen), tokenbasierten Annotationen (positionellen Annotationen) und strukturellen Annotationen.

Headerinformationen enthalten Informationen über jeden in einem Korpus enthaltenen Text. Welche Headerinformationen erhoben und angegeben werden, hängt wieder von der Forschungsfrage ab. Bei Lernerkorpora sind hier zum Beispiel Informationen über die L1, die Lernbiographie der jeweiligen Textproduzenten, sowie Angaben zur Aufgabenstellung, Textsorte oder zu den verwendeten Hilfsmitteln sinnvoll. Wenn systematisch Headerinformationen angegeben sind, können bei der Auswertung *ad hoc* Subkorpora erstellt werden (zum Beispiel alle Essays von Lernenden mit der L1 Spanisch).

Daneben können einzelne Token (kleinste Einheiten, oft (aber nicht notwendigerweise) so etwas wie graphemische Wörter) oder Spannen von Token annotiert werden. Für L1-Korpora des Deutschen gibt es frei verfügbare Werkzeuge für die automatische Annotation von Wortart und Lemma. Die Werkzeuge arbeiten fast alle mit einer Kombination aus Lexikonzugriff und statistischen Verfahren, die auf handannotierten Trainingskorpora trainiert werden. Die Qualität der Ergebnisse ist daher immer abhängig von den ge-

nutzten Lexika, der Qualität der Trainingskorpora und der statistischen Ähnlichkeit zwischen Trainingskorpus und zu annotierendem Text (vgl. Lemnitzer und Zinsmeister 2006; Schmid 2008). Inzwischen gibt es auch einige Korpora, die phonetisch und phonologisch, syntaktisch (mit Bäumen oder Graphen), mit Koreferenzinformationen oder informationsstrukturell annotiert sind.

Allgemein ist zu sagen, dass Annotationen für viele Auswertungsfragen unverzichtbar sind, dass aber *jede* Form von Annotation eine Interpretation der Daten bedeutet und daher für eine sinnvolle Auswertung sowohl die Tagsets als auch die Vergaberichtlinien unbedingt angegeben werden müssen. Man kann Annotationen in unterschiedlichen Formaten speichern; in den letzten Jahren haben sich XML als Datenformat und die Standards der Text Encoding Initiative (TEI Guidelines, http://www.tei-c.org/index.xml) durchgesetzt. Es ist sinnvoll, diesen Standards zu folgen, da viele Verarbeitungs- und Suchwerkzeuge mit solchen Standards arbeiten und nur weit verbreitete Formate die Nachhaltigkeit der Daten sichern.

Neben einer inhaltlichen Festlegung auf Annotationsebenen und Tagsets/Richtlinien spielt für eine sinnvolle Annotation auch die Korpusarchitektur eine Rolle. Lange Zeit wurden Annotationen inline (also innerhalb der Textdaten) annotiert. In den letzten Jahren ist man dazu übergegangen, Korpora (zumindest kleine Korpora) in mehreren Ebenen (multilevel) zu annotieren, im Idealfall so, dass die Daten getrennt von den Annotationen gespeichert sind (standoff). Das hat die Vorteile, dass unterschiedliche Modi (gesprochene Sprache, Textebenen, Annotationen etc.) im selben Korpus repräsentiert werden können und die Annotationsebenen sich nicht gegenseitig beeinflussen, so dass jederzeit neue Annotationsebenen hinzugefügt werden können (Carletta et al. 2005, Lüdeling et al. 2005, Stede 2007, Wittenburg 2008).

5 Zusammenfassung

Dieser Artikel sollte einen Überblick darüber geben, wie Korpora in der Sprachvermittlung eingesetzt werden und welche Rolle sie bei der Erforschung von Lernersprache spielen können. Alle qualitativen und quantitativen Methoden aus der Korpuslinguistik können in beiden Bereichen sinnvoll eingesetzt werden. Wir sind daher neben vermittlungs- und erwerbsspezifischen

Themen auch kurz allgemeine Bemerkungen zu Korpusdesign, Annotation und Auswertung eingegangen.

Noch ein Wort zum Status von Korpusdaten: Korpora sind immer endlich und bilden daher immer nur einen Ausschnitt aus ‚der Sprache' (was auch immer das sein mag) ab. Gerade durch die Korpuslinguistik ist deutlich geworden, welchen Einfluss Parameter wie Register, Modalität oder Autoreneigenschaften auf *alle* sprachlichen Ebenen haben. Daher ist wahrscheinlich die für eine Forschungsfrage jeweils geeignete Auswahl (oder Erstellung) eines Korpus die wichtigste Entscheidung, die getroffen werden muss.

Es gibt bereits viele L1-Korpora für das Deutsche. Bei L2-Korpora des Deutschen als Fremdsprache sieht es schlechter aus. Außerdem haben Lernerkorpusdaten – wie jeder andere Datentyp auch – Beschränkungen. Beispielsweise können bestimmte Strukturen nicht untersucht werden, weil sie schlicht nicht produziert werden (die Vermeidungsstrategien an sich sind natürlich auch interessant). Im Idealfall ergänzen sich daher Korpusdaten und experimentelle Daten. So können aus Korpusdaten Hypothesen über bestimmte Schwierigkeiten und Muster entwickelt werden, die dann experimentell überprüft werden.

Der Einsatz von Korpora im Unterricht bietet viele Möglichkeiten, die in den nächsten Jahren empirisch untersucht werden müssen. Mit dem Einsatz von Korpora gehen die Fokussierung auf den Sprachgebrauch und das Prinzip der Authentizität einher, die neue Perspektiven einer modernen Sprachvermittlung aufzeigen: Dies beginnt in der präzisen Beschreibung des Sprachgebrauchs und endet in der autonomen Handhabung dieses mächtigen Werkzeugs durch die Lernenden selbst. Um die Korpora effizient einzusetzen, ist es notwendig, dass Lehrende (und selbstverständlich auch Forschende) diese Ressourcen kennen und sich über die Möglichkeiten aber auch die Grenzen ihres Einsatzes bewusst werden. Die neue Generation von Lehrenden, die derzeit an den Universitäten ausgebildet werden, kann mit diesem Wissen in die Vermittlungspraxis einsteigen. Dazu ist es notwenig, ‚corpus literacy' zu vermitteln. Ansätze hierzu haben wir in unserem Artikel aufgezeigt.

6 Referenzen

Alle URLs wurden überprüft am 23.04.2009.

6.1 Links

Wir haben nicht alle URLs für deutsche Korpora aufgeführt. Für einen Überblick über verfügbare Korpora des Deutschen siehe Lemnitzer und Zinsmeister (2006) oder
http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links/.

Linksammlungen, die auf Sprachvermittlung und Spracherwerb spezialisiert sind (jedoch viel zum Englischen) sind:

- Bibliographie zu Lernerkorpora des Center for English Corpus Linguistics, Universität Louvain-la-Neuve
  http://cecl.fltr.ucl.ac.be/learner%20corpus%20bibliography.html
- Bibliographie zu Lernerkorpora von Yukio Tono, Meikai Universität,
  http://leo.meikai.ac.jp/~tono/

Konferenz

- Teaching and Language Corpora (Konferenzreihe, 2008 in Lissabon
  http://talc8.isla.pt/index.html, mit Verweis auf alle früheren TaLC-Konferenzen)

6.2 Angesprochene Korpora[3]

Akademisches Deutsch 2006 (Korpus mit Zusammenfassungen akademischer Dissertationen), 841483 Token http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/institutkorpora

Bundestagsreden (1996 – 2003), 36723139 Token http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/institutkorpora

Cobuild Projekt, Bank of English, http://www.collins.co.uk/Corpus/CorpusSearch.aspx

c't-Korpus, Zeitschrift c't (1998-2002), 14596537 Token http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/institutkorpora

DeReKo (Deutsches Referenzkorpus), IDS Mannheim, http://www.ids-mannheim.de/kl/projekte/dereko_I

EK-Korpus (Elizitierte Konfliktgespräche zwischen Müttern und jugendlichen Töchtern), 73007 Token, IDS Mannheim, http://agd.ids-mannheim.de/html/korpora/korpus-ek.shtml

Falko (Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache), 232564 Token, HU Berlin, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

Textbasis des Digitalen Wörterbuchs der deutschen Sprache des zwanzigsten Jahrhunderts, Berlin-Brandenburgische Akademie der Wissenschaften, http://www.dwds.de

[3] Die Institutskorpora des Instituts für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin sind für die Forschung frei zugänglich. Für einige Korpora ist ein (kostenloser) Accountantrag erforderlich, den Sie unter http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/institutkorpora finden.

6.3 Literatur

Aguado, Karin (2002) Formelhafte Sequenzen und ihre Funktionen für den L2-Erwerb. *Zeitschrift für Angewandte Linguistik* 37: 27–49.

Ahrenholz, Bernt (2008) Zum Erwerb zentraler Wortstellungsmuster. In: Bernt Ahrenholz, Ursula Bredel, Wolfgang Klein, Martina Rost-Roth und Romuald Skiba (Hg.), 165-177.

Ahrenholz, Bernt, Ursula Bredel, Wolfgang Klein, Martina Rost-Roth und Romuald Skiba (Hg.) (2009) *Empirische Forschung und Theoriebildung. Beiträge aus der Soziolinguistik, Gesprochene-Sprache- und Zweitspracherwerbsforschung*. Berlin u.a.: Peter Lang.

Altenberg, Bengt und Mats Eeg-Olofsson (1990) Phraseology in Spoken English. In: Jan Aarts und Willem Meijs (Hg.) *Theory and Practice in Corpus Linguistics*, 1-26. Amsterdam: Rodopi.

Aston, Guy (2002) The learner as corpus designer. In: Bernhard Kettemann und Georg Marko (Hg.), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000*, 9-25. Amsterdam: Rodopi.

Baayen, R. Harald (2001) *Word Frequency distributions.* Dordrecht: Kluwer.

Barlow, Michael (2004) Software for corpus access and analysis. In:
John Sinclair (Hg.), *How to Use Corpora in Language Teaching*, 225-250. Amsterdam/Philadelphia: John Benjamins.

Baroni, Marco und Silvia Bernardini (2006) A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21: 259-274.

Bernardini, Silvia (2004) Corpora in the classroom: An overview and some reflections on future developments. In: John Sinclair (Hg.), *How to use corpora in language teaching*, 15-36. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas (1993) Representativeness in Corpus Design. *Literary and Linguistic Computing* 8: 243-257.

Biber, Douglas und Susan Conrad (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad und Viviana Cortes (2004) *If you look at...:* Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25: 371-405.

Biber, Douglas, Susan Conrad und Randi Reppen (1998) *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd und Marie Helt (2002) Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly* 36: 9-48.

Bley-Vroman, Robert (1983) The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity. *Language Learning. A Journal of Applied Linguistics* 33: 1-17.

Bowker, Lynne und Jennifer Pearson (2002) *Working with Specialized Language - A Practical Guide to Using Corpora*. London: Routledge.

Byrnes, Heidi und Castle Sinicrope (2008) Advancedness and the Development of Relativization in L2 German: A Curriculum-based Longitudinal Study. In: Lourdes Ortega und Heidi Byrnes (Hg.), 109-138.

Carletta, Jean (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22: 249–254.

Carletta, Jean, Stefan Evert, Ulrich Heid und Jonathan Kilgour (2005) The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal* 39: 313-334.

Choueka, Yaacov (1988) Looking for Needles in a Haystack. In: *Proceedings of RIAO '88*. Cambridge, MA, 609-623.

Conrad, Susan (2008) Myth 6: Corpus-based research is too complicated to be useful for writing teachers. In: Joy M. Reid (Hg.): *Writing Myths: Applying Second Language Research to Classroom Teaching.* Ann Arbor, Michigan: University of Michigan Press, 115-139.

Corder, Stephen Pit (1981) *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

Coxhead, Averil (2000) A new academic word list. *TESOL Quarterly* 34, 213-238.

Dagneaux, Estelle, Sharon Denness, Sylviane Granger und Fanny Meunier (1996) *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catolique de Louvain.

Doolittle, Seanna (2008) *Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner*. Magisterarbeit, Humboldt-Universität zu Berlin.

EAGLES (1996) *Preliminary recommendations on corpus typology. EAG-TCWG-CTYP/P*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
Online unter http://www.ilc.cnr.it/EAGLES96/corpintr/corpintr.html

Edelhoff, Christoph (Hg.) (1985) *Authentische Texte im Deutschunterricht.* München: Hueber.

Edmondson, Willis und Juliane House (2000) *Einführung in die Sprachlehrforschung.* 2. überarbeitete Auflage. Tübingen u.a.: Francke.

Ellis, Rod und Gary Barkhuizen (2005) *Analysing Learner Language*. Oxford u.a.: Oxford University Press.

Evert, Stefan (2006) How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54: 177-190.

Evert, Stefan (2009) Corpora and Collocations. In: Anke Lüdeling und Merja Kytö (Hg.), 1212 – 1248.

Fandrych, Christian und Erwin Tschirner (2007) Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. *Deutsch als Fremdsprache* 44: 195-204.

Frankenberg-Garcia, Ana (2005) Pedagogical uses of monolingual and parallel concordances. *ELT Journal* 59: 189-198.

Gaskell, Delian und Thomas Cobb (2004) Can learners use concordance feedback for writing errors? In: *System* 32: 301–319.

Gellerstam, Martin (1986) Translationese in Swedish Novels Translated from English. In: Lars Wollin und Hans Lindquist (Hg.) *Translation Studies in Scandinavia*, 88-95. Lund: CWK Gleerup.

Granger, Sylviane (2002) A bird's-eye view of learner corpus research. In: Sylviane Granger, Joseph Hung und Stephanie Petch-Tyson (Hg.), 3–33.

Granger, Sylviane (2008) Learner corpora. In: Anke Lüdeling und Merja Kytö (Hg.), 259-275.

Granger, Sylviane, Joseph Hung und Stephanie Petch-Tyson (Hg.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam/Philadelphia: John Benjamins.

Gut, Ulrike (2007) Sprachkorpora im Phonetikunterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 12. Online unter http://zif.spz.tu-darmstadt.de/jg-12-2/docs/Gut.pdf

Gut, Ulrike (2008) Phonology in advanced learners of German. In: Maik Walter und Patrick Grommes (Hg.), 189-207.

Handwerker, Brigitte und Karin Madlener, Karin (2009) *Chunks für DaF. Theoretischer Hintergrund und Prototyp einer multimodalen Lernumgebung (inklusive DVD).* Baltmannsweiler: Schneider Verlag Hohengehren.

Hunston, Susan (2008) Collection Strategies and Design Decisions. In: Anke Lüdeling und Merja Kytö (Hg.), 154-168.

Izumi, Emi, Kiyotaka Uchimoto und Hitoshi Isahara (2005) Error Annotation for Corpus of Japanese Learner English. In: *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*: 71-80.

Johansson, Stig (1998) On the Role of Corpora in Cross-Linguistic Research. In: Stig Johansson und Signe Oksefjell (Hg.): *Corpora and Cross-Linguistic Research*, 134-147. Amsterdam: Rodopi.

Johns, Tim (2000) Data-Driven Learning: The Perpetual Challenge. In: Kettemann, Bernhard und Georg Marko (Hg.), *Language and Computers, Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000*, 107-117. Rodopi: Amsterdam.

Johns, Tim und Philip King (Hg.) (1991) *Classroom Concordancing.* Birmingham University: English Language Research Journal 4.

Jones, Randall und Erwin Tschirner (2006) *Frequency dictionary of German: Core vocabulary for learners.* London: Routledge.

Jordens, Peter (1983) *Das deutsche Kasussystem im Fremdspracherwerb. Eine psycholinguistische und fehleranalytische Untersuchung zum interim-*

sprachlichen Kasusmarkierungssystem niederländisch- und englischsprachiger Deutschstudierender. Tübingen: Narr.

Kallmeyer, Werner und Gisela Zifonun (Hg.) (2007) *Sprachkorpora - Datenmengen und Erkenntnisfortschritt.* (Jahrbuch des Instituts für Deutsche Sprache 2006.) Berlin/New York: Walter de Gruyter.

Kilgarriff, Adam (2007) Googleology is Bad Science. *Computational Linguistics* 33: 147-151.

Kleppin, Karin (1997) *Fehler und Fehlerkorrektur.* Berlin u.a.: Langenscheidt.

Koch, Peter und Wulf Oesterreicher (1985) Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgebrauch. *Romanistisches Jahrbuch* 36: 15- 43.

Lemnitzer, Lothar und Heike Zinsmeister (2006) *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Leech, Geoffrey (2001) The role of frequency in ELT: New corpus evidence brings a re-appraisal. In: Hu Wenzhong (Hg.) *ELT in China 2001: Papers presented at the 3rd International Symposium on ELT in China*, 1-23. Beijing: Foreign Language Teaching and Research Press.

Lehmberg, Timm, Christian Chiarcos, Georg Rehm und Andreas Witt (2007) Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In: Georg Rehm, Andreas Witt und Lothar Lemnitzer (Hg.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, 93-102. Tübingen: Gunter Narr.

Lüdeling, Anke (2008) Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Walter, Maik und Patrick Grommes, (Hg.), 119-140.

Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt und Maik Walter (2008) Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 45: 67-73.

Lüdeling, Anke und Merja Kytö (Hg.) (2008) *Corpus Linguistics. An International Handbook.* Vol 1. Berlin/New York: Mouton de Gruyter.

Lüdeling, Anke und Merja Kytö (Hg.) (2009) *Corpus Linguistics. An International Handbook.* Vol 2. Berlin/New York: Mouton de Gruyter.

Lüdeling, Anke, Maik Walter, Emil Kroymann und Peter Adolphs (2005) Multi-level error annotation in learner corpora. In: *Proceedings of Corpus Linguistics 2005*, Birmingham.
Online unter http://www.corpus.bham.ac.uk/pclc/index.shtml.

Maden-Weinberger, Ursula (2008) Modality as Indicator of L2 Proficiency? A corpus-based investigation into advanced German interlanguage. In: Maik Walter und Patrick Grommes (Hg.), 141-164.

McEnery, Anthony, Zonghua Xiao und Yukio Tono (2006) *Corpus-based Language Studies: An Advanced Resource Book.* London: Routledge.

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

Meunier, Fanny (2002) The pedagogical value of native and learner corpora in EFL grammar teaching. In: Sylviane Granger, Joseph Hung und Stephanie Petch-Tyson (Hg.), 119-141.

Meyer, Charles (2008) Pre-electronic corpora. In: Anke Lüdeling und Merja Kytö (Hg.), 1-14.

Möllering, Martina (2004) *The Acquisition of German Modal Particles. A corpus-based approach.* Berlin u.a.: Peter Lang.

Mukherjee, Joybrato (2002) *Korpuslinguistik und Englischunterricht: Eine Einführung.* Berlin u.a.: Peter Lang.

Mukherjee, Joybrato (2008) *Anglistische Korpuslinguistik. Eine Einführung.* Berlin: Erich Schmidt Verlag.

Mukherjee, Joybrato und Jens-Martin Rohrbach (2006) Rethinking Applied Corpus Linguistics from a Language-pedagogical Perspective: New Departures in Learner Corpus Research. In: Kettemann, Bernhard und Georg Marko (Hg.), *Planning, Gluing, and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*, 205-232. Berlin u.a.: Peter Lang.

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

O'Keeffe, Anne, Michael McCarthy und Ronald Carter (2007) *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Olohan, Maeve (2004) *Introducing Corpora in Translation Studies.* London/New York: Routledge.

Ortega, Lourdes und Heidi Byrnes (Hg.) (2008) *The Longitudinal Development of Advanced L2 Capacities.* New York: Routledge.

Römer, Ute (2005) *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics.* Amsterdam/Philadelphia: John Benjamins.

Römer, Ute (2006) Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik* 54: 121-134.

Römer, Ute (2008) Corpora and language teaching. In: Lüdeling, Anke und Merja Kytö (Hg.), 112-131.

Rothenhöfer, Andreas (erscheint) New developments in learner's dictionaries II: German. In: Rufus Gouws, Ulrich Heid, Wolfgang Schweickard und Herbert Ernst Wiegand (Hg.), *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie. IV. Ergänzungsband / Supplementary volume.* Berlin/New York: Mouton de Gruyter.

Rug, Wolfgang und Andreas Tomaszewski (2001) *Grammatik mit Sinn und Verstand: Übungsgrammatik Mittel- und Oberstufe.* Neufassung. Stuttgart: Ernst Klett.

Salem, André (1987) *Pratique des segments répétés*. Paris: Institut National de la Langue Française.

Scherer, Carmen (2006) *Korpuslinguistik.* Heidelberg: Winter.

Schmid, Helmut (2008) Tokenizing and part-of-speech tagging. In: Anke Lüdeling und Merja Kytö, (Hg.), 527-551.

Seidlhofer, Barbara (2003) *Controversies in Applied Linguistics.* Oxford: Oxford University Press.

Selinker, Larry (1972) Interlanguage. *IRAL* 10: 209-231.

Skiba, Romuald (2008) Korpora in der Zweitspracherwerbsforschung. Internetzugang zu Daten des ungesteuerten Zweitspracherwerbs. In: Bernt Ahrenholz, Ursula Bredel, Wolfgang Klein, Martina Rost-Roth und Romuald Skiba (Hg.), 21-30.

Skiba, Romuald, Jana Bressem und Norbert Dittmar (2008) Planning, collecting, exploring and archiving longitudinal data of naturalistic L2 acquisition: The contribution of the Berlin P-MoLL project. In: Lourdes Ortega und Heidi Byrnes (Hg.), 73-88.

Stede, Manfred (2007) *Korpusgestützte Textanalyse: Grundzüge der Ebenenorientierten Textlinguistik.* Tübingen: Narr.

Tenfjord, Kari, Jon Erik Hagen und Hilde Johansen (2006) The «Hows» and the «Whys» of Coding Categories in a Learner Corpus (or «How and Why an Error-Tagged Learner Corpus is not 'ipso facto' One Big Comparative Fallacy»). *Rivista di psicolinguistica applicata* 3: 93-108.

Tono, Yukio und Megumi Aoki (1998) Developing the optimal learning list of irregular verbs based on the native and learner corpora. In: Sylviane Granger und Joseph Hung (Hg.), *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, 14-16 December, 1998*, 113-118. The Chinese University of Hong Kong: Symposium Proceedings.

Tschirner, Erwin (2005) Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Antje Heine, Mathilde Hennig und Erwin Tschirner (Hg.): *Deutsch als Fremdsprache – Konturen und Perspektiven eines Fachs*, 133-149. München: Iudicium.

Tschirner, Erwin (2008) Das professionelle Wortschatzminimum im Deutschen als Fremdsprache. *Deutsch als Fremdsprache* 45: 195–208.

Walter, Maik und Patrick Grommes (Hg.) (2008) *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung.* Tübingen: Niemeyer.

Walter, Maik und Karin Schmidt (2008) "Und das ist auch gut so". Der Gebrauch des satzinitialen 'und' bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. In: Bernt Ahrenholz, Ursula Bredel, Wolfgang Klein, Martina Rost-Roth und Romuald Skiba (Hg.), 331-342.

Wichmann, Anne (2008) Speech Corpora and Spoken Corpora. In: Anke Lüdeling und Merja Kytö (Hg.), 187-207.

Wiechmann, Daniel und Stefan Fuhs (2006) Concordancing Software. *Corpus Linguistics and Lingustic Theory* 2: 107–127.

Wittenburg, Peter (2008) Preprocessing multimodal corpora. In: Anke Lüdeling und Merja Kytö (Hg.), 664-685.

Wynne, Martin (2008) Searching and Concordancing. In: Anke Lüdeling und Merja Kytö (Hg.), 706-737.

Year, JungEun (2004) Instances of the Comparative Fallacy. *Columbia University Working Papers in TESOL & Applied Linguistics* 4. Online unter http://www.tc.columbia.edu/academic/tesol/WJFiles/pdf/JungEun2004.pdf

Zeldes, Amir, Anke Lüdeling und Hagen Hirschmann (2008) What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data. Vortrag auf *Quantitative Investigations in Theoretical Linguistics 3 (QITL-3).* Helsinki, Finnland, 2.-4. Juni 2008.
Online unter
http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes_et_al.pdf

Zipf, George Kingsley (1949) *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* Cambridge, MA: Addison-Wesley.