Erscheint in Patrick Grommes & Maik Walter (vorauss. 2007)
Fortgeschrittene Lernervarietäten. Niemeyer, Tübingen

Vorversion – kann sich noch ändern

Anke Lüdeling

Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora[1]

1. Einleitung

Auch fortgeschrittene Lerner des Deutschen als Fremdsprache haben zuweilen Probleme bei der korrekten Setzung des Artikels. Diese Beobachtung wird häufig gemacht, kann aber nur korpusbasiert, d.h. auf Grundlage einer systematischen Sammlung von Lernerdaten, präzisiert und quantifiziert werden. Dies könnte durch explizite Fehlerannotation erfolgen, d.h., indem man in einem Korpus mit Lerneräußerung alle Artikelfehler markiert. In der Lerneräußerung[2] in (1) zum Beispiel fehlt ein definiter Artikel vor *Kunstmärchen* (in dem Text handelt es sich um eine Diskussion der Gattung Kunstmärchen). Man könnte also hier ein Fehlertag ‚fehlender definiter Artikel' hinzufügen.

(1) *Kunstmärchen ist weiterhin abhängig vom Volksmärchen.* (Falko, Text 48)

Wenn man alle solchen Fehler getaggt hat, könnte man die Fehlertags quantitativ auswerten. Das Ergebnis könnte sein:

> A: Der Lernertext, aus dem (1) stammt, enthält 10 Artikelfehler (sämtlich fehlende definite Artikel) in 17 Sätzen.
>
> B: Der Lernertext, aus dem (1) stammt, enthält 3 Artikelfehler (sämtlich fehlende definite Artikel) in 17 Sätzen.

Beide Aussagen treffen zu: Der Text enthält gleichzeitig 10 Artikelfehler und 3 Artikelfehler. Die Zählung der Fehler ist nicht objektiv, sie hängt davon ab, was man als ‚korrekt' ansehen würde, d.h., gegen welchen Standard man evaluiert und wie man die Fehler klassifiziert.

In meinem Beitrag möchte ich mich nicht mit der Artikelsetzung bei Lernern beschäftigen (siehe dazu zum Beispiel die Beiträge von Bongartz und Lippert in diesem Band), sondern mit dem sich aus der unterschiedlichen Analyse ergebenden methodologischen Problem der Korpusarbeit: den unterschiedlichen Zielhypothesen und dem Fehlertagging, das heißt der expliziten oder impliziten Auszeichnung von Fehlern in Lernertexten. Diese Schwierigkeiten werden zwar im Prinzip in fast jeder Arbeit zur Fehleranalyse angemerkt, aber es fehlen Studien, die sagen, wie groß die Effekte sind. Dazu habe ich eine kleine Fragebogenstudie durchgeführt, die ich in Abschnitt 4 diskutiere – dort wird auch Beispiel (1) genauer erläutert. Davor werde ich in Abschnitt 2 Lernerkorpora einführen und das

[1] Für die Mitarbeit bei der Fragebogenstudie möchte ich mich herzlich bedanken bei Grit Mehlhorn, Max Möller, Sabine Schmidt, Anne Thyrolf und Katharina Wieland. Ohne sie und die ‚Falkos', insbesondere Karin Schmidt, Peter Siemen und Maik Walter hätte ich diesen Aufsatz nicht schreiben können. Vielen Dank auch an einen anonymen Gutachter für wertvolle Hinweise.

[2] Alle Lerneräußerungen sind aus dem Falko-Korpus entnommen, das in Abschnitt 2.4 vorgestellt wird. Die meisten Lerneräußerungen enthalten mehrere Fehler.

fehlerannotierte Lernerkorpus Falko vorstellen. In Abschnitt 3 gehe ich genauer auf die Evaluierung von Taggingverfahren ein. Im letzten Abschnitt möchte ich dann mögliche Konsequenzen für die Fehleranalyse in Korpora erörtern.

2. Lernerkorpora & Fehlertagging

Um Spracherwerbsverläufe zu dokumentieren und Erwerbsprobleme zu entdecken kann man im Grunde nur auf experimentell erhobene Daten und auf systematisch erhobene ‚authentische' Lerneräußerungen zurückgreifen. In Studien zum Erst- oder Zweitspracherwerb wird daher schon seit langem korpusbasiert gearbeitet (siehe Diessel, erscheint, für einen Überblick über Korpora im Erstspracherwerb), wobei die ersten Korpora natürlich nicht elektronisch vorlagen (so die berühmte Tagebuchstudie von Stern & Stern 1907). Während im Erstspracherwerb bereits relativ früh auch elektronische Korpora eine Rolle spielten und sich ein *de facto* Korpusstandard herausgebildet hat (das CHILDES-System, siehe MacWhinney & Snow 1985 und MacWhinney 1996)[3], ist die Entwicklung von Lernerkorpora des Zweit- oder Fremdspracherwerbs noch nicht so einheitlich. Für das Englische liegen bereits relativ große Lernerkorpora vor (Überblicksdarstellungen finden sich zum Beispiel in Pravec 2002, Tono 2003, Nesselhauf 2004, Römer 2006 und Granger, erscheint); für viele andere Sprachen gibt es solche nur in deutlich kleinerem Umfang. Die bisherigen Lernerkorpora sind in unterschiedlichen Formaten und Architekturen gespeichert und die Diskussion über gemeinsame Standards hält an.

In der Forschung und Vermittlung des Deutschen als Fremdsprache werden Korpora generell (also auch Muttersprachlerkorpora) bisher wenig eingesetzt. Es gibt kaum frei zugängliche Lernerkorpora.[4] Ausnahmen sind die ESF-Korpora[5] mit gesprochenen Daten aus dem ungesteuerten Zweitspracherwerb und das LeaP-Korpus[6], das gesprochene Daten von sehr fortgeschrittenen Lernern des Deutschen enthält, das tief annotiert ist und für prosodische und phonologische Untersuchungen genutzt werden kann (siehe Milde & Gut 2002; Gut, in diesem Band).

---

[3] CHILDES steht für Child Language Exchange System und bietet eine Architektur für das Speichern von Korpusdaten und eine Reihe von Auswertungstools. Es wird für einen Großteil der Erstspracherwerbsstudien eingesetzt, so dass ein einfacher Austausch der Daten möglich ist. Erstspracherwerbsdaten sind zumeist Dialogdaten (zwischen Kind und Elternteil oder Interviewer) – CHILDES ist also für solche Daten optimiert. Leider sind die CHILDES-Daten nicht TEI-konform und nicht in XML kodiert, das heißt, dass viele Such- und Auswertungsprogramme nicht damit umgehen können. Außerdem sind die meisten Daten in Lernerkorpora für Zweit- oder Fremdspracherwerb sind Textdaten – hier kann man andere Standards und Tools verwenden.

[4] Neben den hier besprochenen öffentlich zugänglichen Lernerkorpora gibt es sicher sehr viele (und zum Teil sicher auch umfangreiche) eher private Lernerdatensammlungen an Sprachvermittlungsinstitutionen und Universitäten. Belz (2004) beschreibt ein Korpus mit Daten, die in Email-Diskursen zwischen Deutschlernern und Englischlernern entstanden sind, das aber nicht öffentlich verfügbar ist. Die Daten aus Weinberger (2002, in diesem Band) – Essays von Lernern mit englischer Muttersprache – stehen bisher nicht zur Verfügung, sollen aber in Kürze über die Falko-Webseite öffentlich gemacht werden.

[5] http://www.mpi.nl/world/data/esf_archive/html/esf.html.

[6] http://www.phonetik.uni-freiburg.de/leap/corpus.html.

Für die Auswertung von Lernerkorpora können alle qualitativen und quantitativen Verfahren der Korpusauswertung genutzt werden (siehe zum Beispiel Biber, Conrad & Reppen 1998, Manning & Schütze 1999, Lemnitzer & Zinsmeister 2006). Man kann zum einen den ‚nackten' Text nach bestimmten Wörtern oder Konstruktionen durchsuchen und statistische Vergleiche mit Muttersprachlerdaten oder kontrastiven Lernerdaten (zum Beispiel Daten von Lernern mit einer anderen Muttersprache oder einem anderen Sprachstand) vergleichen, um bestimmte Charakteristika herauszuarbeiten. Man kann den Text auch mit weiteren Informationen annotieren, wie zum Beispiel Wortarten und Fehlerklassen (siehe Abschnitt 2.2). Diese kann man dann genauso wie die Wortformen quantitativ auswerten. Der statistische Vergleich von Lernerkorpora mit anderen Korpora wird oft Contrastive Interlanguage Analysis (CIA) oder kontrastive Analyse genannt.[7] Ein wesentliches Ziel der CIA ist dabei die Ermittlung von Wörtern oder Konstruktionen, die von Lernern im Vergleich zu Muttersprachlern (oder Lernern mit einer anderen Muttersprache) zu selten (underuse) oder zu häufig (overuse) gebraucht werden.
Damit ein Lernerkorpus nützlich ist, muss es im Design, in der Dokumentation und Annotation und in der Kodierung und Veröffentlichung gewisse Kriterien erfüllen, die ich in den Abschnitten 2.1-2.3 erläutern möchte. In Abschnitt 2.4 stelle ich dann anhand der aufgestellten Kriterien das Lernerkorpus Falko vor.

2.1 Design

Granger (2002, 7) definiert Lernerkorpora wie folgt: "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".[8]
Die Zusammenstellung von Lernertexten zu einem Korpus ist also abhängig von vorher festgelegten Kriterien, die wiederum von den Fragestellungen bestimmt werden, für die das Korpus als Datengrundlage dienen soll (siehe Tono 2003). Die wesentlichen Parameter sind laut Granger (2002, erscheint) die Menge der vertreten Muttersprachen (sie unterscheidet, etwas verwirrend, zwischen einsprachigen und mehrsprachigen Korpora), die Aufgabenstellung und Umstände der Aufnahme, der Sprachstand der Lerner, der Kommunikationsmodus (gesprochen oder geschrieben), sowie die zeitliche Dimension (Longitudinalkorpora in denen Lerner zu verschiedenen Zeitpunkten aufgenommen werden vs. Querschnittskorpora).

---

[7] Die Begriffe gehen manchmal etwas durcheinander. Zunächst bezeichnete man mit ‚kontrastive Analyse' den Vergleich von Lernern und Muttersprachlern in einer bestimmten Methodik. Granger (1996) führt den Begriff ‚Contrastive Interlanguage Analysis' ein, in der unterschiedliche Interimssprachen miteinander verglichen werden können – also auch zum Beispiel Interimssprachen von Lernergruppen mit unterschiedlichen L1 oder unterschiedlichen Lernständen. Heute wird ‚kontrastive Methode' zuweilen auch als Übersetzung von CIA verwendet. Die korpusbasierte CIA unterscheidet sich in Grundannahmen und Methoden von der ‚traditionellen' kontrastiven Analyse, wie sie zu Recht oft kritisiert worden ist (siehe zum Beispiel Granger 1996, erscheint und Maden-Weinberger, in diesem Band).

[8] FL steht für foreign language, SL für second language, SLA für second language acquisition und FLT für foreign language teaching.

Eine weitere Designentscheidung ist, wie stark die Datenerhebung kontrolliert wird. Dürfen die Lerner Hilfsmittel nutzen oder die Texte zu Hause produzieren? Werden die Daten digital oder handschriftlich erhoben? Während eine digitale Erhebung für die Korpuserstellung natürlich weniger aufwändig ist, kann dabei in vielen Situationen nicht ausgeschlossen werden, dass automatische Rechtschreibkontrollen oder Online-Ressourcen benutzt werden. Handschriftlich erhobene Texte hingegen müssen manuell digitalisiert werden.

2.2 Vorverarbeitung

Zur Vorverarbeitung eines Korpus gehören neben der eventuellen Digitalisierung nicht digital vorliegender Daten (zum Beispiel handschriftlich vorliegende Daten) die Tokenisierung sowie die Hinzufügung von Headerdaten, struktureller und positioneller Annotation. Alle Entscheidungen, die hier getroffen werden, wirken auf die späteren Analysemöglichkeiten ein.
Bereits bei der Digitalisierung von handschriftlichen Daten müssen Entscheidungen getroffen werden. Ist das dritte Wort in Abb. 1 ‚unterscheiden'? Oder ‚umterscheiden'? Oder noch etwas anderes?

Abb. 1: Ausschnitt aus den Vorlagen für das Lernerkorpus Falko

Wenn die Texte digital vorliegen, müssen sie weiterverarbeitet werden. Bestimmte Fragestellungen – alle, die auf die Binnenstruktur des Korpus zugreifen, also zum Beispiel alle kontrastiven Untersuchungen zu Lernern verschiedener Muttersprachen oder verschiedener Sprachniveaus – können in einem Korpus nur bearbeitet werden, wenn Daten über die im Korpus vorhandenen Texte vorliegen (so genannte Headerdaten). In Lernerkorpora müssen also strukturiert detaillierte Angaben zu allen Designparametern (den Lernern, ihrer Erwerbsbiographie, der Erhebungssituation etc.) sowie zur Aufbereitung erhoben werden. Falls das Korpus weiter annotiert ist, gehört eine Dokumentation der verwendeten Tags, der Vergaberichtlinien und idealer Weise auch eine Evaluation der Annotation dazu.[9] Wie wichtig dieser Punkt für Lernerkorpora ist, soll in den Abschnitten 3 und 4 deutlich werden. Bisher gibt es meines Wissens kein Lernerkorpus, für das eine solche Evaluation vorgenommen wurde.
Oft ist es wünschenswert, strukturelle Eigenschaften der Texte wie zum Beispiel die Unterteilung von Texten in ihre Binnenstruktur (Überschriften, Aufgabentexte etc.) zu markieren, da man für die Auswertung systematische Unterscheidungen treffen möchte. Dies ist nur in einigen Korpusarchitekturen einfach möglich (siehe Abschnitt 2.3). Die meisten großen Korpora sind mit Wortarten und Lemmanamen getaggt – dies ist hilfreich, um bei Auswertungen den Suchraum einzuschränken und um bestimmte Muster zu finden.

[9] Für viele Tagsets fehlt eine ausführliche Dokumentation. Wie eine gute Dokumentation für ein Tagset aussehen kann, wird in Johansson et al. (1986) vorgeführt.

Hierzu werden statistische Tagger verwendet (Manning & Schütze 1999). Da die meisten Wortarttagger auf Zeitungskorpora trainiert werden und ihre Qualität nachlässt, wenn sie auf Texte angewendet werden, die weniger standardisiert sind, wie zum Beispiel Lernertexte, muss man bei hier eine Evaluation vornehmen (Meunier 1998, van Rooy & Schäfer 2003; siehe auch Pankow & Pettersson 2006 für eine Auswertung der Qualität von Wortarttaggern für Korpora gesprochener Sprache).
Die wichtigste Art der Annotation von Lernertexten ist die Fehlerannotation. Dabei werden Tokens oder Tokenfolgen systematisch mit Fehlerkategorien versehen. Wie bei allen anderen Annotationsverfahren müssen auch hier Tagset und Annotationsrichtlinien festgelegt werden. Das Annotieren von Fehlern erfolgt heute fast immer manuell (auch wenn es einige Arbeiten zur automatischen Fehlererkennung gibt, siehe zum Beispiel Menzel & Schröder 1998, Reuer & Kühnberger 2005, Arrieta et al. 2003, das Flag-Projekt[10]).

2.3 Kodierung

Wenn das Design und die Annotationsebenen festgelegt sind, muss eine Kodierung und Architektur gewählt werden. Ein Desiderat der empirischen Arbeit ist die prinzipielle Reproduzierbarkeit der Ergebnisse. Dazu gehört, dass die Datengrundlage allgemein zur Verfügung steht. Für Korpora bedeutet das, dass sie entweder über eine Webschnittstelle abfragbar sein müssen oder in einer Korpusversion inklusive Abfragewerkzeugen distribuiert werden. Dies ist umso einfacher (da allgemein verfügbare Architekturen und Werkzeuge genutzt werden können), je mehr die Kodierung allgemeinen Korpusstandards (zum Beispiel den Standards der Text Encoding Initiative[11]) und Kodierungsstandards (zum Beispiel XML) genügt.
Große standardisierte Korpora sind meistens in ‚flachen' Dateien gespeichert, d.h., die Texte und die Annotationen befinden sich in derselben Datei, entweder in einem Tabellenformat oder in einem (XML)-Baumformat (Lüdeling et al. 2005). Solche Formate sind schnell indizier- und damit durchsuchbar. Für kleinere Korpora wie Lernerkorpora spielt die Optimierung der Geschwindigkeit keine Rolle, viel wichtiger ist hingegen eine Architektur, in der verschiedene Annotationsebenen abgelegt werden können. Solche so genannten Mehrebenen-Architekturen stehen seit einigen Jahren zur Verfügung. Sie wurden für multimodale Korpora entwickelt, in denen Sprachsignal, Transkription, Bildmaterial (z.B. Videos der begleitenden Gestik) und Annotationsebenen gemeinsam gespeichert werden sollen (Carletta et al. 2003, Wörner et al. 2006). In Abschnitt 5 werde ich dafür argumentieren, Lernerkorpora in Mehrebenen-Architekturen abzulegen.
Die Veröffentlichung aller Daten inklusive der Dokumentation der Vorverarbeitungsschritte ist umso wichtiger, je kontroverser die Ergebnisse sind und je subtiler die Untersuchungen sind, die anhand der Daten durchgeführt werden sollen.

[10] http://flag.dfki.de/
[11] http://www.tei-c.org/

Erscheint in Patrick Grommes & Maik Walter (vorauss. 2007)
Fortgeschrittene Lernervarietäten. Niemeyer, Tübingen

2. 4 Das Lernerkorpus Falko

Da es, wie erwähnt, für das Deutsche bisher kaum frei verfügbare fehlerannotierte Lernerkorpora gibt, wird zur Zeit an der Humboldt-Universität zu Berlin in Kooperation mit der Freien Universität Berlin das Korpus **f**ehler**a**nnotierte **L**erner**ko**rpus Falko entwickelt. Falko ist bisher noch klein, wächst aber ständig und steht jederzeit online über ein Webinterface zur Verfügung.[12] Falko enthält schriftlich produzierte Texte von fortgeschrittenen Lernern des Deutschen als Fremdsprache. Es besteht aus mehreren Subkorpora mit unterschiedlichen Textsorten. Alle Texte haben bestimmte Headerdaten zur Muttersprache, Alter, Geschlecht etc. des Lerners und zur Aufgabenstellung. Die Texte sind automatisch mit dem DecisionTreeTagger (Schmid 1994) mit Wortarten annotiert (wie oben erwähnt, ist die Qualität des Wortarttaggens bei Lernerdaten schlechter als bei Zeitungskorpora). Einige Subkorpora sind außerdem mit weiteren Annotationsebenen (beispielsweise zum Ausdruck der Definitheit oder zur Ausgestaltung der syntaktischen/topologischen Felder) versehen; auch in diesem Bereich wird das Korpus ständig erweitert.

Am besten annotiert ist ein Teilkorpus mit Zusammenfassungen von linguistischen oder literaturwissenschaftlichen Fachtexten[13], die von nichtmuttersprachlichen Studierenden der Freien Universität Berlin im Rahmen einer Sprachstandsüberprüfung zum Zeitpunkt der Zwischenprüfung produziert wurden. Die Erhebung ist stark kontrolliert: die Vorlagentexte sind bekannt, die Texte werden handschriftlich in einer Klausursituation erhoben, keinerlei Hilfsmittel sind zugelassen. Wie in der Einleitung zu diesem Buch erläutert, ist die Einstufung von Lernern in Fortgeschrittenheitsklassen nicht einfach, da sprachexterne oder sprachinterne Kriterien herangezogen werden können. Für das Zusammenfassungskorpus berufen wir uns auf externe Kriterien: die Lerner haben die DSH-Prüfung abgelegt und ein Grundstudium der Germanistik an der Freien Universität absolviert oder wurden im Hauptstudium eingestuft. Die Texte sind aber in ihrer Qualität sehr unterschiedlich! Es gibt außerdem zwei weitere kleine Teilkorpora mit Texten, die nach der gleichen Aufgabenstellung erstellt wurden, nämlich ein weiteres Lernerkorpus von dänischen Lernern, das in Kopenhagen erhoben wurde und ein Muttersprachlerkorpus. Bei allen Subkorpora, die Zusammenfassungen enthalten, wurde die automatische Wortartannotation manuell korrigiert und eine Zielhypothese (siehe unten) eingefügt. Außerdem sind sie auf mehreren Ebenen fehlerannotiert. Das Zusammenfassungskorpus wurde bereits in mehreren Studien verwendet, unter anderem in Hirschmann (2005), Lippert (2005, in diesem Band).

Um andere linguistische Fragestellungen bearbeiten zu können, gibt es ein weiteres Teilkorpus mit Essays. Die Lerner können dabei (ebenfalls in stark kontrollierten Situationen) aus mehreren kontroversen Themen[14] auswählen. Die Texte im Essaykorpus sind bisher noch nicht weiter annotiert, werden aber im Moment bearbeitet. Die Lerner, die die Texte für das Essaykorpus beisteuern, sind viel heterogener als die oben angesprochenen Lerner. Daher ist es hier nicht möglich, Fortgeschrittenheit nach formalen

---

[12] http://www.linguistik.hu-berlin.de/korpuslinguistik/projekte/falko/index.php

[13] In einigen bisherigen Publikationen wie Lüdeling et al. (2005) Kernkorpus genannt.

[14] Die Themen sind angelehnt an die für das ICLE (International Corpus of Learner English) vorgegebenen Essaythemen, siehe http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm.

Kriterien wie Einstufung im Studium festzustellen. Wir erheben deshalb von jedem Lerner einen standardisierten C-Test und geben das Testergebnis im Header an.
Zusätzlich gibt es weitere Teilkorpora, wie zum Beispiel ein Longitudinalkorpus mit Texten, die an der Georgetown University erhoben wurden.
Die Korpusarchitektur ist in Lüdeling et al. (2005) und Lüdeling (2006) beschrieben, das Suchinterface in Siemen et al. (2006).

3 Annotationsevaluation

Um die Qualität von automatischer Annotation zu bewerten und um die Möglichkeiten manueller Annotation einschätzen zu können, werden in der Computerlinguistik und in der Korpuslinguistik verschiedene Evaluationsverfahren eingesetzt. Bei automatischen Annotationsverfahren wird dabei meist ein so genannter Goldstandard erstellt, das ist ein manuell annotiertes Teilkorpus von hoher Qualität. Dasselbe Teilkorpus wird automatisch (oder manuell) annotiert und dann werden beide Versionen verglichen. Je genauer die Annotation dem Goldstandard folgt, desto besser ist sie. Eine andere Möglichkeit der Evaluation ist die Ermittlung der Urteilerübereinstimmung (auch Inter Annotator Agreement oder Inter Rater Reliability genannt), die angewendet wird, wenn kein Goldstandard existiert; zumeist um manuelle Analysen zu vergleichen. Hier wird derselbe Text von zwei oder mehreren Annotatoren bearbeitet. Dann wird ermittelt, inwieweit die Annotatoren in ihrer Analyse übereinstimmen.[15] Während bei einigen Annotationsaufgaben hohe Übereinstimmung erzielt werden kann, gibt es auf vielen linguistischen Ebenen Schwierigkeiten – nicht überraschend sind das natürlich diejenigen Gebiete, für die in der Linguistik selbst keine einheitliche Kategorisierung vorgeschlagen werden konnte, z.B. das Lesartentagging (Kilgarriff 1998, Veronis 2001).

In fast allen Arbeiten zur Fehleranalyse bei Lernertexten wird angemerkt, dass es große Interpretationsspielräume gebe (Corder 1973, 1981, Cherubim 1980, Lennon 1991, Dagneau, Denness & Granger 1998). Die Forschung, die sich auf Lernerkorpora stützt, muss sich also diesem Problem stellen. Wünschenswert wäre eine hohe Urteilerübereinstimmung in der Fehleranalyse – die kleine Studie in Abschnitt 4 zeigt jedoch, wie wenig dieses Desiderat erreicht werden kann. Sehr oft wird das Problem in der korpuslinguistischen Forschung ignoriert (so zum Beispiel in Abe 2003, Arrieta et al. 2003, Garnier et al. 2003, Izumi et al. 2003, Granger, erscheint) oder heruntergespielt. In Abschnitt 5 will ich dann kurz erläutern, wie man mit in modernen Korpusarchitekturen mit dem Problem der fehlenden Urteilerübereinstimmung umgehen kann.
Fehlerzählungen können durch zwei Faktoren beeinflusst werden, nämlich durch (a) die Kategorisierung und (b) die Zielhypothese. Die Faktoren sind nicht unabhängig voneinander; ich möchte sie im Folgenden dennoch so getrennt wie möglich betrachten, um ihren jeweiligen Einfluss herauszuarbeiten. Die Kategorisierung, d.h., die Erstellung von Tagsets und Vergaberichtlinien ist in vielen Arbeiten diskutiert worden (siehe z.B.

[15] Für zwei Annotatoren rechnet man den sogenannten κ-Wert aus, siehe Carletta (1996), Jurafsky & Martin (2000). Bei mehr Annotatoren wird es etwas schwieriger, siehe Bortz, Lienert, & Boehnke (2000). Für meine Studie konnte ich keine Werte ausrechnen, da die Ergebnisse keine einfache Kategorisierung darstellen.

Granger 2002, Tono 2003). Hier müssen Entscheidungen zur Granularität (will man ‚general' Tagsets wie Izumi et al. (2003) vorschlagen oder sehr detaillierte wie in Lippert, in diesem Band?) genauso getroffen werden wie Entscheidungen darüber, ob man linguistische Ebenen (Morphologie, Orthographie, Syntax etc.) trennen möchte oder die formale Art des Fehlers (z.B. Auslassung, Einfügung, Vertauschung; siehe auch Abschnitt 4). Die Entscheidungen über alle diese Fragen hängen sicher im Wesentlichen von der jeweiligen Forschungsfrage ab.

Ein Goldstandard kann nur erstellt werden, wenn vorher ein Tagset und Vergaberichtlinien festgelegt sind. Dies wäre im Prinzip für Fehlerannotationen in Lernerkorpora genauso möglich wie für andere Kategorisierungen, allerdings ist im Unterschied zu anderen Annotationsaufgaben die Kategorisierung abhängig von der jeweiligen Zielhypothese. Der Faktor ‚Zielhypothese' ist bisher in der korpuslinguistischen Literatur weniger diskutiert worden. Man kann einen Fehler nur identifizieren, wenn man die Lerneräußerung mit einer angenommen ‚korrekten' Äußerung vergleicht. Laut Ellis (1994: 54) ist die Zielhypothese von Lerneräußerungen die "reconstruction of those utterances in the target language". Auch wenn in einigen Lernerkorpora die Zielhypothese explizit angegeben wird (wie z.B. bei Izumi et al. 2005; in den meisten ist sie implizit), wird bisher kaum diskutiert, wie stark sich Zielhypothesen unterscheiden und so die Zählung beeinflussen können. Man hat also mehrere Evaluationsaufgaben:

(a) Übereinstimmung in der Kategorisierung (Festlegung von Tagset und Vergaberichtlinien)
(b) Urteilerübereinstimmung in der Annotation nach Festlegung eines Tagsets
(c) Urteilerübereinstimmung in der Zielhypothese
(d) Urteilerübereinstimmung in der Kategorisierung in Abhängigkeit von der Zielhypothese.

4. Befragung zur Fehlerkategorisierung und Zielhypothese

Um herauszufinden, wie stark sich die unterschiedlichen Kategorisierungen und Zielhypothesen tatsächlich in der Zählung von Lernerfehlern auswirken, habe ich fünf erfahrene DaF-LehrerInnen in einer kleinen Befragung gebeten, Lernertexte zu annotieren.[16] Die TeilnehmerInnen hatten beliebig viel Zeit, die Texte zu bearbeiten. Der Fragebogen hat zwei Teile. Im ersten Teil wurde ein Lernertext (Falko L2, Text 80, eine Zusammenfassung eines literaturwissenschaftlichen Texts) präsentiert, zu dem Fehler markiert werden sollten – hier wollte ich eruieren, ob sich aus den Fehlertags eine implizite Zielhypothese ablesen lässt, ob ungefähr die gleichen Fehler angestrichen würden und welche Fehler in einer Fehlerklasse zusammengefasst werden.

Im zweiten Teil sollten zu einem anderen Lernertext (Falko L2, Text 48, ebenfalls eine Zusammenfassung eines literaturwissenschaftlichen Texts) explizite Zielhypothesen

[16] Die genaue Aufgabenstellung, alle Daten und die Auswertungen sind online unter http://www.linguistik.hu-berlin.de/korpuslinguistik/projekte/falko/index.php abrufbar.
Alle Teilnehmer haben mehrjährige Erfahrung im DaF-Unterricht. Ich möchte betonen, dass meine Problematisierung der Uneinheitlichkeit der Kategorisierung ein allgemeines Problem beschreibt, das nicht zu umgehen ist, und nicht als eine Kritik an den Annotatoren hier aufgefasst werden soll!

angegeben werden. Das Ziel war, herauszufinden, ob und wie stark sich auch explizite Zielhypothesen unterscheiden.

4.1 Fehlertags

Im ersten Teil des Fragebogens wurden die Teilnehmer gebeten, in einem Lernertext alle Fehler so zu markieren, wie sie es normalerweise in einer Klausur tun würden. Die Teilnehmer mussten also sowohl das Tagset selber wählen und erläutern als auch entscheiden, welche Fehler sie anstreichen wollten. Dies ist eine informelle Variante des Vorgehens, das in vielen Lernerkorpora angewendet wird – bei der Korpuserstellung wird ein Tagset mit Vergaberichtlinien erarbeitet .

Ich möchte anhand einiger Beispiele zeigen, dass sich die verwendeten Tagsets, ihre Granularität und die zugrundeliegenden Fehlerklassen deutlich voneinander unterscheiden. Beispiel (2) zeigt, dass die Fehlerzählung (zwischen 1 und 4 Fehlern) und die Fehlerkategorisierung schon voneinander abweichen. Annotator 5 ist sehr explizit, wahrend Annotator 1 recht unspezifisch bleibt. Ohne weitere Markierung könnte man aus der Angabe von Annotator 1 zum Beispiel nicht erkennen, wo genau der Fehler liegt.

(2) *Es gab enger Zusammenhang zwischen Wissenschaft vom Menschen und der Poesie.*

Annotator 1	GR (Grammatik)
Annotator 2	+ (Morphologiefehler (z.B. falsches Genus, falsche Flexion)) S (Syntaxfehler (Wortstellungsfehler; fehlendes oder überflüssiges Wort; unvollständiger Satz))
Annotator 3	Gr √ Art. (Grammatik, fehlender Artikel), Gr Dekl. (Grammatik Deklination), Gr √ Art.
Annotator 4	Art (fehlender oder falscher Artikel)
Annotator 5	V (Wort fehlt) GR (Grammatik) V GR

Tabelle 1: Fehlerauszeichnung zu Satz (2)

In Beispiel (3) wird von vier der fünf Annotatoren[17] die fehlende Verbpartikel angestrichen (da ich in diesem Teil des Fragebogens nicht nach einer expliziten Zielhypothese gefragt hatte, muss ich dies in einigen Fällen interpretieren; einige Annotatoren haben die Partikel *ab* direkt eingefügt). Allerdings wird dies von manchen Annotatoren als Lexikonfehler angesehen und von anderen als Syntax- oder Grammatikfehler (die Erläuterung hinter dem Zeichen ist von den Annotatoren selbst gegeben).

[17] Annotator 5 hat eine explizite Zielhypothese angegeben und den Satz folgendermaßen korrigiert (hier wären also zwei Fehler): *Das hängt vor allem mit der Kultur der Neuzeit zusammen.*

(3) *Das hängt vor allem von der Kultur der Neuzeit.* (Falko L2, Text 80)

Annotator 1	GR (Grammatik)
Annotator 2	S (Syntaxfehler (Wortstellungsfehler; fehlendes oder überflüssiges Wort; unvollständiger Satz))
Annotator 3	A (Ausdruck oder Wortwahl)
Annotator 4	W (fehlendes Wort (auch Verbpartikel) oder falsches Wort)

Tabelle 2: Fehlerauszeichnung zu Satz (3)

Beispiel (3) zeigt, dass derselbe Fehler mit derselben impliziten oder expliziten Zielhypothese unterschiedlich bewertet wird. Nun könnte man vermuten, dass die Annotatoren vielleicht nur verschiedene Namen vergeben hätten, aber zugrunde liegend ähnliche Fehlerklassen annehmen – daher ist hier interessant, was genau die Fehlerklassen (zum Beispiel GR oder A) bedeuten, d.h., welche Fehler jeweils unter einem Tag zusammengefasst wurden. Wenn man sich die Fehlerklassen anschaut, sieht man aber, dass jeweils ganz unterschiedliche Fehler in eine Klasse fallen.
Ich möchte dies in Tabelle 3 exemplarisch für die Annotatoren 2 und 3 und die Fehlerklasse zeigen, die für die fehlende Partikel in (3) vergeben wurde (S und A). Beide Annotatoren haben ihre Kategorisierung im Text dokumentiert. Ich habe jeweils die Stellen angestrichen, die explizit als S oder A markiert waren und in eckigen Klammern in recte die angegebene Korrektur eingefügt; alle anderen Fehler und Fehlertags wurden ignoriert (zum Teil wurden von den Annotatoren die gleichen Fehler markiert, aber anders bezeichnet). Man sieht, dass sich S und A nur in Satz (3) treffen. In allen anderen Fällen wurden unterschiedliche Fehler ausgezeichnet.

Annotator 2, Tag S	Annotator 3, Tag A
Nachwievor gilt der Traum in der Literatur und Wissenschaft [als] *ein wenig erforschtes Gebiet.*	
Das hängt vor allem von der Kultur der Neuzeit [ab].	*Das hängt vor allem von der Kultur der Neuzeit* [ab].
Der Traum wurde schon sehr früh, seit der Renaissancezeit unter verschiedenen Gesichtspunkten erforscht und beschrieben. [zu streichen: ‚seit der Renaissancezeit]	
	In der Zeit der Renaissance war der Traum auch ein herrschendes [besser: ‚beherrschendes, vorherrschendes'] *Thema auf der Theaterbühne.*
Im Theater wurde der Traum als etwas Geheimnisvolles [Verb!], *das voller Gelehrsamkeit* [war] *und etwas Unabhängiges vom Körper und von der Seele hatte.* [Wortstellung: ;… vom Körper und von der Seele Unabhängiges hatte']	*Im Theater wurde der Traum als etwas Geheimnisvolles, das voller Gelehrsamkeit und etwas Unabhängiges vom Körper und von der Seele hatte.* [umformulieren: ‚… und unabhängig vom Körper und von der Seele war']
Es wurde in der Literatur die Sprachformen,	

Symbole und Zeichen angedeutet und gepflegt, damit das Seelenleben von Menschen während des Traumes deutlich zu machen. [Wortstellung: ‚In der Literatur wurden die …'] [zu streichen: ‚, damit', einzufügen ‚um']	
Auch in der literarischen Kulturgeschichte erweckte der Traum die Interesse, der die Poesie und [das] *Wissen beeinflusste.*	
In der literarischen Kulturgeschichte wurde der Traum als etwas Kompliziertes [Verb!]*, das die Wechselwirkungen des Wissens und der poetischen Einbildungskraft spiegelte.*	
Dazu hatte man in [der] *literarischen Kulturgeschichte den Traum als ein ästhetisches und intellekturelles Beispiel, gesehen.*	
In der abendländischen Geschichte der Moderne wurde die Literatur mit den Ordnungen der Gelehrsamkeit eng verbunden. [Wortstellung: ‚… eng mit den Ordnungen der Gelehrsamkeit verbunden']	*In der abendländischen Geschichte der Moderne wurde* [besser: ‚war'] *die Literatur mit den Ordnungen der Gelehrsamkeit eng verbunden.*
Was in der Zeit der Antike, des Renaissancehumanismus und der Aufklärung das Gegenteil war. [diesen und den vorigen Satz verbinden]	
	Die Poesie wurde von den gelehrten Diskursen der [besser: ‚an den'] *Universitäten und von den menschlichen Wissen über Natur und Vernunft beeinflusst.*
Es gab [einen] *enger Zusammenhang zwischen* [der] *Wissenschaft vom Menschen und der Poesie.*	
	In der Zeit der Moderne wurden ganze [besser: ‚alle, viele'] *Kenntnisse über das Individuum für* [besser: ‚in'] *die ästhetische Praxis übernommen.*

Tabelle 3: Fehlerkategorien A und S gegenübergestellt

Die Ergebnisse sind wie erwartet: unterschiedliche Annotatoren vergeben verschiedene Fehlertags für die gleichen Fehler und – hier wichtiger – fassen unterschiedliche Fehler in einer Kategorie zusammen. Wenn man dieses (eigentlich triviale) Ergebnis auf die Auswertung von Fehlern (in Lernerkorpora oder auf irgendeine andere Art) überträgt, muss man genauso davon ausgehen, dass alle Ergebnisse, die auf einer Fehlerkategorisierung beruhen, zumindest transparent gemacht werden sollten.

4.2 Zielhypothesen

Im zweiten Teil der Befragung wurden die Annotatoren gebeten, Lernersätze zu verbessern, ohne explizit die Fehler anzustreichen. Hier wollte ich herausfinden, inwiefern sich Zielhypothesen unterscheiden können. Der zu bearbeitende Text hat 17 Sätze und ist eine Zusammenfassung eines literaturwissenschaftlichen Texts. Die Ergebnisse wurden folgendermaßen ausgewertet: zunächst habe ich für jeden Lernersatz alle Korrekturen aufgelistet und zwar so, dass alle Elemente des Lernertexts, die auch in der Korrektur vorkommen, untereinander geschrieben werden. Die ist für den in (1) gegebenen Lernersatz in Tabelle 4 dargestellt. Im zweiten Schritt habe ich mithilfe eines kleinen Tagsets, das eine Mischung aus formalen und ebenenspezifischen Tags enthält, für jede Korrektur die Fehler ermittelt.

L		Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen		.
A1	Das	Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen		.
A2	Das	Kunstmärchen	sollte	weiterhin	nicht losgelöst	vom	Volksmärchen	betrachtet werden	.
A3	Das	Kunstmärchen	ist	außerdem	abhängig	vom	Volksmärchen		.
A4	Das	Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen		.
A5	Das	Kunstmärchen	ist	weitestgehend		vom	Volksmärchen	abhängig	.

Tabelle 4: Satz (1) (Zeile 1) und die alignierten Zielhypothesen

Ich möchte das Tagset zunächst kurz erläutern. Es wurde so gewählt, dass es relativ mechanisch angewendet werden kann, wenn der Lernersatz mit der alignierten Annotation verglichen wird. Ich habe formal Ersetzungen (E), Einfügungen (I, für insert) und Löschungen (L) unterschieden[18] und dann jeweils noch hinzugefügt, ob das ersetzte, eingefügte oder gelöschte Element eine Nominalphrase (PHR), ein Inhaltswort oder eine Gruppe von Inhaltswörtern (WORT-I), ein Funktionswort oder eine Gruppe von Funktionswörtern (WORT-F) oder eine Clause (CL) ist. Eine Untergruppe der Funktionswörter sind die Artikel, die ich noch einmal gesondert gezählt habe, um das in Abschnitt 1 eingeführte Beispiel zu illustrieren. Wortstellungsfehler sind auf diese Art nicht direkt erkennbar, sondern nur als eine Sequenz von Löschung und Einfügung.[19] Für Orthographiefehler gibt es ein gesondertes Tagset. Die Fehlertags für Beispiel (1) und die Zielhypothesen A1 und A2 werden in den Tabellen 5 und 6 dargestellt.

L		Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen	.
A1	Das	Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen	.
	WORT-F-I							

Tabelle 5: Lerneräußerung (1) mit der Zielhypothese A1 und Fehlertagging

[18] Die Abweichungen sind von der Zielhypothese her zu lesen. Wenn also in der Zielhypothese ein Wort mehr steht, vergebe ich einen Einfügungstag. Der Lerner hat dabei ein Wort weggelassen.

[19] Es ist ohne ein zugrunde liegendes theoretisches Modell, wie zum Beispiel ein Feldermodell, schwierig, Wortstellungsfehler an einzelnen Tokens zu taggen, da nicht festgestellt werden kann, welches Wort eigentlich falsch steht.

L		Kunstmärchen	ist	weiterhin	abhängig	vom	Volksmärchen		.
A2	Das	Kunstmärchen	sollte	weiterhin	nicht losgelöst	vom	Volksmärchen	betrachtet werden	.
	WORT-F-I		WORT-I-E		WORT-I-E			WORT-I-I	

Tabelle 6: Lerneräußerung (1) mit der Zielhypothese A2 und Fehlertagging

Hier zeigt sich bereits, wie stark die unterschiedlichen Zielhypothesen auch bei einem Tagset und den gleichen Vergaberichtlinien zu verschiedenen Fehlerzählungen führen: Zielhypothese A1 führt zu einem Einfügungstag für ein Funktionswort, Zielhypothese A2 zu demselben Tag, aber außerdem noch zu zwei Ersetzungstags für Inhaltswörter und einem Einfügungstag für ein Inhaltswort.
Es gibt keinen der 17 Sätze des Lernertexts, für den alle Annotatoren dieselbe Zielhypothese gewählt haben (eine Übereinstimmung von einigen der Annotatoren gibt es manchmal, siehe A1 und A4 in Tabelle 4). Selbst Sätze, die von einigen Annotatoren als vollständig korrekt eingeschätzt wurden, wurden von anderen verbessert, wie in Tabelle 7 dargestellt, wo A2 zwei lexikalische Ersetzungen vornimmt, während alle anderen Annotatoren nichts ändern.

L	Beide	Gruppen	schränken	mit	ihren	Bedeutungserklärungen	den	Gegenstand	ein	.
A1	Beide	Gruppen	schränken	mit	ihren	Bedeutungserklärungen	den	Gegenstand	ein	.
A2	Beide	Gruppen	grenzen	mit	ihren	Bedeutungserklärungen	die	Interpretation	ein	.
A3	Beide	Gruppen	schränken	mit	ihren	Bedeutungserklärungen	den	Gegenstand	ein	.
A4	Beide	Gruppen	schränken	mit	ihren	Bedeutungserklärungen	den	Gegenstand	ein	.
A5	Beide	Gruppen	schränken	mit	ihren	Bedeutungserklärungen	den	Gegenstand	ein	.

Tabelle 7: Lerneräußerung mit Zielhypothesen.

Unterschiedliche Annotatoren haben also verschiedene Ziele: manche beschränken sich darauf, offensichtliche Fehler zu korrigieren, während andere den Lernern helfen wollen, ihren Ausdruck zu verbessern (wahrscheinlich gerade, weil es sich hier offensichtlich um einen fortgeschrittenen Lerner handelt). Da es – wie oft besprochen – keine klare Trennung von strukturellen und nichtstrukturellen Fehlern gibt, können hier selbst ausformulierte Richtlinien nicht immer weiterhelfen.
Bestimmte Entscheidungen in der Korrektur ziehen andere Entscheidungen nach sich. In der Lerneräußerung in Tabelle 8 zum Beispiel ist das Verb falsch gewählt und *Kunst-Novelle*, *Kunst-Ode* und *Kunst-Komödie* können jedenfalls so nicht ohne definiten Artikel stehen. Je nach Wahl des Verbs und der Struktur kann man aber auch hier unterschiedlich vorgehen.

L	Klar	ist	es	,	dass		dem	Kunstmärchen	nicht		Kunst-Novelle	,
A1	Klar	ist		,	dass		dem	Kunstmärchen	nicht		Kunst-Novelle	,
A2	Klar	ist		,	dass		dem	Kunstmärchen	keine		Kunst-Novelle	,
A3	Klar	ist		,	dass		dem	Kunstmärchen	nicht	die	Kunst-Novelle	,
A4	Klar	ist		,	dass	neben	dem	Kunstmärchen	nicht		Kunst-Novelle	,
A5	Klar	ist		,	dass		dem	Kunstmärchen	nicht	die	Kunst-Novelle	,

	Kunst-Ode	,		Kunst-Komödie	nebeneinanderstehen	.
	Kunst-Ode	,		Kunst-Komödie	gegenüber stehen	.
	Kunst-Ode	oder		Kunst-Komödie	gegenüber steht	.
die	Kunst-Ode	oder	die	Kunst-Komödie	zur Seite stehen	.
	Kunst-Ode	,		Kunst-Komödie	bestehen	.
die	Kunst-Ode	,	die	Kunst-Komödie	gegenüber stehen	.

Tabelle 8: Lerneräußerung mit Zielhypothesen

In diesem einen Satz können also je nach Zielhypothese drei Artikelfehler oder kein Artikelfehler gezählt werden. Durch diesen und ähnliche Fälle kommen die unterschiedlichen Zahlen zustande, die ich in Abschnitt 1 erwähnt habe: bei einheitlichem Fehlertagging komme ich – nur durch die verschiedenen Zielhypothesen – bei Annotator 3 auf 10 Artikelfehler und Annotator 4 auf 3 Artikelfehler.

Auch auf allen anderen Fehlerebenen wirken sich die Zielhypothesen zum Teil enorm aus, wie zusammengefasst in Tabelle 9 dargestellt. Hier habe ich alle Inhaltswortfehler (also Einfügung, Auslassung, Ersetzung), alle Funktionswortfehler (inklusive Artikelfehler), alle Phrasenfehler, alle Clausefehler und die Orthographiefehler in zwei verschiedenen Kategorien dargestellt, ORTH betrifft alle Getrennt/Zusammenschreibungsfehler (in beide Richtungen), Groß/Kleinschreibungsfehler (in beide Richtungen) und alle anderen Orthographiefehler ohne Interpunktionsfehler und ORTH-Z alle Interpunktionsfehler. Die Orthographiefehler habe ich extra aufgeführt, da oft angenommen wird, dass sie unkontrovers zu taggen seien. Dass Orthographiefehler sind allerdings auch von der Zielhypothese abhängig, dies gilt vor allem für Interpunktionsfehler (siehe das Beispiel in Tabelle 8, wo einige Annotatoren die Kommata durch *oder* ersetzt haben).

	WORT-I	WORT-F	PHR	CL	ORTH	ORTH-Z
A1	15	13	17	0	16	8
A2	24	26	14	5	23	15
A3	17	25	19	0	22	12
A4	16	12	15	2	19	10
A5	14	22	15	0	24	12

Tabelle 9: Gesamtfehlerzählungen für die verschiedenen Annotatoren

Die Ergebnisse zeigen in allen Spalten große Unterschiede. In fast jeder Spalte liegt der Maximalfehlerwert deutlich höher als (zum Teil mehr als doppelt so hoch wie) der Minimalfehlerwert. Man kann sich vorstellen, wie

5 Diskussion und Zusammenfassung

Am Ende von Abschnitt 3 habe ich vier Evaluationsaufgaben aufgezählt:

(a) Übereinstimmung in der Kategorisierung (Festlegung von Tagset und Vergaberichtlinien)
(b) Urteilerübereinstimmung in der Annotation nach Festlegung eines Tagsets
(c) Urteilerübereinstimmung in der Zielhypothese
(d) Urteilerübereinstimmung in der Kategorisierung in Abhängigkeit von der Zielhypothese.

Aufgabe (a) ist in der Literatur viel diskutiert worden – letztendlich ist hier immer die Aufgabenstellung bestimmend. Abschnitt 4.1 hat gezeigt, dass zumindest bei einem informellen Vorgehen kaum Übereinstimmung erzielt wird. Die Aufgaben (b) und (d) sind im Rahmen einer κ-Evaluation zu behandeln. In diesem Artikel habe ich mich im Wesentlichen der Aufgabe (c) gewidmet, die bisher in der korpuslinguistischen Literatur zu Lernerkorpora kaum behandelt wird. In Abschnitt 4.2 wurde gezeigt, wie sehr sich Zielhypothesen unterscheiden können (es gibt, wie gesagt, keinen Satz in dem betrachteten Lernertext, für den alle Zielhypothesen übereinstimmen). Jede Fehlerzählung ist von der zugrundeliegenden Zielhypothese abhängig. Tabelle 9 illustriert, wie stark Zählungen – die hier nach einem einheitlichen Tagset und Vergaberichtlinien vorgenommen wurden – von der Zielhypothese abhängig sind. Man kann aus den Daten völlig unterschiedliche Schlüsse ziehen. Wie kann man mit diesen Ergebnissen umgehen? Soll man – wie oben angedeutet – vielleicht aufhören, Korpora überhaupt mit Fehlern zu annotieren? Soll man aufhören, quantitative Studien zu Lernertexten durchzuführen? Sind Lernerkorpora dann überhaupt noch hilfreich?

Hier möchte ich dafür argumentieren, dass dieses Problem allgemein ist und es nur durch eine explizite Angabe von Zielhypothesen überhaupt möglich ist, nachprüfbare Ergebnisse bei der Auswertung von Lernerdaten zu erhalten.

Zunächst einmal handelt es sich natürlich nicht um ein Problem, das nur in elektronisch vorliegenden Korpora auftritt. Jede manuelle Auswertung von Lernerdaten ist genauso subjektiv. Nicht elektronisch vorliegendende Daten sind oft nicht allgemein verfügbar, so dass die Kategorisierungen und Ergebnisse nicht reproduzierbar sind. In vielen elektronisch vorliegenden Lernerkorpora können allerdings die Kategorisierungen auch nicht überprüft werden, da die Zielhypothesen nicht explizit gemacht werden. Das heißt, dass man sich bei Fehlerzählungen auf die Zielhypothese des Annotators verlassen muss und diese nicht einmal kennt. Außerdem können in den meisten Lernerkorpora keine Alternativen aufgeführt werden.

Ich möchte dafür argumentieren, dass die Architektur von Lernerkorpora so angelegt sein muss, dass es möglich ist, eine oder mehrere Zielhypothesen explizit anzugeben. Das kann dadurch geschehen, dass das Lernerkorpus in einer Mehrebenen-Architektur gespeichert wird, in der verschiedene Annotationsebenen unabhängig voneinander sind und auf eine Referenzebene (den Lernertext) verweisen. Solche Mehrebenen-Architekturen sind in den letzten Jahren vor allem für multimodale Korpora entwickelt worden und werden mehr und

mehr eingesetzt (siehe zum Beispiel Dipper et al. 2004 für eine Anwendung auf historische Texte oder Wörner et al. 2006 für ein allgemeines Framework; Lüdeling et al. 2005 und Siemen et al. 2006 stellen die technische Realisierung für Falko dar).
Man kann sich das im Prinzip so vorstellen, dass jede Zielhypothese getrennt bearbeitet und fehlergetaggt wird. In Tabelle 10 stelle ich das schematisch für einen der in der Untersuchung betrachteten Sätze vor. Nur wenn die Zielhypothesen angegeben sind, ist eine Fehleranalyse auf Lernerkorpusdaten transparent und reproduzierbar.

Erscheint in Patrick Grommes & Maik Walter (vorauss. 2007) *Fortgeschrittene Lernervarietäten*. Niemeyer, Tübingen

Vorversion – kann sich noch ändern

L	Klar	ist	es	,	dass		dem	Kunstmärchen	nicht		Kunst-Novelle	,		Kunst-Ode	,		Kunst-Komödie	nebeneinanderstehen	.
ZH1	Klar	ist		,	dass		dem	Kunstmärchen	nicht		Kunst-Novelle	,		Kunst-Ode	,		Kunst-Komödie	gegenüber stehen	.
FE1.1																			
…																			
FE1.n			X																
ZH2	Klar	ist		,	dass		dem	Kunstmärchen	keine		Kunst-Novelle	,		Kunst-Ode	oder		Kunst-Komödie	gegenüber steht	.
FE2.1																			
…																			
FE2.n			X																
ZH3	Klar	ist		,	dass		dem	Kunstmärchen	nicht	die	Kunst-Novelle	,	die	Kunst-Ode	oder	die	Kunst-Komödie	zur Seite stehen	.
FE3.1										X			X			X			
…																			
FE3.n			X																
ZH4	Klar	ist		,	dass	neben	dem	Kunstmärchen	nicht		Kunst-Novelle	,		Kunst-Ode	,		Kunst-Komödie	bestehen	.
FE4.1																			
…																			
FE4.n			X																
ZH5	Klar	ist		,	dass		dem	Kunstmärchen	nicht	die	Kunst-Novelle	,	die	Kunst-Ode	,	die	Kunst-Komödie	gegenüber stehen	.
FE5.1										X			X			X			
…																			
FE5.n			X																

Tabelle 10: Schematische Darstellung einer Mehrebenenarchitektur mit einer Lerneräußerung (markiert durch L), sowie verschiedenen Zielhypothesen (ZH1-ZH5) mit jeweils zugeordnetetn Fehlerebenen (FEm.1 .. FEm.n). Hier steht exemplarisch FEm.1 für Artikelfehler und FEm.n für Phrasenfehler; der Fehlerort wird jeweils durch ‚X' markiert. Weitere Fehlerebenen (zum Beispiel für Ausdrucksfehler) können jeweils hinzugefügt werden.

6. Literatur

Alle URLs in diesem Text wurden im November 2006 überprüft.

Abe, Mariko (2003). A corpus-based contrastive analysis of spoken and written learner corpora: The case of Japanese-speaking learners of English. In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 1-9. Technical Papers, Lancaster University.

Arrieta, Bertol, Arantza Díaz de Ilarraza, Koldo Gojenola, Montse Maritxalar & Maite Oronoz (2003). A database system for storing second language learner corpora. In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 33-41. Technical Papers, Lancaster University.

Belz, Judy A. (2004). Learner Corpus Analysis and the Development of Foreign Language Proficiency. *System: An International Journal of Educational Technology and Applied Linguistic*. 32.4: 577-591.

Biber, Douglas, Susan Conrad & Randi Reppen (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bongartz, Christine (in diesem Band). XXX

Bortz, Jürgen, Gustav A. Lienert, & Klaus Boehnke (2000). *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer-Verlag

Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics* 22(2):249-254

Carletta, Jean, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, & Holger Voormann (2003). The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers* 35(3): 353-363

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

Corder, Stephen Pit (1973). *Introducing Applied Linguistics.* Harmondsworth: Penguin.

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

Dagneaux, Estelle, Sharon Denness & Sylviane Granger (1998) Computer-aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics 26(2)*, 163-174.

Diessel, Holger (erscheint). Corpus linguistics and first language acquisition. In *Corpus Linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds). Berlin: Mouton de Gruyter

Dipper, Stefanie, Lukas Faulstich, Ulf Leser & Anke Lüdeling (2004) Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In *Workshop on XML-based richly annotated corpora, Lisbon*.

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

Garnier, Sandrine; Youhanizou Tall, Sisay Fissaha & Johann Haller (2003). Learner corpora: Design, development and applications - development of NLP tools for

CALL based on learner corpora (German as a foreign language). In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 246-252. Technical Papers, Lancaster University.

Granger, Sylviane (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in Contrast. Papers from the Symposium on Text-based Cross-linguistic Studies. Lund 4-5 March 1994*, Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.), 37-5. Lund: Lund University Press.

Granger, Sylviane (2002). A bird's-eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), 3-33. Amsterdam: John Benjamins.

Granger, Sylviane (erscheint). Learner corpora. In *Corpus Linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds). Berlin: Mouton de Gruyter

Gut, Ulrike (in diesem Band). XXX

Hirschmann, Hagen (2005). *Platzhalterphrasen bei fortgeschrittenen Lernern des Deutschen als Fremdsprache*. Unveröffentlichte Staatsexamensarbeit, Humboldt-Universität zu Berlin

Izumi, Emi, Toyomi Saiga, Thepchai Supnithi, Kiyotaka Uchimoto & Hitoshi Isahara (2003). The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques. In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 359-366. Technical Papers, Lancaster University

Izumi, Emi, Kiyotaka Uchimoto & Hitoshi Isahara (2005). Error Annotation for a corpus of Japanese Learner English. In: *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora* (LINC 2005)

Jurafsky, Daniel S. & James H. Martin (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Johansson, S., E. Atwell, R. Garside & G. Leech (1986). The Tagged LOB Corpus: Users' Manual. Bergen: Norwegian Computing Centre for the Humanities.

Kilgarriff, Adam (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language* 12(3): 453-472

Lemnitzer, Lothar & Heike Zinsmeister (2006). *Korpuslinguistik – Eine Einführung.* Tübingen: Gunther Narr Verlag.

Lennon, Paul (1991). Error and the very advanced learner. International Review of *Applied Linguistics* 29, 31-44

Lippert, Eva (2005) *Probleme von Nichtmuttersprachlern mit der Definitheit von Nominalphrasen.* Unveröffentlichte Magisterarbeit, Humboldt-Universität zu Berlin

Lippert, Eva (in diesem Band). Der oder ein – das ist die (eine?) Frage.

Lüdeling, Anke, Maik Walter, Emil Kroymann & Peter Adolphs (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005, Birmingham*. Online verfügbar unter http://www2.hu-berlin.de/korpling/projekte/falko/FALKO-CL2005.pdf

Lüdeling, Anke (2006). Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In *Jahrbuch des Instituts für deutsche Sprache 2006*, Gisela Zifonun & Werner Kallmeyer (Hgg.). Berlin: de Gruyter

MacWhinney, Brian (1996). The CHILDES system. *American Journal of Speech-Language Pathology* 5: 5-14.

MacWhinney, Brian & Catherine Snow (1985). The child language data exchange system. *Journal of Child Language* 12: 271-296.

Maden-Weinberger, Ursula (in diesem Band). Modality as Indicator of L2 Proficiency? A corpus based investigation into advanced German interlanguage

Manning, Christopher & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Meunier, Fanny (1998). Computer tools for the analysis of learner corpora. In *Learner English on Computer*, Sylviane Granger (ed.). London/New York: Addison Wesley Longman.

Milde, Jan-Torsten & Ulrike Gut (2002). A prosodic corpus of non-native speech. Proceedings of the Speech Prosody 2002 conference Aix-en-Provence, Bernard Bel & Isabel Marlien (eds.), 503-506.

Nesselhauf, Nadja (2004). Learner corpora and their potential in language teaching. In *How to Use Corpora in Language Teaching*, John Sinclair (ed.), 125-152. Amsterdam: Benjamins.

Pankow, Christiane & Pettersson, Helena (2006). Auswertung der Leistung von zwei frei zugänglichen POS-Taggern für die Annotation von Korpora des gesprochenen Deutsch. Göteborger Arbeitspapiere zur Sprachwissenschaft. Online verfügbar unter http://hum.gu.se/institutioner/tyska-och-nederlandska/tyska/publikationer/gas/gas_2006/

Pravec, Norma A. (2002). Survey of learner corpora. *ICAME Journal* 26: 81-114. Online verfügbar unter http://nora.hd.uib.no/icame/ij26/

Reuer, Veit & Kai-Uwe Kühnberger (2005). Feature Constraint Logic and Error Detection in ICALL systems. In *Proceedings of the Fifth International Conference on Logical Aspects of Computational Linguistics*. Berlin: Springer

Römer, Ute (2006). Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for the Future. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 121-134

van Rooy, Bertus & Schäfer, Lande (2003). Automatic POS tagging of a learner corpus: the influence of learner error on tagger accuracy. In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 835-844. Technical Papers, Lancaster University

Schmid, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Online verfügbar unter http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Siemen, Peter, Anke Lüdeling & Frank Henrik Müller (2006). FALKO - Fehler-Annotiertes LernerKOrpus des Deutschen. In: *Proceedings of Konvens 2006*, Konstanz. Online verfügbar unter http://www.ub.uni-konstanz.de/kops/volltexte/2006/2013/

Stern, Carola & William Stern (1907). *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung*. Leipzig: Barth.

Tono, Yukio (2003). Learner corpora: design, development, and applications. In *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* Dawn Archer, Paul Rayson, Andrew Wilson, & Tony McEnery (eds), 800-809. Technical Papers, Lancaster University

Veronis, Jean (2001). *Sense tagging: does it make sense?* Proceedings of the Corpus Linguistics 2001 Conference, Lancaster. Online verfügbar unter http://www.up.univ-mrs.fr/veronis/pdf/2001-lancaster-sense.pdf

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

Wörner, Kai, Andreas Witt, George Rehm & Stefanie Dipper (2006). Modelling Linguistic Data Structures. In *Proceedings of 'Extreme Markup Languages'* 2006, Montreal