Zum Themenschwerpunkt „Korpuslinguistik und Deutsch als Fremdsprache", eröffnet in Heft 4/2007, äußern sich in diesem Heft Anke Lüdeling / Seanna Doolittle / Hagen Hirschmann / Karin Schmidt / Maik Walter sowie Claudia Schmidt. Die Redaktion

Anke Lüdeling / Seanna Doolittle / Hagen Hirschmann / Karin Schmidt / Maik Walter[1]

Das Lernerkorpus Falko

aus: Deutsch als Fremdsprache 2/2008, S. 67–73.

0 Lernerkorpora in der Erforschung des Fremdsprachenerwerbs

In diesem Artikel wird das frei zugängliche **fehlerannotierte Lernerkorpus** Falko (http://www.linguistik.hu-berlin.de / institut / professuren / korpuslinguistik / forschung / falko) vorgestellt, das Texte von fortgeschrittenen Lernern des Deutschen als Fremdsprache enthält. Auch wenn es bisher noch kaum elektronisch verfügbare Lernerkorpora für DaF gibt, steht natürlich die empirische Beschäftigung mit Sammlungen von Lernerdaten in langer Tradition. Viele Aspekte des Erst- oder Zweit-/Fremdsprachenerwerbs können nur auf der Basis von tatsächlich vorkommenden Lerneräußerungen untersucht werden. Dies betrifft einerseits die genaue Betrachtung von Fehlern bzw. Abweichungen (in einem Kontext) und andererseits vor allem auch die Ermittlung quantitativer Eigenschaften von Lernersprache (im Vergleich zur L1 oder im Vergleich mit anderen Lernervarietäten). Lerneräußerungen können entweder in (mehr oder weniger) authentischen Situationen entstehen oder in gezielten Fragebogenstudien oder psycholinguistischen Studien elizitiert werden.[2] Seit sich die Wissenschaft systematisch mit der Erforschung des Spracherwerbs beschäftigt, werden Lerneräußerungen gesammelt und ausgewertet (eine frühe Sammlung von Erstspracherwerbsdaten ist Stern/Stern 1907; zur Rolle von Daten im Zweitsprachenerwerb vgl. z. B. Dietrich 2006). Viele Sammlungen von Lerneräußerungen sind aber nicht systematisch – die Erkenntnisse, die sich daraus ableiten lassen, sind daher oft nicht ohne weiteres generalisierbar. Außerdem sind die Daten oft nicht allgemein zugänglich, sodass Ergebnisse nicht reproduzierbar sind. In den letzten Jahren werden deshalb Lernerkorpora – also Sammlungen von Lernerdaten, nach genauen Kriterien und mit ausführlichen Metadaten – zunehmend kontrolliert erhoben.[3] Im Folgenden werden wir uns nur auf Lernerkorpora mit Daten von Fremdsprachenlernern beziehen (wir kürzen diese Lernerdaten als L2-Daten ab, unabhängig von der Anzahl der vorher gelernten Sprachen).

Während es für das Englische als Fremdsprache bereits mehrere Lernerkorpora und viele Studien gibt,[4] liegen für das Deutsche bisher nur sehr wenige frei verfügbare und dokumentierte Lernerkorpora vor.[5] Um diese

[1] Unser Dank geht an Emil Kroymann, Marc Reznicek und alle anderen Falkos für ihre Hilfe.

[2] Ergänzend dazu gibt es weitere psycholinguistische Studien, wie zum Beispiel die Studien zur Konzeptualisierung in der L2 (vgl. von Stutterheim/Carroll 2006).

[3] Neben Lernerkorpora spielen natürlich auch Muttersprachlerkorpora (L1-Korpora) im Fremdsprachenbereich eine Rolle, allerdings eher im Bereich Sprachvermittlung. Wir können hier nicht auf L1-Korpora oder L2-Korpora in der Sprachvermittlung eingehen; vgl. dazu zum Beispiel Nesselhauf (2004); Fandrych/Tschirner (2007); Römer (i. Dr.).

[4] Das bekannteste Lernerkorpus des Englischen als Fremdsprache ist sicher das International Corpus of Learner English (ICLE), das an der Universität Louvain-la-Neuve erstellt wird (vgl. Granger/Dagneaux/Meunier 2002). Das Korpus besteht aus Essays von Englischlernern mit unterschiedlichen Muttersprachen und ist Grundlage für sehr viele Untersuchungen (eine Sammlung von Literaturhinweisen findet sich unter http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm). Für weitere Korpora siehe Granger (2003); Granger (i. Dr.) und die Proceedings der TALC(Teaching and Language Corpora)-Konferenzen.

[5] Neben privaten, nicht öffentlich zugänglichen Sammlungen (zum Beispiel in Wegener 1995; Birkner/Dimroth/Dittmar 1995; Weinberger 2002; Belz 2004) gibt es, soweit wir wissen, nur die folgenden: das phonologisch annotierte LEAP-Korpus (vgl. Milde/Gut 2002) und die ESF-Korpora des MPI Nijmegen (http://corpus1.mpi.nl/ds/imdi_browser).

Lücke zu schließen, wurde das Lernerkorpus Falko in einer Zusammenarbeit der Freien Universität (FU) Berlin und der Humboldt-Universität (HU) zu Berlin konzipiert und erhoben.

Wir möchten zunächst in Abschn. 1 einige Aspekte der Erstellung von Lernerkorpora ansprechen und erläutern, wie wir uns jeweils für das Falko-Korpus entschieden haben. Fehlerannotation ist für viele Forschungsfragen hilfreich, aber konzeptuell sicherlich problematisch. Daher werden wir in Abschn. 2 auf die Fehlerannotation in Falko genauer eingehen. In Abschn. 3 werden dann exemplarisch zwei kurze Studien mit Falkodaten vorgestellt.

1 *Design und Architektur von Falko*

1.1 *Korpusdesign*

Es gibt viele Definitionen des Begriffs „Korpus". In der Korpuslinguistik hat sich aber in den letzten Jahren weitgehend durchgesetzt, dass eine Textmenge dann als Korpus bezeichnet wird, wenn sie für einen bestimmten Forschungszweck nach vorher definierten externen oder internen Kriterien zusammengestellt wurde (vgl. zum Beispiel die EAGLES-Definition unter http://nl.ijs.si/et/talks/korpus/eagles-corpus.html; Hunston i. Dr.). Für Lernerkorpora einer gegebenen L2 sind hier – je nach Forschungsfrage – externe Kriterien (mehr als eine L1, gesprochene oder geschriebene Texte, Erhebungssituation, Lernstand) genauso wie interne Kriterien (Texte, die bestimmte sprachliche Strukturen aufweisen) vorstellbar (vgl. die Diskussion in Granger/Hung/Petch-Tyson 2002: 9f.).

Das Falko-Korpus ist nach externen Kriterien zusammengestellt worden. Es enthält Texte von fortgeschrittenen Lernern[1] in zwei Subkorpora, die von der FU und der HU erhoben wurden, und ein Longitudinalkorpus, das an der Georgetown University in Washington erhoben wurde. Im Folgenden beschreiben wir nur die durch die FU und die HU erhobenen Subkorpora (die Korpusgrößen für die einzelnen Subkorpora finden sich in Tab. 1, s. S. 69). Lernerspezifische Metadaten wie z. B. Lernbiographie, L1, Alter oder Geschlecht sowie die Metadaten über die jeweilige Erhebung können abgefragt und zur Bildung von eigenen speziellen Korpora verwendet werden.

Die Erhebungen für beide Subkorpora waren stark kontrolliert: Es durften keine Hilfsmittel verwendet werden, und die Zeit war vorgegeben.[2]

Das Zusammenfassungskorpus enthält Zusammenfassungen von linguistischen oder literaturwissenschaftlichen Fachtexten, die im Rahmen einer obligatorischen Sprachprüfung von nichtmuttersprachlichen Studierenden der Germanistik an der FU erstellt wurden. Die Fortgeschrittenheit wird hier dadurch definiert, dass die Studierenden die DSH-Prüfung bestanden und ihr Grundstudium in Deutschland absolviert haben. Zugeordnet ist ein Kontrollkorpus mit Daten von Muttersprachlern, das unter vergleichbaren Bedingungen (dieselben Vorlagentexte, gleiche Bearbeitungszeit etc.) erhoben wurde. Außerdem sind die Vorlagentexte abfragbar. Das Zusammenfassungskorpus ist abgeschlossen. Als externes Vergleichskorpus haben wir ein Korpus mit deutschsprachigen Dissertationsabstracts zusammengestellt.[3]

Das Essaykorpus enthält Essays zu vier kontrovers zu diskutierenden Themen, die sich an Themen des ICLE-Korpus anlehnen. Die Daten wurden in unterschiedlichen Ländern erhoben,[4] das Korpus wächst noch. Die Fortgeschrittenheit wird mit Hilfe eines C-Tests überprüft; nur Texte von Lernern, die einen bestimmten Wert überschreiten, werden aufgenommen. Zugeordnet ist wieder ein L1-Korpus, das an Brandenburger und Berliner Gymnasien (12./13. Klasse) erhoben wurde.

[1] Die Fortgeschrittenheit wird unterschiedlich festgestellt, s. unten; zur Diskussion von Fortgeschrittenheit vgl. auch Walter/Grommes (2008a).

[2] Einige Texte wurden handschriftlich erhoben, andere am Computer. Bei handschriftlicher Erhebung fallen Tippfehler weg. Allerdings ist es oft kaum möglich, die Handschrift eindeutig zu entziffern. In der Digitalisierung müssen daher Entscheidungen getroffen werden. Außerdem können hier Übertragungsfehler entstehen. Direkt am Computer entstandene Texte können Tippfehler enthalten. In unseren Erhebungen wurde immer darauf geachtet, dass kein Textverarbeitungsprogramm mit eingebauter Rechtschreibkorrektur verwendet wurde.

[3] Das frei zugängliche Korpus Akademisches Deutsch umfasst 1,8 Millionen Tokens. Es ist automatisch mit Wortarten annotiert und lemmatisiert, ein nach Fachgebieten ausgewogenes Subkorpus Akademisches Deutsch 2006 umfasst 841.483 Tokens.

[4] Bisher haben wir Daten aus Dänemark, Kenia, Usbekistan, der Türkei und Südafrika sowie Daten aus Ferienkursen an der HU und der FU. Weitere Erhebungen sind geplant.

Subkorpus	Texte	Tokens	Types
Zusammenfassungen L2, Version 1.1	107	41075	5181
Zusammenfassungen L1, Version 1.1	57	21211	4099
Essays L2, Version 1.0	132	65387	7720
Essays L1, Version 0.5	39	33806	5831

Tab. 1: Korpusgrößen für die einzelnen Subkorpora in Falko
(Das Zusammenfassungskorpus ist abgeschlossen; das Essaykorpus wird weiter wachsen.)

1.2 *Korpusarchitektur*

Die Forschungsfragen beeinflussen nicht nur das Korpusdesign, sondern auch die Korpusarchitektur, d. h. das formale Modell, in dem das Korpus und seine Annotationen gespeichert sind. Unterschiedliche Architekturen ermöglichen bzw. verhindern bestimmte Annotationen. Die meisten Korpora sind flach annotiert, d. h., die Annotation erfolgt mit besonderer Kennzeichnung in derselben Datei wie die Daten, sei es in einem Tabellenmodell oder in einem Baummodell (XML).[1] Das ist für große, einfach annotierte Korpora eine gute Option. Bestimmte Möglichkeiten gibt es aber in derart annotierten Korpora nicht (vgl. Lüdeling/Walter/Kroymann/Adolphs 2005; Lüdeling 2007). Ein großes Problem ist, dass alternative Analysen für die gleichen Daten auf diese Weise nicht annotiert werden können. Für ein kleines, spezialisiertes Korpus wie Falko mit Daten, deren Interpretation oft umstritten sein kann (wie zum Beispiel die Zielhypothese, s. Abschn. 2), bietet sich daher eine Korpusarchitektur an, in der die Annotationen getrennt von den eigentlichen Daten gespeichert werden. Mit einer solchen Mehrebenenarchitektur können zu den Daten beliebig viele voneinander unabhängige Annotationsebenen eingefügt werden.[2] Dies ist insbesondere für die Fehlerannotation wichtig, auf die wir in Abschn. 2 eingehen. Alle Annotationsebenen können bei der Suche kombiniert werden.

[1] Dies gilt für alle uns bekannten Lernerkorpora außer dem LEAP Korpus (vgl. Milde/Gut 2002).

[2] Wir verwenden zur Annotation das an der Universität Hamburg entwickelte Programm EXMARaLDA (vgl. Schmidt/Wörner 2005).

1.3 *Annotation*

In nichtannotierten Korpora kann man nur nach Zeichenketten suchen. Um Strukturen oder Muster suchbar zu machen, werden Korpora annotiert, d. h. mit Metainformationen versehen. Neben Header-Informationen, die sich auf ganze Dokumente beziehen, können auch kleinere Einheiten annotiert werden. Dazu muss ein Korpus zunächst in kleinste annotierbare Einheiten (Tokens) zerlegt werden. Wir haben uns in Falko für eine automatische Tokenisierung in graphemische Wörter (Abrücken von Satzzeichen, stattdessen Zeichenketten zwischen jeweils zwei Leerzeichen) entschieden, weil das in den meisten Korpora europäischer Sprachen so gemacht wird. Die bekannten Probleme einer solchen Tokenisierung (Mehrwortlexeme und mehrdeutige Zeichen; vgl. Schmid i. Dr.) nehmen wir in Kauf.

Falko wird automatisch mit Wortarten annotiert. Dazu verwenden wir den TreeTagger (vgl. Schmid 1994) mit dem Stuttgart-Tübingen-Tagset (http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html). Die Fehlerrate im Wortarten-Tagging ist in einem Lernerkorpus wegen der orthographischen Fehler und der Wortstellungsfehler höher als bei Zeitungskorpora. Aus diesem Grund wurde in Teilen des Korpus die automatische Annotation manuell nachkorrigiert (und diese zusätzliche Annotationsebene kann bei Suchanfragen berücksichtigt werden).

Ebenfalls automatisch erfolgt die Annotation der Lemmata. D. h., jeder Wortform wird eine Grundform zugewiesen, sodass nach allen Formen, die zu derselben Grundform gehören, gesucht werden kann.

In Falko wurden darüber hinaus bisher Zielhypothesen (s. Abschn. 2) und topologische Felder (vgl. Höhle 1986) annotiert. Mit Letzteren ist es beispielsweise möglich, Wortstellungsfehler, die Besetzung von Vorfeldern oder von Satzklammern effizient zu untersuchen. Durch eine Kombination der einzelnen Ebenen können auch Lemmata in den einzelnen Feldern (z. B. alle Flexionsformen des Nomens *Problem* im Vorfeld eines Matrixsatzes) oder aber spezifische Wortartbesetzungen in den Feldern (z. B. alle attributiven Adjektive,

Tokens	Und	immer	noch	kann	man	eine	unzufriedenheit	spüren
Lemma	und	immer	noch	können	man	ein	unknown	spüren
Wortart	KON	ADV	ADV	VMFIN	PIS	ART	NN	VVINF
Zielhypothese							Unzufriedenheit	
Topologische Felder	Vorfeld			linke Satzklammer	Mittelfeld			rechte Satzklammer

Tab. 2: Schematische Darstellung eines Ausschnitts aus dem annotierten Lernertext Falko Essays L2 1.0, fk024, 247[1]

die im Mittelfeld eines Konstituentensatzes erscheinen) herausgefiltert werden.

In Tab. 2 wird eine Lerneräußerung auf den verschiedenen Ebenen dargestellt. Die Notation wird in der Dokumentation, die sich auf der Falko-Homepage befindet, ausführlich beschrieben. In der ersten Zeile befinden sich die einzelnen Tokens, in der zweiten und dritten Zeile das Ergebnis der automatischen Lemmatisierung bzw. der Wortartenerkennung. Vom Lemmatisierer nicht erkannte Einheiten (z. B. *unzufriedenheit*) werden als „unknown" markiert. Die Zielhypothese und die topologischen Felder werden manuell in den beiden letzten Zeilen hinzugefügt.

2 *Zielhypothese und Fehlerannotation*

Wie in der Einleitung bereits angesprochen, kann man die Lernerdaten verwenden, um Abweichungen (Fehler) zu untersuchen und um quantitative Eigenschaften der Lernersprache zu ermitteln (vgl. z. B. Granger 2002: 11ff.; Walter/Grommes 2008a: 15ff.). Für die Fehlerforschung ist eine Fehlerannotation hilfreich, mithilfe deren Fehler eines bestimmten Typs zuverlässig gefunden werden können.

An dieser Stelle wollen wir die Diskussion um den Begriff „Fehler" nicht aufgreifen (vgl. aber z. B. Cherubim 1980; Ellis 1994: 50ff.; Lennon 1991). Wir definieren „Fehler" als Abweichung von einer Zielhypothese, welche von uns auf der Basis der Lerneräußerung rekonstruiert wird (s. unten). Die Entwicklung von Fehler-Tagsets kann nach unterschiedlichen Kriterien (linguistische Ebene, formaler Fehlertyp, Prinzipienverletzung etc.) und in unterschiedlicher Granularität erfolgen. Die meisten Fehler-Tagsets, wie z. B. das Tagset für ICLE (vgl. Dagneaux/Denness/Granger/Meunier 1996) oder das Tagset in Weinberger (vgl. 2002), sind Mischungen aus mehreren Kriterien und daher nicht immer leicht nachvollziehbar. In einer Mehrebenenarchitektur, in der beliebig viele Annotationsebenen eingefügt werden können, muss man sich nicht von vornherein auf ein Fehler-Tagset festlegen.

Eine notwendige Voraussetzung für jede Fehlerannotation ist die Annahme einer im Kontext „korrekten" Form, die wir in Anlehnung an Ellis (vgl. 1994) „Zielhypothese" nennen. In den meisten Fehlerstudien und fehlerannotierten Korpora bleibt diese Form implizit.[2] Ein Problem ist aber, dass es zu jeder abweichenden Äußerung mehrere Zielhypothesen geben kann und dass natürlich die angenommene Zielhypothese die Fehlerart und -zahl bestimmt (das Problem ist größer, als vielfach angenommen; vgl. Lüdeling 2008). In Falko wird die Zielhypothese daher explizit angegeben. Alle Fehlerannotationen erfolgen dann in Bezug auf diese Zielhypothese. Die Erstellung der Zielhypothesen ist demnach der wichtigste Bearbeitungsschritt bei der Annotation von Abweichungen; zusammengefasst erfüllt er folgende zwei Hauptzwecke:

- die Kennzeichnung der Abweichungen sowie
- die Grundlegung einer transparenten Basis für die Identifizierung und Klassifizierung der Fehler.

Durch die Mehrebenenarchitektur ist es möglich, mehrere konkurrierende Zielhypothesen anzugeben. Dieses Vorgehen dient darüber hinaus auch einer Qualitätssicherung in der deskriptiven Analyse.

[1] „fk024", „247" und analoge Angaben im folgenden Text sind Referenzen auf Korpusstellen (hier: Text 24, Token 247).

[2] Es gibt einige wenige Korpora, in denen die Korrektur auch explizit angegeben ist, vgl. zum Beispiel Izumi/Uchimoto/Isahara (2005) oder das norwegische ASK-Korpus (http://decentius.aksis.uib.no/corpus/askdemohome.xml). Da diese Korpora keine Mehrebenenarchitektur haben, kann jeweils nur eine Zielhypothese angegeben werden.

3 *Die Analyse von Lernersprachen – zwei Beispiele*

Im Folgenden werden exemplarisch zwei Lernerkorpusanalysen skizziert, die mit Falko-Daten vorgenommen wurden. Zunächst werden in einer quantitativen Vergleichsstudie anhand der satzinitialen Konjunktion *und* mögliche Präferenzen (der Lerner) untersucht, anschließend wird eine qualitative und quantitative Studie im Bereich der Orthographie vorgestellt.

3.1 *Das satzinitiale „und"*

Die Sprache in L2-Korpora weicht in vielen Fällen quantitativ von der Sprache vergleichbarer L1-Korpora ab (vgl. Walter/Grommes 2008a). Die Unterschiede liegen häufig nicht mehr im Bereich der Grammatik, sondern in Bereichen, die traditionell als stilistische Fehler (oder Ausdrucksfehler) bezeichnet werden. Die Lernertexte sind hierbei grammatisch korrekt, aber Muttersprachler würden andere Konstruktionen wählen (das Problem der „nativelike selection"). In Walter/Schmidt (i. Dr.) wird eine in der L1 stilistisch markierte Konstruktion in der Lernersprache untersucht, die Verwendung des satzinitialen *und*, die zwar nicht ungrammatisch ist, aber normalerweise nur in bestimmten Kontexten auftritt (vgl. z. B. Baumgarten 2006; Bell 2007):

(1) Und wenn man einen anderen Beispiel nemen wurde, konnten Frauen zu dieser Zeit ihre Sexualität nicht kontrolieren. (Falko Essays L2 1.0, fk010, 27)

Fremdsprachenlerner verwenden diese Konstruktion auf eine sehr spezifische Weise. Bei einem Vergleich der L2-Zusammenfassungen und der L1-Zusammenfassungen kann bei Ersteren ein deutlicher Overuse (die Lerner verwenden ein Wort / eine Konstruktion häufiger als Muttersprachler) des satzinitialen *und* festgestellt werden.[1] Im L2-Korpus ist die Konstruktion achtmal häufiger. Darüber hinaus werden die L2-Lernerdaten auf das Genre bezogen ausgewertet, mit Daten aus dem L2-Essay-Korpus kontrastiert und bezüglich ausgewählter Kontextmerkmale analysiert. Besonders häufig formulieren die Lerner mithilfe der Konstruktion in den Essays rhetorische Fragen, wie das folgende Beispiel illustriert:

(2) Und wo haben diese berühmte Menschen ihre Ideen entwickelt, wenn nicht in der Universität? (Falko Essays L2 1.0, fk012, 630)

Weitere Faktoren, z. B. das Auftreten der Konstruktion in einer elliptischen Struktur oder aber die unterschiedliche Verwendung in Konstituenten- bzw. Matrixsätzen, werden in der Studie ausgewertet und diskutiert; für eine vertiefte Darstellung der Ergebnisse vgl. Walter/Schmidt (i. Dr.). Studien wie diese zeigen, dass Lernerkorpora verwendet werden können, um Präferenz (Overuse), aber auch Vermeidung (Underuse) von ausgewählten Konstruktionen bei den Lernern festzustellen. Die Ergebnisse dieser Analysen können beispielsweise in der Lernerlexikographie genutzt werden, indem Lerner in Lernerwörterbüchern explizit auf diese lernersprachlichen Besonderheiten aufmerksam gemacht werden. Für das Englische wurde dies auch schon richtungsweisend im Macmillan English Dictionary für das satzinitiale *and* in der Auswertung der ICLE-Daten vollzogen (vgl. Rundell/Fox 2007: IW4).

3.2 *Orthographie*

Während die Orthographie lange als „idiosynkratisch" oder als für strukturelle Untersuchungen uninteressant galt, wurden in den letzten Jahren mehr und mehr Regelmäßigkeiten gefunden und Verbindungen zu den anderen grammatischen Gebieten gezeigt. Daher kann die Orthographie auch in der Spracherwerbsforschung interessante Hinweise auf phonologische, morphologische und syntaktische Hypothesen der Lerner liefern. In Hirschmann (i. Vorb.) werden die Orthographiefehler im L2-Essay-Subkorpus analysiert und mit L1-Vergleichsdaten[2] kontrastiert. Derzeit wird darauf aufbauend ein Fehler-Tagset bzw. ein Annotationsschema entwickelt, um eine effiziente automatische Abfrage der Orthographiefehler in Falko zu ermöglichen.

Die Orthographiefehler werden zunächst in kontextunabhängige (Wortschreibungs-)Fehler und kontextabhängige (syntaktische) Fehler kategorisiert. So beinhalten die Schreibun-

[1] Hierzu wurde das auf S. 68 in Fußnote 3 bereits angeführte Korpus Akademisches Deutsch 2006 verwendet.

[2] Als L1-Vergleichskorpus dienten die Daten aus Falko Essay L1 0.5, die von Leistungskursschülern des Faches Deutsch produziert wurden.

gen *<nemmen>* und *<kontrolieren>* in (1) orthographische Fehler, die ohne den syntaktischen Kontext diagnostiziert werden können. Die Wortfolge *<in dem>* in (3) hingegen ist nur unter der Berücksichtigung eines Kontextes (wie in Beispiel (3)) als (Getrenntschreibungs-)Fehler zu identifizieren:

(3) ..., aber auf dieser mitleidlosen Welt sollen sie ihre Zierlichkeit nicht verlieren, in dem sie alles versuchen zu machen. (Falko Essays 1.0, trk001, 369)

Die grundlegende Unterscheidung zwischen diesen beiden Fehlertypen wird häufig übergangen (vgl. z.B. Eisenberg 2004: 332), ist jedoch gerade hinsichtlich des Schriftspracherwerbs (in L1 und L2) von hoher Relevanz.

Ein wesentliches Ergebnis der Studie ist, dass das Lernerkorpus mit 111 Fehlern pro 10.000 Wörter (alle Fehlerzahlen sind normalisiert auf 10.000 Einheiten) unerwartet wenig Orthographiefehler enthält, dass der Schnitt sogar deutlich niedriger als in dem L1-Kontrollkorpus (168 Fehler) liegt. Weiterhin weist das Lernerkorpus wesentlich mehr Wortschreibungsfehler als syntaktische Orthographiefehler auf (das Verhältnis beträgt 92 zu 19), wohingegen im L1-Korpus die Relation umgekehrt (67 zu 101) ist.

In Hirschmann (i.Vorb.) werden verschiedene Fehlerkategorien qualitativ und quantitativ analysiert und lerntheoretische Erklärungsvorschläge unterbreitet. Studien wie diese zeigen, dass korpusbasierte (Fehler-)Untersuchungen zu neuen, differenzierten Ergebnissen führen können: Wortschreibungsfehler sind oft eigentlich phonologische oder morphologische Fehler (sehr oft Transferfehler), während kontextabhängige Fehler auf syntaktische Probleme hinweisen können. In der Vermittlung können solche Probleme dann gezielt besprochen werden.

4 Zusammenfassung

Kontrolliert erhobene und transparent annotierte Lernerkorpora werden in der Untersuchung der Lernersprache immer wichtiger (vgl. Fandrych/Tschirner 2007). Bestimmte komplexe Erwerbsphänomene, die quantitative und qualitative Merkmale von Lernern betreffen, können wahrscheinlich am besten anhand von Lernerkorpus-Studien erforscht werden. Die aus den Korpusstudien gewonnenen Hypothesen müssen dann oft in gezielten Erhebungen überprüft werden. Wir denken zum Beispiel an integrierte Registeruntersuchungen, die mehrere Eigenschaften der Lernersprache gleichzeitig betreffen (das oben beschriebene „mündliche" satzinitiale *und* kann dann mit weiteren Beobachtungen verbunden werden).

Das in diesem Artikel vorgestellte Lernerkorpus Falko kann aufgrund seiner Designkriterien und seiner vielschichtigen Annotation mit einer expliziten Zielhypothese für viele Zwecke genutzt werden. Das Korpus ist zwar noch klein, aber es wächst![1]

[1] Über Unterstützung bei der Datenerhebung würden wir uns sehr freuen.

Literatur

(Alle genannten URLs wurden am 23.11.2007 überprüft.)

Baumgarten, Nicole (2006): Konventionen der Kohäsionsbildung in deutschen und englischen Texten. AND und UND als makrosyntaktische Verknüpfung in populärwissenschaftlichen Zeitschriftenartikeln. In: D. Wolff (Hg.), Mehrsprachige Individuen – vielsprachige Gesellschaften. Frankfurt a.M., 133–153.

Bell, David (2007): Sentence-initial *and* and *but* in academic writing. In: Pragmatics 17/2, 183–201.

Belz, Judy A. (2004): Learner Corpus Analysis and the Development of Foreign Language Proficiency. In: System. An International Journal of Educational Technology and Applied Linguistics 32/4, 577–591.

Birkner, Karin et al. (1995): Der adversative Konnektor *aber* eines italienischen und zweier polnischer Lerner des Deutschen. In: B. Handwerker (Hg.), Fremde Sprache Deutsch. Tübingen, 65–118.

Cherubim, Dieter (Hg.) (1980): Fehlerlinguistik. Beiträge zum Problem der sprachlichen Abweichung. Tübingen.

Dagneaux, Estelle et al. (1996): Error Tagging Manual Version 1.1. Centre for English Corpus Linguistics, Université Catholique de Louvain. Louvain-la-Neuve.

Dietrich, Rainer (2006): Second Language Acquisition in the 20th Century. In: S. Auroux et al. (Hg.), History of the Language Sciences. Vol. 3. Berlin, 2705–2728.

Eisenberg, Peter (2004): Grundriß der deutschen Grammatik. Das Wort. 2. Aufl. Stuttgart/Weimar.

Ellis, Rod (1994): The Study of Second Language Acquisition. Oxford.

Fandrych, Christian / Tschirner, Erwin (2007): Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. In: DaF 4, 195–204.

Granger, Sylviane (2002): A bird's-eye view of learner corpus research. In: S. Granger / J. Hung / S. Petch-Tyson (Hg.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam/Philadelphia, 3–33.

Granger, Sylviane (2003): The International Corpus of Learner English. A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. In: TESOL Quarterly 37/3, 538–546.

Granger, Sylviane (i.Dr.): Learner Corpora. In: A. Lüdeling / M. Kytö (Hg.).

Granger, Sylviane / Dagneaux. Estelle / Meunier, Fanny (Hg.) (2002): The International Corpus of Learner English. Louvain-la-Neuve.

Hirschmann, Hagen (i.Vorb.): Orthographische Kompetenz zwischen Deutsch als L1 und Deutsch als L2.

Höhle, Tilman (1986): Der Begriff „Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. In: Akten des Siebten Internationalen Germanistenkongresses. Göttingen, 329–340.

Hunston, Susan (i.Dr.): Collection strategies and design decisions. In: A. Lüdeling / M. Kytö (Hg.).

Izumi, Emi et al. (2005): Error Annotation for a Corpus of Japanese Learner English. In: Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005). Jeju Island, Südkorea. ACL Anthology IOC 6009, 72–80 (online verfügbar unter: http://acl.ldc.upenn.edu/I/I05/I05-6009.pdf).

Lennon, Paul (1991): Error and the very advanced learner. In: International Review of Applied Linguistics 29/1, 31–44.

Lüdeling, Anke (2007): Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In: W. Kallmeyer / G. Zifonun (Hg.), Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Berlin / New York, 28–48 (Institut für Deutsche Sprache, Jahrbuch 2006).

Lüdeling, Anke (2008): Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: M. Walter / P. Grommes (Hg.) (2008b), 119–140.

Lüdeling, Anke / Kytö, Merja (Hg.) (i. Dr.): Corpus Linguistics. An International Handbook. Berlin / New York.

Lüdeling, Anke et al. (2005): Multi-level error annotation in learner corpora. In: Proceedings of Corpus Linguistics 2005, Birmingham (online verfügbar unter: http://www.corpus.bham.ac.uk/PCLC/).

Milde, Jan-Thorsten / Gut, Ulrike (2002): A prosodic corpus of non-native speech. In: B. Bel / I. Marlien (Hg.), Proceedings of the Speech Prosody 2002 Conference. Aix-en-Provence, 503–506.

Nesselhauf, Nadja (2004): Learner Corpora and their Potential in Language Teaching. In: J. Sinclair (Hg.), How to Use Corpora in Language Teaching. Amsterdam, 125–152.

Römer, Ute (i.Dr.): Corpora and Language Teaching. In: A. Lüdeling / M. Kytö (Hg.).

Rundell, Michael / Fox, Gwyneth (Hg.) (2007): Macmillan English Dictionary For Advanced Learners. Second edition. Oxford.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, 44–49 (erweiterte Fassung verfügbar unter: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf).

Schmid, Helmut (i.Dr.): Tokenizing and part-of-speech tagging. In: A. Lüdeling / M. Kytö (Hg.).

Schmidt, Thomas / Wörner, Kai (2005): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. In: Gesprächsforschung. Online-Zeitschrift zur verbalen Interaktion 6, 171–195 (online verfügbar unter: http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf).

Stern, Clara / Stern, William (1907): Die Kindersprache. Eine psychologische und sprachtheoretische Untersuchung. Leipzig.

von Stutterheim, Christiane / Carroll, Mary (2006): The Impact of Grammatical Temporal Categories on Ultimate Attainment in L2 Learning. In: H. Byrnes et al. (Hg.), Educating for Advanced Foreign Language Capacities. Washington, D.C., 40–53.

Walter, Maik / Grommes, Patrick (2008a): Die Entdeckung des fortgeschrittenen Lerners in der Varietätenlinguistik. In: M. Walter / P. Grommes (Hg.) (2008b), 3–27.

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

Walter, Maik / Schmidt, Karin (i.Dr.): *Und das ist auch gut so!* Der Gebrauch des satzinitialen *und* bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. In: B. Ahrenholz et al. (Hg.), Theorie und Empirie. Berlin.

Wegener, Heide (1995): Die Nominalflexion des Deutschen – verstanden als Lerngegenstand. Tübingen.

Weinberger, Ursula (2002): Error Analysis with Computer Learner Corpora. A corpus-based study of errors in the written German of British University Students. Lancaster University (MA thesis).