Lernerkorpora:
Ressourcen für die Deutsch-als-Fremdsprache-Forschung

Karin Schmidt[1]

Abstract

The article addresses the growing importance of corpus-based research in the field of German foreign language acquisition. German corpora in general and learner corpora in particular are briefly introduced. A short overview of existing German learner corpora is followed by a detailed description of the error-annotated learner corpus Falko, a learner corpus of advanced learner German, which is accessible via internet (without any prior registration) and free of charge. Finally, a short example analysis demonstrates some of the functionalities of Falko. The aim of the article is to encourage researchers to employ corpora as helpful tools in their own work.

1. Einleitung

Seit der Etablierung der Korpuslinguistik, die vor allem durch die technologischen Fortschritte in der Hard-, Software-, und Netzwerktechnik sowie der Sprachtechnologie in den letzten Jahrzehnten möglich wurde, sind Korpora zu wichtigen und unentbehrlichen Hilfsmitteln in der linguistischen Forschung geworden. Die Einführung eines Lehrstuhls für Korpuslinguistik an der Humboldt-Universität (2002 Juniorprofessur, 2008 Professur) und das Erscheinen der ersten beiden deutschsprachigen Einführungen in die Korpuslinguistik (Lemnitzer/Zinsmeister 2006; Scherer 2006) dokumentieren die wachsende Bedeutung korpusbasierter Ansätze in der Linguistik allgemein. Im deutschsprachigen Raum wird die Bedeutung der Korpuslinguistik für die Zweit- und Fremdsprachenerwerbsforschung seit einigen Jahren diskutiert (vgl. Fandrych/Tschirner 2007, Lüdeling et al. 2008, Walter/Grommes 2008, Lüdeling/Walter 2009). Lernerkorpora kommt dabei für die empirische Erforschung des Zweit- bzw. Fremdsprachenerwerbs eine besondere Rolle zu, wie dieser Artikel exemplarisch zeigen möchte.

Die Struktur des Artikels ist wie folgt: In den ersten beiden Abschnitten werden die zentralen Begriffe „Korpus“ (Korpuslinguistik) und „Lernerkorpus“

[1] Dokuz Eylül Üniversitesi. Buca Eğitim Fakültesi. Alman Dili Bölümü

umrissen; im folgenden Punkt wird ein Überblick über bisher existierende Lernerkorpora des Deutschen als Zweit- und Fremdsprache gegeben; anschließend wird das Lernerkorpus „Falko“ vorgestellt; im letzten Punkt werden anhand einer einfachen Beispielanalyse die Funktionalitäten der Suchabfragen in einigen Subkorpora illustriert.

2. Korpora und Korpuslinguistik

Korpus meint in seiner Grundbedeutung nichts anderes als „a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language“ (EAGLES 1996)[2]. Korpora sind damit keine wahllosen, sondern zielgerichtete und bestimmten Kriterien folgende Zusammenstellungen von Sprachbeispielen oder Texten (mündlich oder schriftlich). Damit unterscheiden sich Korpora noch nicht unbedingt von schriftlichen Textsammlungen, die seit jeher Grundlage linguistischer Forschung waren. In einer Präzisierung muss also ergänzt werden, dass in der Regel – und so auch in diesem Artikel – der Begriff Korpus als gleichbedeutend mit einem Computerkorpus verstanden werden soll: „Computer corpus: a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks“[3] (EAGLES 1996). Gerade die elektronische und digitale Verfügbarkeit ermöglicht einerseits eine rasche Analyse und andererseits auch die (wiederholte) Analyse größerer Datenmengen und damit sprachlich seltener auftretender Phänomene (man denke z.B. an bestimmte, seltene Konnektoren wie *obgleich* oder Modalpartikeln wie *bloß)*. Je größer die enthaltenen Textmengen in den Korpora sind, desto besser sind im Allgemeinen statistische Verfahren anwendbar, um Hypothesen überprüfen zu können.

Das Institut für Deutsche Sprache in Mannheim ist die zentrale Adresse für Korpora der deutschen Gegenwartssprache (www.ids-mannheim.de), die dort seit Mitte der sechziger Jahre erstellt und archiviert werden. Knapp 40 Korpora mit mehr als 2,5 Milliarden laufenden Textwörtern zur Erforschung synchronen (und begrenzt diachronen) Sprachgebrauchs sind dort öffentlich für Forschungszwecke zugänglich.

[2] „eine Sammlung von Sprachstücken (Sprachbeispielen), die nach expliziten linguistischen Kriterien ausgewählt und angeordnet sind, um als Beispiel einer Sprache genutzt zu werden“ (eigene Übertragung).

[3] „Computer-Korpus: ein Korpus, das in einer standardisierten und gleichförmigen (einheitlichen, homogenen) Art und Weise für unbefristet mögliche Abfragen kodiert wurde“ (eigene Übertragung).

Die Korpuslinguistik beschäftigt sich mit dem Aufbau, der Auszeichnung und der Auswertung von Korpora (Lüdeling/Walter 2009: 1).[4] Die Auszeichnung von Korpora durch Annotationen, d.h. zusätzlichen Merkmalen, wie z.B. die Wortartenzuweisung, ist hierbei von besonderem Interesse und Vorteil für Suchabfragen und linguistische Analysen. Suchabfragen können sich so nicht nur auf das Sprachmaterial selbst (auf Wörter), sondern auch auf die durch die Annotation zugewiesenen Merkmale beziehen (also z.B. bei einer Wortartenzuweisung auf alle Wörter einer Wortart, wie Modalverben oder Artikel).

Zwei beliebte Funktionalitäten von Korpora mit Relevanz nicht nur für die Spracherwerbsforschung, sondern auch für die Sprachvermittlung (die im Artikel nur gestreift werden kann) seien kurz demonstriert: a) Häufigkeiten – also z.B. **Wort(form)frequenzen** - lassen sich in Korpora gut und schnell ermitteln, man vergleiche z.B. das „Wort des Tages" (http://wortschatz.uni-leipzig.de/wort-des-tages/2009/06/21/), das tagesaktuell die jeweils häufigsten Wörter in der deutschsprachigen Presse kompiliert (am 21.06. waren dies u.a. Altpapier, Moschee, Bundestagsmandat, Tränengas etc.). Eine Nutzung für die Sprachvermittlung illustriert der Langenscheidt-Service, der zu einer Auswahl dieser Wörter ebenfalls tagesaktuell die englischen Äquivalente anbietet (http://www.langenscheidt.de/service/wort_des_tages_1049.html).

b) **Kookkurrenzanalysen** können ebenfalls bequem anhand von Korpora durchgeführt werden, so kann man beispielsweise überprüfen, welches Wort am häufigsten in Verbindung mit einem bestimmten Wort auftaucht. Das Wort „es" beispielsweise steht am häufigsten in Verbindung mit „gibt", „gab" und „heißt"(http://corpora.ids-mannheim.de/ccdb/, Abfrage am 21.06.2009); der häufigste Partner des Wortes „friedlich" hingegen ist das Wort „Lösung" (Abfrage am 22.06.2009). Solche Kollokationen ermitteln zu können hat einen unbestreitbaren Wert für die Sprach- und speziell für die Wortschatzvermittlung (und natürlich auch für die Lexikographie)[5].

c) **Konkordanzen** (auch als kwic-Liste bezeichnet = keyword in context) ermöglichen es, für ein beliebiges Suchwort (oder für eine Kette von Suchwörtern) alle Vorkommen in einem zuvor definierten Kontext zu erstellen und in einem Listenformat übersichtlich zur Hand haben zu können.

[4] Der Frage, ob es sich bei der Korpuslinguistik um eine selbständige Teildisziplin der Linguistik handelt (Fandrych/Tschirner 2007: 195; Köhler 2005) oder vielmehr um eine linguistische Methode (Stefanowitsch 2005: 141, 147), soll an dieser Stelle nicht nachgegangen werden.

[5] Vgl. auch Miklitz 2002 zur Nutzung der IDS-Korpora und des „Wortschatz" der Universität Leipzig für DaF-Studierende und Lehrer; http://www.lernforum.uni-bonn.de/corpora.html.

Üblicherweise steht das Suchwort mittig mit den jeweiligen Kontexten rechts und links (Abfrage im Korpus Akademisches Deutsch 2006, am 26.06.2009)[6]:

Kontext	Wort	Kontext
. Ein möglicher Grund	**dafür**	könnte die Tatsache
Sie sprechen nämlich zusammengenommen	**dafür**	, daß der
liefern alle Ergebnisse Evidenz	**dafür**	, dass die
betrachten. Als Grund	**dafür**	kann genannt werden
sind einige weitere Gründe	**dafür**	dass die
. Die Untersuchungsergebnisse sprechen	**dafür**	, dass sich
ist damit als Evidenz	**dafür**	zu werten,
werden konfrontiert mit Indizien	**dafür**	, dass Hören
geworden, ohne die	**dafür**	erforderliche Kompetenz zu
dem Gläubiger die Beweislast	**dafür**	, daß der

Diese Form der Ergebnispräsentation ermöglicht es Lehrenden u.a. auch, das Formbewusstsein der Lernenden für typische Verwendungskontexte bestimmter Wörter oder Formen (ggfs. in bestimmten Textsorten) durch die Arbeit mit authentischen Beispielen zu schärfen (Schmidt, C. 2008: 78).

3. Lernerkorpora und Spracherwerbsforschung

Die Diskussion darüber, wie auch lernersprachliche Korpora und korpuslinguistische Herangehensweisen die Fremd- und Zweitsprachenerwerbsforschung des Deutschen erleichtern und unterstützen können, wurde durch einen Artikel von Fandrych/Tschirner 2007 angeregt und fortgeführt in der Zeitschrift „Deutsch als Fremdsprache“ 2008, Heft 2 (vgl. ibid Lüdeling et al.; Schmidt, C.). Der ebenfalls 2008 erschienene Sammelband von Walter/Grommes demonstriert in seinen Beiträgen die fruchtbare Verbindung, die Korpuslinguistik und Fremdsprachenerwerbsforschung gerade auch in der Erforschung fortgeschrittener Lernervarietäten eingehen können. Während der zweibändige HSK-Sammelband „Deutsch als Fremdsprache“ (19.1; 19.2) aus dem Jahr 2001 (Helbig et al.) noch keinen Artikel zur Relevanz von Korpuslinguistik für Deutsch als Fremdsprache beinhaltete, wird die Neuausgabe (hg. von Krumm et al.; voraussichtliches Erscheinungsdatum 2010) diesem Thema einen Grundlagenartikel widmen, der vorab in einer ausführlicheren und gut verständlichen Version online verfügbar ist (Lüdeling/Walter 2009). Auch die Aufnahme des Themas in den genannten Sammelband zeigt, dass Spracherwerbsforscher im Bereich Deutsch als

[6] Näheres zum Korpus unter https://korpling.german.hu-berlin.de/cqpwi/corpora.php (Registrierung erforderlich).

Fremdsprache sich zukünftig mit den Möglichkeiten und Grenzen von korpuslinguistischen Analysen generell und von Lernerkorpora speziell auseinandersetzen müssen.
Unter Lernerkorpora versteht man analog zu oben angeführter Korpusdefinition die Zusammenstellungen von Lernertexten, speziell von Fremdsprachenlernern (Nicht-Muttersprachlern) (Granger et al. 2002: 7):

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.[7]

Ausschlaggebend sind hier die Arbeiten von Sylviane Granger et al, die das erste, elektronisch verfügbare Lernerkorpus für die Zielsprache Englisch zusammengestellt haben (zur Korpusdarstellung vgl. http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm; eine fortlaufend aktualisierte Lernerkorpusbiographie, in der u.a. auch die einschlägigen Arbeiten von Granger et al. zusammengestellt sind, findet sich unter http://cecl.fltr.ucl.ac.be/learner%20corpus%20bibliography.html).
Das International Corpus of Learner English (inzwischen in einer zweiten Version erhältlich) enthält schriftliche Lernertexte (Essays) von Lernern des Englischen mit 16 verschiedenen Herkunftssprachen, darunter sowohl Türkisch als auch Deutsch (mit den beiden kooperierenden Hochschulen Çukurova Üniversitesi und Universität Augsburg). Insgesamt umfasst das ICLE-Korpus mehr als 3 Millionen Wortformen (ca. 200.000 Wortformen pro Sprache und Subkorpus).
Granger et al. zufolge sollen bei der Erstellung von Lernerkorpora bestimmte Design-Kriterien kontrolliert werden, damit die Daten für Analysezwecke nutzbar gemacht werden können: Dazu gehören u.a. die Ausgangssprachen, der Aufgabentyp, das Genre (oder allgemeiner gesprochen die Textsorte), der Lernerkontext (gesteuert vs. ungesteuert) und das Stadium (Sprachniveau). Idealiter soll ein parallel strukturiertes Korpus mit Texten muttersprachlicher Sprecher/Schreiber verfügbar sein. Nur so sind Aussagen validierbar, die Phänomene als lernersprachlich bedingt interpretieren. Zentral für Lernerkorpora ist ferner, dass die biographischen Daten der Lerner und die Umstände der Erhebung möglichst genau erfasst werden. Aussagen über

[7] FL steht für „Foreign Language"; SL für „Second Language; SLA für „Second Language Acquistion" (Zweitspracherwerb), FLT für Foreign Language Teaching (Fremdsprachenvermittlung).

Einflüsse der Muttersprache oder gelernter Fremdsprachen sind natürlich nur dann überhaupt belegbar, wenn diese lernbiographischen Metadaten zusätzlich zu den Sprachdaten erhoben wurden und auch erfragt und durchsucht werden können (vgl. zu lerner-/aufgabenbezogenen Kriterien Granger 2008: 263-65).

4. Lernerkorpora für Deutsch als Fremd- und Zweitsprache

Einige wichtige Korpora des Deutschen als Zweit- und Fremdsprache sollen im Folgenden vorgestellt werden. Die Zusammenstellung umfasst sowohl öffentlich zugängliche als auch nicht öffentlich zugängliche Korpora. Die öffentliche Zugänglichkeit von Korpora stellt dabei ein klares Desiderat dar, denn sie ist unerlässlich, um Analysen überhaupt nachvollziehbar und überprüfbar machen zu können (Lüdeling/Walter 2009: 15).

Im Bereich des ungesteuerten Spracherwerb des Deutschen ist das European Science Foundation-Projekt „Second Language Acquistion of Adult Immigrants“ (kurz **ESF-Projekt**) einflussreich. Im Rahmen der groß angelegten Studie wurden in den achtziger Jahren Longitudinaldaten von Lernern mit 6 Ausgangssprachen (Punjabi, Italienisch, Türkisch, Arabisch, Spanisch, Finnisch) und 5 verschiedenen Zielsprachen (Englisch, Deutsch, Niederländisch, Französisch, Schwedisch) zusammengestellt (Perdue 1993, Klein/Perdue 1992). Das Korpus ist – nach einer vorherigen Registrierung – über das Max-Planck-Institut Nijmegen durchsuchbar (vgl. http://www.mpi.nl/resources/data/browsable-corpora-at-mpi und http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI1%23).

Die im Folgenden genannten DaZ-Korpora beinhalten Daten türkischsprachiger Sprecher und sind ebenfalls über das Max-Planck-Institut zugänglich: Das **Augsburger DaZ-Korpus** (Heide Wegener) enthält Daten von 12 Kindern typologisch verschiedener Ausgangssprachen (L1 Türkisch, Russisch, Polnisch), die in Deutschland Deutsch als L2 erwerben. Das **Dimroth-Korpus** (Christine Dimroth) enthält L2-Daten von Erwachsenen mit den Ausgangssprachen Türkisch, Russisch und Kroatisch (sowie ein L1-Kontrollkorpus). Darüber hinaus sind über das Max Planck-Institut weitere DaZ-Korpora (vorwiegend mit Sprechern slawischer Ausgangssprachen wie Polnisch, Tschechisch und Russisch) zugänglich.[8]

Zusammenfassend kann festgehalten werden, dass Korpora im Bereich Deutsch als Zweitsprache den ungesteuerten Spracherwerb im Zielsprachenland

[8] Vgl. zu einer etwas ausführlicheren Darstellung der MPI-Korpora im Bereich DaZ vor allem Skiba 2008: 23-25).

fokussieren und in der Regel Longitudinaldaten nur weniger Sprecher beinhalten.

Im Bereich Deutsch als Fremdsprache, also dem gesteuerten Spracherwerb, kommt den Lernern mit der Ausgangssprache Englisch besondere Bedeutung zu. Für schriftliche Lernertexte sind zwei Korpora – die beide bisher nicht öffentlich zugänglich sind – zu nennen:

Das Telecollaborative Learner Corpus of English and German (**Telecorp**) der Penn State University wurde von Judy Belz in den Jahren 2000-2005 erstellt (insgesamt 6 Erhebungen). Es enthält Lernertexte der Textsorte E-Mail von ca. 200 Studierenden, sowohl Englisch lernenden Deutschen (Deutsch L1, Englisch L2) als auch Deutsch lernenden Amerikanern (Englisch L1, Deutsch L2). Das Korpus umfasst insgesamt 1,5 Millionen Token (vgl. Belz 2005).

Das Corpus of Learner German (**CLeG-Korpus**) wurde von Ursula Maden-Weinberge erstellt und enthält 451 Essays von 126 anglophonen Studierenden aus 3 Studienjahren, die Deutsch als Fremdsprache an der Universität Lancester studierten. Der Korpusumfang beträgt 198.301 Token (Maden-Weinberger 2002).

Im Bereich der gesprochenen Sprache ist das Korpus „Learning Prosody in a Foreign Language (**LeaP Corpus)** von Ulrike Gut zu nennen, das 359 annotierte Aufnahmen (mehr als 12 Stunden) von 131 Sprechern enthält. Die Daten umfassen Aufnahmen zum L2-Deutsch-Erwerb (Subkorpus Deutsch; 183 Aufnahmen) und auch zum L2-Englisch-Erwerb (Subkorpus Englisch; 176 Aufnahmen) sowie 18 Aufnahmen von L1-Sprechern (8 für L2 Englisch; 10 für L2 Deutsch). Die Aufnahmen variieren in der Länge zwischen 2 und 30 Minuten je nach Aufgabentyp. Das Korpus ist nach Anmeldung für Forschungszwecke zugänglich und eignet sich vor allem für Fragestellungen zu prosodischen Phänomenen bei L2-Sprechern (http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/applied/Research/leap/).

Des weiteren soll auch auf das leider noch nicht öffentlich zugängliche Lernerkorpus **Varcom** hingewiesen werden[9], das an der Universität Barcelona erstellt wurde (Marta Fernandez-Villanueva, Oliver Strunk, vgl. auch http://www.ub.edu/lada/f_projekte.htm). Das Korpus enthält Videoaufnahmen gesprochener Sprache in Deutsch, Spanisch/Katalanisch und umfasst 1520 Texte von 66 Studierenden (2009: 68). Das Korpus ist in zwei Subkorpora unterteilt; ein Lernerkorpus DaF von Sprechern mit Spanisch/Katalanisch als L1/L2 sowie ein Referenzkorpus mit Deutsch als L1 sowie Spanisch als Fremdsprache. Jeder Informant produzierte dialogische und monologische

[9] Es ist jedoch beabsichtigt, das Korpus auch anderen Forschern ab 2009 zur Verfügung zu stellen (Fernandez-Villanueva/Strunk 2009: 72).

Texte (Narration, Deskription, Instruktion, Argumentation, Exposition). Die Daten sind mit weiterführenden Annotationsebenen versehen, z.B. für Gestik, Wortarten, Interaktion. Es handelt sich bei Varkom um ein für vielseitige Forschungs- und Lehrzwecke einsetzbares Korpus von gesprochenen Lernerdeutsch, das hoffentlich bald auch anderen Forschern wie angekündigt zugänglich gemacht werden wird.
Zusammenfassend kann festgehalten werden, dass verschiedene Korpora im Bereich Deutsch als Zweitsprache (also zum ungesteuerten Spracherwerb im Zielsprachenland) mit Daten von erwachsenen Sprechern und mit Daten zum kindlichen Spracherwerb vor allem über das Max Planck-Institut in Nijmegen für Analysezwecke zugänglich sind. Für den gesteuerten DaF-Erwerb wurde bis dato jedoch nur eine geringe Anzahl meist kleiner Korpora erstellt, die in der Regel nicht öffentlich zugänglich sind. Dies unterscheidet die Situation des Deutschen als Fremdsprache von beispielsweise der Situation des Englischen als Fremdsprache, wo mit dem ICLE-Korpus ein breit angelegtes Korpus für Forschungszwecke zur Verfügung steht (wenn auch nur nach vorherigem Ankauf und nicht unentgeltlich). Zudem sind für den Bereich des Englischen als Fremdsprache bereits eine Reihe zusätzlicher Lernerkorpora entstanden (Pravec 2002). Vor diesem Hintergrund begann vor einigen Jahren die Entwicklung eines öffentlich frei und unentgeltlich zugänglichen, fehlerannotierten Lernerkorpus für Deutsch als Fremdsprache, das im Folgenden kurz vorgestellt werden soll[10].

5. Das Lernerkorpus „Falko“

5.1. Zusammensetzung

Seit dem Jahre 2004 entsteht in Kooperation zwischen dem Studiengebiet Deutsch als Fremdsprache an der Freien Universität und der Korpuslinguistik der Humboldt-Universität sowie unter Beteiligung der Georgetown University, Washington D.C. das frei zugängliche Lernerkorpus **F**ehler**a**nnotierte **L**erner**ko**rpus (**Falko**). Das Korpus ist ohne vorherige Registrierung unter online http://korpling.german.hu-berlin.de/falko/index.jsp verfügbar. Das Korpus enthält schriftliche Texte fortgeschrittener Lerner des Deutschen als Fremdsprache. Bei der Erstellung der beiden Kern-Korpora wurde auf die Einhaltung der folgenden Design-Kriterien für Lernerkorpora, so wie sie durch Granger et al. formuliert wurden, geachtet:

[10] Angesichts der Geschwindigkeit, mit der weltweit Lerner Korpora erstellt werden, kann natürlich jeder Versuch eines Überblicks nur als vorläufig betrachtet werden (vgl. Granger (im Druck): 261).

Tabelle 1: Design-Kriterien beim Aufbau des Falko-Korpus

	Falko Summary	**Falko Essay**
Aufgabe	(hand)schriftliche Prüfung	am PC
Textsorte	Zusammenfassung (eines germanistischen Fachtextes)	argumentatives Essay
Erhebungs-umstände	Zeitbegrenzung: 90 Min. keine zugelassenen Hilfsmittel (Wörterbücher etc.) unter Aufsicht	
Lernkontext	Deutsch als Fremdsprache	
Stadium	fortgeschritten: Kriterium DSH	fortgeschritten: Kriterium 60% C-Test
L1	typologisch verschieden	
Kontrolle	Vergleichskorpus mit nativen Schreibern	

Wichtig ist uns, dass „Fortgeschrittenheit" für Falko nicht über Lernjahre operationalisiert wurde (wie im ICLE-Korpus), sondern über ein formales Kriterium (DSH, C-Test) (vgl. zu Kriterien für Fortgeschrittenheit auch Walter/Grommes 2008; sowie Granger (im Druck: 265). Die Lerner des Falko-Summary-Korpus sind daher alle auf einem Sprachlevel von mindestens C1 gemäß Europäischem Referenzrahmen einzuordnen; die Lerner des Essay-Korpus hingegen mindestens auf einem Sprachlevel des Niveaus B1.
Drei der verschiedenen Subkorpora, die Falko umfasst, seien im Folgenden kurz skizziert:
Falko-Summary, das Zusammenfassungskorpus, enthält Textzusammenfassungen germanistischer Fachtexte (linguistisch oder literaturwissenschaftlich), die von ausländischen Germanistikstudierenden an der Freien Universität Berlin im Rahmen einer Sprachstandfeststellung erstellt wurden. Dieses Korpus ist abgeschlossen. Es gibt zu diesem Korpus ein Vergleichskorpus mit Texten muttersprachlich deutscher Germanistikstudierender (Falko-Summary L1) und ein getrenntes Subkorpus, in dem die Vorlagentexte, zu denen die Zusammenfassungen erfasst wurden, kompiliert sind (Falko-Source). Dieses Subkorpus ist hilfreich, wenn Leseverstehen oder auch Strategien der Textübernahme durch die Studierenden im Vordergrund stehen.
Falko-Essay, das Essaykorpus, enthält Essays, die von fortgeschrittenen Deutschlernern im Ausland (u. a. Adana, Türkei; Taschkent, Usbekistan; Kopenhagen, Dänemark; Turin, Italien, Nairobi, Kenia; Stellenbosch, Südafrika etc.) oder an Sommerkursen in Deutschland erhoben wurden. Die Essays wurden zu vorgegebenen Schreibimpulsen erstellt, die in Anlehnung an die

Themen des ICLE-Korpus übersetzt und übernommen wurden. Die vier Schreibimpulse, unter denen die Studierenden wählen konnten, lauten: Kriminalität zahlt sich nicht aus; Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt; Die meisten Universitätsabschlüsse sind nicht praxisorientiert und bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert; Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat. Das Essay-Korpus wird weiter ausgebaut.[11] Auch zu diesem Korpus liegt ein Vergleichskorpus vor (Falko Essay L1), dessen Daten an Berliner Gymnasien erhoben wurde.

Falko-GU ist ein Longitudinalkorpus, das an der Georgetown University in Washington D.C. zwischen 2001 und 2004 erhoben wurde. Es enthält schriftliche Texte amerikanischer Deutsch-Studierender in unterschiedlichen Studienjahren. Zu der Testsorte Buchrezension existiert ebenfalls ein L2-Kontrollkorpus. Die genauen Erhebungsumstände und die Zusammensetzung des GU-Korpus sowie der anderen Subkorpora sind auf der Falko-Homepage nachlesbar (http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko; sowie Lüdeling et al. 2008; Lüdeling et al. 2005; Siemen et al. 2006. Folgende Tabelle gibt abschließend die Korpusgröße der genannten drei L2-Korpora an (Token und Texte).

Tabelle 2: Text- und Tokenzahl in ausgewählten Falko-Subkorpora

	Falko-Summary L2[12]	**Falko-Essay L2**	**Falko-GU**
Token	41.075	91.112	78.132
Texte	107	187	92[13]

5.2. Korpusarchitektur und Annotation

Die Falko-Subkorpora sind auf verschiedenen Ebenen annotiert. Falko besitzt eine flexible Architektur, das bedeutet, dass jederzeit neue Ebenen eingefügt und getrennt bearbeitet werden können (multi-layer stand-off annotation).[14] Automatisch annotiert werden die Ebenen Lemma und Wortartenzuweisung

[11] Wenn Sie uns beim Ausbau von Falko-Essay unterstützen wollen, so nehmen Sie bitte mit uns Kontakt auf: falko-korpus@hu-berlin.de.

[12] Bezogen auf Version 1.1 bei Falko Summary und Falko Essay (zu beiden Korpora gibt es noch ältere Vorversionen).

[13] Von insgesamt 28 Lernern aus mindestens drei Lernjahren in Folge.

[14] Vergleiche zur Korpusarchitektur und zur Annotation vor allem Lüdeling 2008; Lüdeling et al. 2008, Lüdeling et al. 2005; Hirschmann et al. 2007.

(POS = Part of Speech). Der folgende Screenschot zeigt eine Lerneräußerung mit den Ebenen Lerneräußerung (word); Wortartenzuweisung (kpos), Lemma; Zielhypothese (target hypothesis; dazu folgend mehr) und Belegreferenz. Basis für die Wortartenzuweisung ist das Stuttgart-Tübingen-Tag-Set (1995/1999); kpos verweist darauf, dass die automatische Zuweisung manuell nachkorrigiert wurde.

Bildschirmfoto einer Lerneräußerung (Falko Summary L2):

word	Dabei	ist	es	zu	beachten	,
kpos	PAV	VAFIN	PPER	PTKZU	VVINF	$,
lemma	dabei	sein	es	zu	beachten	,
target_hypothesis	Dabei ist zu beachten					
ref	70	71	72	73	74	75

Besonderes Augenmerk ist auf die dritte Zeile, die „Zielhypothese", zu richten. In dieser wird eine zielsprachlich angemessene Rekonstruktion der Lerneräußerung versucht. Diese Reparatur oder Rekonstruktion der Lerneräußerung geschieht manuell. Als Voraussetzung für eine angestrebte Fehlerannotation ist sie von größter Bedeutung, da ein „Fehler" oder eine Abweichung stets in Hinblick auf eine implizite Zielversion geschieht, also im Hinblick darauf, was der Lehrende oder Forschende glaubt, habe der Lerner eigentlich sagen bzw. schreiben wollen. Ambivalente Lerneräußerungen erlauben aber oft verschiedene Rekonstruktionen, die unterschiedliche Fehlerannotationen zur Folge haben kann (vgl. ausführlich dazu Lüdeling 2008; Lüdeling/Walter 2009). Die Zielhypothese muss also expliziert werden, um die Fehlerannotation – die auf ihr basiert – nachvollziehbar machen zu können (Zielhypothesen liegen derzeit für Falko-Summary und Falko-GU vor). Daneben liegen eine syntaktische Felderdauszeichnung (vgl. Doolittle 2008) und eine Fehlermarkierung vor (ebenfalls für Falko-Summary und Falko-GU). Weitere Fehlerannotationen werden erarbeitet.

6. Beispielabfragen im Lernerkorpus „Falko"

Einleitend sei bemerkt, dass bereits eine Reihe von korpusbasierten Abschlussarbeiten und Artikeln auf der Grundlage von Falko entstanden sind (Auswahl: Walter/Schmidt 2008 zum satzinitialen „und"; Hirschmann 2005 zu Platzhalterphrasen, Kroymann 2008 zu Wortstellungsvarianz im Mittelfeld;

Lippert 2005 zur Definitheit von Nominalphrasen) und weitere verfasst werden (ausführlicher Überblick unter http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko). Arbeiten mit Lernerkorpora können qualitativ oder quantitativ erfolgen; kontrastive Arbeiten sind ebenfalls möglich, da Metadaten (wie Ausgangssprache) getrennt durchsucht werden können.

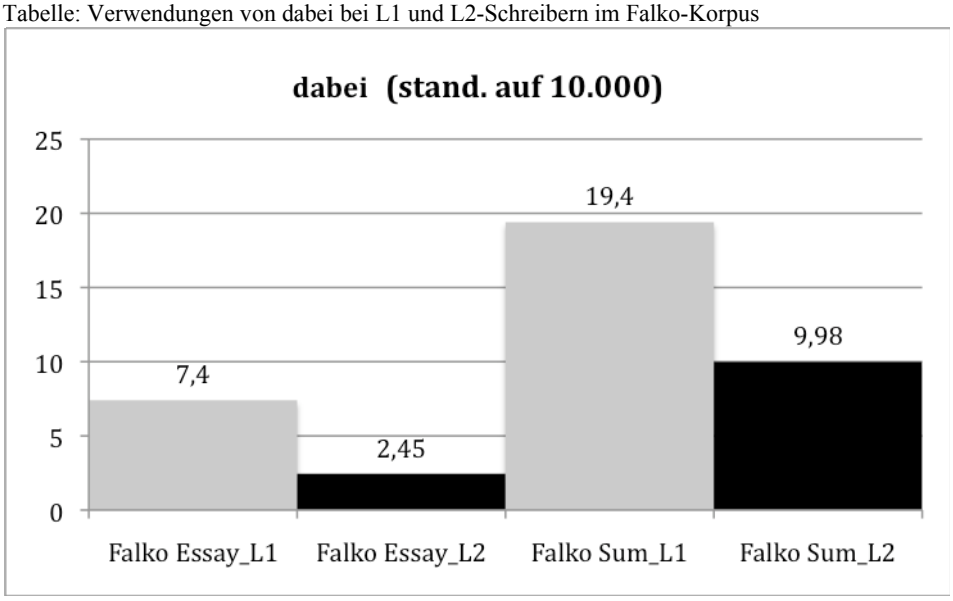
Anhand einiger kurzer und sehr einfacher Beispielabfragen zum Gebrauch des Pronominaladverbs *dabei* soll illustriert werden, wie Suchabfragen zur Hypothesenüberprüfung genutzt werden können. Das Pronominaladverb *dabei* ist das frequenteste unter den Pronominaladverbien (vgl. Schmidt, K. in Vorbereitung) und hat – neben anderen Funktionen – vor allem die Funktion eines Konnektors (HdK 2003). Der Konnektorengebrauch unterscheidet sich hinsichtlich der präferierten Elemente zwischen nativen und nichtnativen Schreibern einerseits und hinsichtlich der Präferenz der Konnektoren insgesamt, wie mehrere Studien zum L2-Englischen gezeigt haben (Field/Yip 1992; Bolton/Nelson/Hung 2002), aber auch zum Deutschen (Walter/Schmidt 2008). Häufig wurde eine Overuse festgestellt, d.h. ein quantitativ stärkerer Einsatz der Konnektoren, ein Mehrgebrauch, im Vergleich zu Muttersprachlern. Wir formulieren in erster Annäherung die Hypothese, dass *dabei* häufiger von L2-Schreibern eingesetzt wird als von L1-Schreibern.

Um eine Suchabfrage formulieren zu können, muss man sich zuvor mit der Abfragesprache vertraut machen. Diese einfach zu erlernende Abfragesprache kann mittels eines Manuals schnell selbständig erlernt und geübt werden (vgl. Beispielabfragen unter http://korpling.german.hu-berlin.de/falko/queryHelp.do; ein einführendes Tutorial unter http://korpling.german.hu-berlin.de/korpus-docs/cqp-tutorial.pdf und ein ausführliches Manual unter http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/cqpman.html) . Wichtig und hilfreich für Abfragen und Analysen ist, dass einerseits einfache Wörter oder Wortfolgen gesucht werden können, z.B. [word="dabei"] oder auch [word="die"][word="schöne"][word="Türkei"]. Auf der anderen Seite können auch Muster gesucht werden, z.B. alle Nominalphrasen, die aus Artikel plus Adjektiv plus Substantiv aufgebaut sind (Suchabfrage: [pos="ART"][pos="ADJA"][pos= „NN"]. Dabei können Suchabfragen auf mehreren Ebenen kombiniert werden, beispielsweise verhindert die Abfrage: [pos="PAV" & word= „damit"], dass Belege von *damit* in seiner Funktion als finaler Subjunktor angezeigt werden. Mittels der angeführten Suchabfrage werden nur diejenigen Belege angezeigt, in denen das Wort *damit* als Pronominaladverb auftritt. PAV ist hierbei das Wortartenkürzel für Pronominaladverb (vgl. STTS 1995/6). Die erste, sehr einfache Wortform-Suchabfrage müsste daher lauten: [word="dabei" %c]:

Bildschirmfoto einer Suchabfrage in Falko:

Die Ergebnisse zum Gebrauch von *dabei* in den beiden Subkorpora Falko-Zusammenfassung und Falko-Essay lassen sich wie folgt darstellen:

Tabelle: Verwendungen von dabei bei L1 und L2-Schreibern im Falko-Korpus



Entgegen der formulierten Annahme wird deutlich, dass L2-Schreiber in beiden der herangezogenen Subkorpora – dem Zusammungsfassungs- und dem Essaykorpus – *dabei* weniger häufig verwenden als die nativen Kontrollgruppen (vgl. die jeweils grauen Spalten mit den schwarzen). Ein Overuse dieses Konnektors kann demnach nicht konstatiert werden[15]. Gleichzeitig wird ein genrespezifischer Unterschied deutlich: Sowohl L1- als auch L2-Schreiber verwenden *dabei* häufiger in Zusammenfassungen als in Essays (vgl. die beiden linken mit den beiden rechten Spalten). Dieser Befund könnte Anlass für weitere Hypothesen zur textlinguistischen Spezifik der einzelnen Textsorten sein.

Wir möchten jedoch zunächst als Beispiel eine weitere Analyse zu den unterschiedlichen Verwendungsweisen von *dabei* durchführen. *Dabei* fungiert häufig als Konnektor und kann als solcher verschiedenen Lesarten haben, unter anderem eine temporale oder eine adversative[16]. Im Allgemeinen kann *dabei* (wie auch viele andere Pronominaladverbien) im Vorfeld und im Mittelfeld stehen. Die adversative Lesart ist jedoch an ein bestimmtes Stellungsmuster – die Vorfeldstellung - gebunden, wie folgende Beispiele illustrieren:

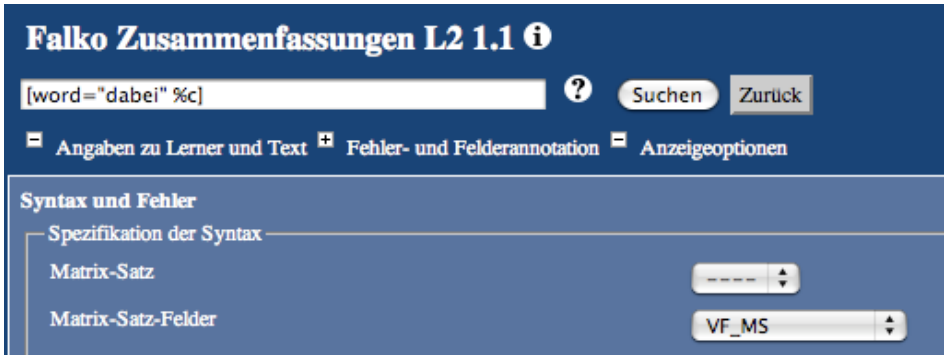
1. Er isst viel. Er schmatzt *dabei* sehr laut. (Temporalität, Gleichzeitigkeit)
2. Er isst viel. *Dabei* schmatzt er sehr laut. (Temporalität, Gleichzeitigkeit)
3. Er isst viel. *Dabei* hat er gar keinen Hunger. (Adversativität)
4. Er isst viel. Er hat *dabei* gar keinen Hunger. (?Temporalität?)

Eine temporale Lesart von 3 ist nicht möglich, wohingegen eine adversative von 4. ebenfalls ausgeschlossen scheint. Bezüglich der Verwendungen von *dabei* liegt die Vermutung nahe, dass a) adversative Verwendungen von *dabei* insgesamt nicht häufig in den Korpora zu finden sein werden, da sie nicht der prototypischen Verwendungsweise entsprechen (vgl. auch temporales und adversatives *während)*. Des weiteren kann vermutet werden, dass b) adversative Verwendungen – wenn überhaupt – häufiger bei L1-, als bei L2-Sprechern auftreten. Dazu müssen sämtliche Vorfeld-Belege von *dabei* vergleichend untersucht werden. Die Suchabfrage (Wortsuche kombiniert mit Feldersuche) kann dazu dienen, alle Einzelbelege aufzuführen:

[15] Die Frage, ob insgesamt Konnektoren mehr oder weniger häufiger durch L2-Sprecher gebraucht werden, also ein Over- oder Underuse konstatiert werden kann, kann selbstverständlich nur unter Einbeziehung aller oder zumindest mehrerer Konnektoren beantwortet werden.

[16] Die komitative Lesart darf natürlich nicht übersehen werden (vgl. HdK 2003: 560), die Differenzierung zwischen der temporalen und komitativen Lesart interessiert hier jedoch nicht.

Bildschirmfoto einer kombinierten Suchabfrage in Falko (Wort plus Vorfeld)[17]:



Die anschließende Desambiguierung der Lesarten, d.h. die Bestimmung, welcher Beleg eine adversative Lesart hat oder haben kann, muss jedoch manuell durch den Forschenden geschehen. Die Suchabfragen helfen uns demnach, Belege schnell und strukturiert zu finden, sie nehmen uns jedoch nicht die Aufgabe ab, diese Belege zu analysieren und interpretieren. Dies ist und bleibt Aufgabe des Forschenden. Die Vorfeld-Abfrage (nur im Zusammenfassungskorpus) und die anschließende Lesartenbestimmung ergeben keinen Beleg eines adversativ gebrauchten *dabei* – weder bei den L1-, noch bei den L2-Schreibern. Während unsere erste Hypothese (a) damit bestätigt werden kann (adversativer Gebrauch ist sehr selten), kann die zweite (b) auf Grund der vorliegenden Datenlage nicht beantwortet werden. Hierzu müsste entweder ein größeres Korpus hinzugezogen werden oder es müssten Hypothesen zum adversativen Erwerb von *dabei* durch die Lerner experimentell erhoben werden. Beispielsweise könnten wir vermuten, dass die adversative Lesart erst spät von Lernern rezeptiv verstanden wird (und ergo noch später produziert wird). Korpusrecherchen bilden daher oft den Anfang von Untersuchungen, die im Anschluss daran durch andere Erhebungsmethoden sinnvoll ergänzt werden können (und müssen). Die Beispielabfrage illustriert außerdem, dass es wichtig ist, in kleinen Korpora – wie in Falko - möglichst frequente Phänomene zu untersuchen, um zu verwertbaren Ergebnissen zu kommen (ein Beispiel wäre der Artikelgebrauch); dies kann unter Ausnutzung der jeweiligen Subkorpora beispielsweise auch im Hinblick darauf geschehen, ob es Unterschiede zwischen Lernern verschiedener Herkunftssprachen gibt (artikellose Sprachen wie Turksprachen und slawische Sprachen im Kontrast zu artikelhaltigen

[17] VF_MS steht für Vorfeld im Matrixsatz.

Sprachen wie germanischen oder romanischen Sprachen etc.). Für selten auftretende Phänomene hingegen sind große Korpora besser geeignet.

7. Ausblick

Öffentlich zugängliche Korpora – speziell Lernerkorpora - können gerade für Auslandsgermanisten eine sinnvolle Hilfe in der Forschung sein:

• Zusammenarbeit mit anderen Linguisten ist standortunabhängig problemlos möglich (die Daten können gleichzeitig an verschiedenen Universitäten und Standorten analysiert werden).

• Lernerkorpora können quantitativ, qualitativ oder auch sprachkontrastiv analysiert werden, d.h. türkische Spracherwerbsforscher können die L2-Daten Deutsch der türkischen L1-Sprecher und russische oder englische KollegInnen die Daten der russischen oder englischen Deutsch-Lernenden unter derselben Fragestellung vergleichen. Der L1-Einfluss typologisch unterschiedlicher Erst- und Zweitsprachen kann gut herausgearbeitet werden.

• Lernerkorpora mit einem muttersprachlichen Vergleichskorpus (wie z.B. Falko) verringern den „Standortnachteil" der Auslandsgermanisten, die im Ausland oft nur erschwert bzw. gar nicht parallele Kontrolldaten von L1-Sprechern erheben können.

Korpora und Lernerkorpora haben darüber hinaus wichtige Anwendungsgebiete auch in der Lehrmaterialerstellung und in der Fremdsprachenvermittlung, auf die aus hier Platzgründen nicht eingegangen werden kann (vgl. hierzu auch Lüdeling/Walter 2009; Römer 2008). Forscher und Lehrende, die für eine bestimmte Untersuchung gerne ein eigenes Lernerkorpus erstellen möchten (diese Aspekte konnten hier nicht behandelt werden), seien vor allem verwiesen auf Hunston 2008, Granger 2008.

Literatur

Ahrenholz, Bernt/ Ursula Bredel/ Wolfgang Klein/ Martina Rost-Roth/ Romuald Skiba (Hg.) (2008). Empirische Forschung und Theoriebildung. Beiträge aus der Soziolinguistik, Gesprochene-Sprache- und Zweitspracherwerbsforschung. Berlin u.a.: Peter Lang.

Belz, Julie A. (2005). Telecollaborative language study: A personal overview of praxis and research. Selected papers from the 2004 NFLRC Symposion. Online unter http://nflrc.hawaii.edu/networks/nw44/belz.htm.

Bolton, K./ G. Nelson/ J. Hung (2002). A corpus-based study of connectors in student writing. International Journal of Corpus Linguistics 7,2: 165-182.

EAGLES (1996). Preliminary recommendations on corpus typology. EAGTCWG-CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di

Linguistica Computazionale. Online unter http://www.ilc.cnr.it/EAGLES96/corpintr/corpintr.html
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. Humboldt-Universität zu Berlin: Unveröffentlichte Magisterarbeit.
Fandrych, Christian/ Erwin Tschirner (2007). Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. Deutsch als Fremdsprache 44, 4: 195-204.
Fernandez-Villanueva, Marta/ Oliver Strunk (2009). Das Korpus Varkom. Variation und Kommunikation in der gesprochenen Sprache. Deutsch als Fremdsprache 46, 2: 67-73.
Field, Y/ L. Yip (1992). A comparison of internal conjuncitve cohesion in the English essay writing of Cantonese speakers and native speakers of English. RELC Journal 23,1: 15-28.
Granger, Sylviana (2008). Learner Corpora. In: Lüdeling, Anke/ Merja Kytö (Hg.) 259-275.
Granger, Sylviane/ Joseph Hung/ Stephanie Petch-Tyson (Hg.) (2002). Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam/Philadelphia: John Benjamins.
Helbig, Gerhard/ Lutz Götze/ Gert Henrici (Hg.) (2001). Deutsch als Fremdsprache (HSK). Ein internationales Handbuch. 2 Bde. Berlin u.a.: Walter de Gruyter.
Hirschmann, Hagen (2005). Platzhalterphrasen bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. Humboldt-Universität zu Berlin: Unveröffentliche Zulassungsarbeit (Staatsexamen).
Hirschmann, Hagen/ Seanna Doolittle/ Anke Lüdeling (2007). Syntactic Annotation of Non-Canonical Linguistics Structures. Proceedings of Corpus Linguistics 2007. Birmingham. Online unter http://www.corpus.bham.ac.uk/corplingproceedings07/paper/128_Paper.pdf
Hunston, Susan (2008). Collection strategies and design decision. In: Lüdeling/Kytö (Hg.), 154-168.
Klein, Wolfgang/ Clive Perdue (1992). Why does the production of some learners not grammaticalize? Studies in Second Language Acquisition 14: 259-272.
Köhler, Reinhard (2005). Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. Zeitschrift für Computerlinguistik und Sprachtechnologie 20, 2: 1-16. Online unter http://sirao.kgf.uni-frankfurt.de/gldv/2005_Heft2/Reinhard_Koehler.pdf.

Kroymann, Emil (2008). Wortstellungsvarianz im Mittelfeld bei Fremdsprachenlernern und bei Muttersprachlern des Deutschen. Humboldt-Universität zu Berlin: Unveröffentlichte Magisterarbeit.
Krumm, Hand-Jürgen(Christian Fandrych/ Britta Hufeisen/ Claudia Riemer (Hg.) (im Druck). Deutsch als Fremd- und Zweitsprache. Berlin u.a.: Walter de Gruyter.
Lemnitzer, Lothar/ Heike Zinsmeister (2006). Korpuslinguistik. Eine Einführung. Tübingen: Narr.
Lippert, Eva (2005). Probleme von Nichtmuttersprachlern mit der Definitheit von Nominalphrasen. Humboldt-Universität zu Berlin: Unveröffentlichte Magisterarbeit.
Lüdeling, Anke (2008). Merhdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter/ Grommes (Hg.), 119-140.
Lüdeling, Anke/ Seanna Doolittle/ Hagen Hirschmann/ Karin Schmidt/ Maik Walter (2008). Das Lernerkorpus Falko. Deutsch als Fremdsprache 45, 2: 67-73.
Lüdeling, Anke/ Merja Kytö (Hg.) (2008) Corpus Linguistics. An International Handbook. Vol 1. Berlin/New York: Mouton de Gruyter.
Lüdeling, Anke/ Merja Kytö (Hg.) (2009) Corpus Linguistics. An International Handbook. Vol 2. Berlin/New York: Mouton de Gruyter.
Lüdeling, Anke/ Maik Walter(2009). Korpuslinguistik und Deutsch als Fremdsprache. Online unter https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/anke_veroeffentlichungen.
Lüdeling, Anke/ Walter, Maik (im Druck). Korpuslinguistik und Deutsch als Fremdsprache. In: Krumm et al. (Hg.).
Lüdeling, Anke/ Maik Walter/ Emil Kroymann/ 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 (o.J.): Modality in Learner German. Online-Folienpräsentation unter http://209.85.135.132/search?q=cache:oflwbQK1OeUJ:userpage.fu-berlin.de/~maik/dgfs/maden_folien.pdf+%22CLeG%22+Maden-Weinberger&cd=1&hl=tr&ct=clnk&client=safari).
Miklitz, G. (2002). Online-Recherche in Corpora - Hinweise für DaF-Lehrer und Studierende von DaF. Bonn: Studienkolleg. Online unter http://www.lernforum.uni-bonn.de/coropora.html.
Pasch, Renate/ Ursula Brauße/ Eva Breindl/ Ulrich Hermann Waßner (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der

Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). Berlin: Walter De Gruyter.
Perdue, Clive (1993). Adult language acquistion. Crosslinguistic perspecktives. 2 Bde. Cambridge: Cambridge University Press.
Pravec, Norma A. (2002). Survey of learner corpora. ICAME Journal 26: 81-114.
Römer, Ute (2008). Corpora and language teaching. In: Lüdeling, Anke/ Merja Kytö (Hg.), 112-131.
Scherer, Carmen (2006). Korpuslinguistik. Heidelberg: Winter.
Schmidt, Claudia (2008). Grammatik und Korpuslinguistik. Überlegungen zur Unterrichtspraxis DaF. Deutsch als Fremdsprache 45, 2: 74-80.
Schmidt, Karin (in Vorbereitung). Pronominaladverbien im Kontext Deutsch als Fremdsprache. Diss. FU Berlin.
Siemen, Peter/ Anke Lüdeling/ Frank Henrik Müller (2006). Falko - ein fehlerannotiertes Lernerkorpus des Deutschen. In: Proceedings of Konvens 2006, Konstanz. Pdf-Download unter http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko
Skiba, Romuald (2008). Korpora in der Zweitspracherwerbsforschung. Internetzugang zu Daten des ungesteuerten Zweitspracherwerbs. In: Ahrenholz, Bernt et al. (Hg.), 21-30.
Solte-Gresser, Christiane/ Karin Struve/ Natascha Ueckmann (Hg.) (2005).Von der Wirklichkeit zur Wissenschaft. Aktuelle Forschungsmethoden in den Sprach-, Literatur- und Kulturwissenschaften, 147-161. Hamburg: LIT-Verlag.
Stefanowitsch, Anatol (2005). Quantitative Korpuslinguistik und sprachliche Wirklichkeit. In: Solte-Gresser, Christiane/Karin Struve/Natascha Ueckmann (Hg), Von der Wirklichkeit zur Wissenschaft. Aktuelle Forschungsmethoden in den Sprach-, Literatur- und Kulturwissenschaften, 147-161. Hamburg: LIT-Verlag. Online unter http://www-user.uni-bremen.de/~anatol/docs/ms_korpuslinguistik.pdf.
Stuttgart Tübingen Tag Set (1995/1999). Entwickelt am Institut für maschinelle Sprachverarbeitung Stuttgart und am Seminar für Sprachwissenschaft an der Universität Tübingen. Online unter http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html
Walter, Maik/ Patrick Grommes (Hg.) (2008). Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Niemeyer.
Walter, Maik/ Karin Schmidt (2008). "Und das ist auch gut so". Der Gebrauch des satzinitialen 'und' bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. In: Ahrenholz, Bernt et al. (Hg.), 331-342.