

Erschienen in: Kallmeyer, Werner/Zifonun, Gisela (Hrsg.): Sprachkorpora. Datenmengen und Erkenntnisfortschritt. – Berlin, New York: de Gruyter, 2007. S. 28-48. (Institut für Deutsche Sprache. Jahrbuch 2006), <https://doi.org/10.1515/9783110439083-004>

ANKE LÜDELING

## Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik

### Abstract

Es gibt viele linguistische Forschungsfragen, für deren Beantwortung man Korpusdaten qualitativ und quantitativ auswerten möchte. Beide Auswertungsmethoden können sich auf den Korpustext, aber auch auf Annotationssebenen beziehen. Jede Art von Annotation, also Kategorisierung, stellt einen kontrollierten und notwendigen Informationsverlust dar. Das bedeutet, dass jede Art von Kategorisierung auch eine Interpretation der Daten ist. In den meisten großen Korpora wird zu jeder vorgesehenen AnnotationsEbene, wie z. B. Wortart-Ebene oder Lemma-Ebene, genau eine Interpretation angeboten. In den letzten Jahren haben sich neben den großen, ‚flach‘ annotierten Korpora Korpusmodelle herausgebildet, mit denen man konfligierende Informationen kodieren kann, die so genannten Mehrebenen-Modelle (*multilevel standoff corpora*), in denen alle Annotationssebenen unabhängig vom Text gespeichert werden und nur auf bestimmte Textanker verweisen. Ich argumentiere anhand der Fehlerannotation in einem Lernerkorpus dafür, dass zumindest Korpora, in denen es stark variierende Annotationsbedürfnisse und umstrittene Analysen geben kann, davon profitieren, in Mehrebenen-Modellen kodiert zu werden.

### 1. Einleitung

Korpora werden zur Beantwortung von linguistischen Fragestellungen immer wichtiger. In einigen Bereichen wie der historischen Linguistik oder der Dialektologie wird natürlich traditionell schon immer mit Korpora gearbeitet (es gibt ja oft keine andere Datenquelle), aber in den letzten Jahren finden Korpora als Datenquelle auch in eher theoretisch ausgerichteten Arbeiten immer häufiger Anwendung (siehe zum Beispiel Gries 2003; mehrere Artikel in Kepser & Reis 2005; Featherston, in diesem Band; Müller, in diesem Band). Neben der Verwendung von Korpora als eine Art ‚Beispielbank‘ und rein qualitativen Analysen wird auch immer häufiger quantitativ gearbeitet. Neben relativ einfachen beschreibenden statistischen Techniken werden auch statistische Tests, Kollokationsanalysen, multivariate Techniken und Verfahren aus dem maschinellen Lernen eingesetzt. Die mathematischen Verfahren und deren Voraussetzungen sind in vielen Arbeiten beschrieben worden (Biber 1988; Oakes 1998; Manning & Schütze 1999; Baayen 2001; Evert 2005; Gries 2003 und viele andere).

Im Gegensatz zu anderen empirischen Gebieten wie der Psycholinguistik befindet sich die Entwicklung von Standards für die quantitative Analyse von Korpusdaten noch in der Diskussion. Neben der Frage, *wie* gezählt und ge-

rechnet wird, ist hier die Frage entscheidend, *was* eigentlich gezählt wird. In diesem Papier möchte ich nicht auf die mathematischen Techniken selbst eingehen, sondern auf die Grundlage jeder quantitativen Analyse, nämlich die qualitative Analyse oder Kategorisierung der Daten.

Jede quantitative Analyse ist abhängig von einer vorherigen Kategorisierung. Allerdings ist eine solche Kategorisierung oft nicht unproblematisch und unumstritten – diese Tatsache ist bekannt und eigentlich trivial. Trotzdem findet man in sehr vielen Arbeiten, die Korpora als Datenquelle für quantitative Untersuchungen verwenden, kaum Hinweise auf die zugrunde liegende Kategorisierung – oft werden weder die verwendeten Kategorien (Tagsets) noch die Richtlinien zur Vergabe der Kategorien angegeben. Noch seltener findet man Angaben zum so genannten „Inter Annotator Agreement“, der Zuverlässigkeit bei der Vergabe von Kategorien. Ohne diese Angaben sind quantitative Ergebnisse eigentlich nicht zu bewerten.

Ich möchte in diesem Papier zunächst verdeutlichen, dass es unmöglich ist, nicht zu kategorisieren (Abschnitt 2), auch wenn keine explizite Kategorisierung vorliegt. Dann zeige ich exemplarisch anhand eines Lernerkorpus, wie stark Kategorisierungen die Zählungen beeinflussen können (Abschnitt 4) und welche Schwierigkeiten sich durch fehlende Evaluierungen ergeben. Das verwendete Lernerkorpus Falko wird kurz in Abschnitt 3 vorgestellt. Im letzten Teil (Abschnitt 5) möchte ich für eine Mehrebenenkorpusarchitektur argumentieren, in der die unterschiedlichen Hypothesen zumindest explizit gemacht werden können, so dass die Grundlage für jede quantitative Auswertung sichtbar ist.

## 2. Kategorisierung und quantitative Analyse

Ein Grunddatum der empirischen Linguistik ist die sprachliche Äußerung. Korpora sind immer interpretierte Repräsentationen solcher Äußerungen und unterscheiden sich als solches bereits von den ursprünglichen Äußerungen. So muss man unterscheiden zwischen einer sprachlichen Äußerung, die in einer Zeitung steht und einem Korpus, in dem diese Äußerung repräsentiert ist (siehe auch Moisl, *erscheint*). Zum einen unterscheiden sich die Reaktions- und Analysemöglichkeiten (so kann man z. B. keinen Leserbrief an das Korpus schreiben, aber evtl. statistische Auswertungen vornehmen), zum anderen unterscheiden sich die Kontextinformationen. Noch deutlich wird dieser Schritt, wenn man an elektronische Wiedergaben von historischen Manuskripten oder Transkriptionen gesprochener Sprache denkt. In jedem Fall werden verschiedene Formen im Original notwendigerweise auf gleiche Formen im Korpus zurückgeführt – selbst, wenn man sehr diplomatisch arbeitet.<sup>1</sup>

---

<sup>1</sup> Bloomfield (1933) unterscheidet in diesem Zusammenhang zwischen ‚primary data‘ und ‚secondary data‘, Himmelmann (1998) zwischen ‚raw data‘, ‚primary data‘ und ‚secondary data‘. Siehe auch Lehmann (in diesem Band).

Die Auswertung von Korpusdaten erfolgt dann oft auf einer weiteren Interpretationsstufe, der Annotation.

In beiden Interpretationsstufen werden Entscheidungen getroffen. Wir haben es also nie mit uninterpretierten Daten zu tun. Die Interpretation von Korpusdaten stellt in jedem Fall eine Kategorisierung dar, also kontrollierten und erwünschten Informationsverlust. Dies gilt für die qualitative genauso wie für die quantitative Auswertung, für die Arbeit ohne explizite Annotationen genauso wie für die Arbeit mit expliziten Annotationskategorien, unabhängig davon, wie die Korpusdaten verarbeitet werden.

Für viele Fragestellungen genügen einige wenige Beispiele – man braucht keine quantitativen Aussagen. Wenn man zum Beispiel nur illustrieren möchte, dass Lerner des Deutschen als Fremdsprache Schwierigkeiten bei der Verwendung des Korrelat-*es* haben, indem sie ein *es* einfügen, wo keines hingehört (1a) oder ein obligatorisches *es* nicht setzen ((1b); für eine detaillierte Analyse siehe Hirschmann 2005), genügt es, einige solche Fälle zu finden – man verwendet das Korpus als eine Art ‚Beispielbank‘.<sup>2</sup> Wichtig ist aber auch in diesen Fällen, dass die Beispiele nach externen Kriterien ausgesucht und gruppiert werden.<sup>3</sup> Man hat vorher festgelegt, welche Konstruktionen betrachtet werden und auch, wie man solche Konstruktionen finden kann. Die Kriterien müssen so explizit sein, dass für Fälle wie (1c), wo das *es* entweder als stilistisches Problem oder als grammatischer Fehler gewertet werden kann, eine eindeutige und nachvollziehbare Entscheidung getroffen werden kann.

(1a) *Z. B. wenn es das konventionelle Wort „Lehrer“ existiert, sollte man nicht das Wort „Unterrichter“ verwendet.* (Falko L2, Text 36)

(1b) *Eine Vielzahl syntaktischer Mittel, z. B. Relativsatz, Präpositionalphrase, Genitivergänzung, adjektivische Modifikationen, ermöglichen ein bestimmtes Mitglied dieser Klasse hervorzuheben.* (Falko L2, Text 17)

(1c) *Das zweite Paradox ist es, dass der Mann sich weigert hereinzutreten, obwohl das Tor offensteht, und den Befehl des Türhüters befolgt.* (Falko L2, Text 5)

Bei der quantitativen Auswertung muss genauso nach externen Kriterien festgelegt werden, was gezählt wird. In der Korrelatstudie von Hirschmann beispielsweise sind nur sehr wenige echte Korrelatfehler zu finden, es gibt aber einige Zweifelsfälle (analog zu (1c)). Wenn die Zweifelsfälle nicht als Fehler gewertet werden, kann man schließen, dass Lerner in dieser Konstruktion kaum mehr Fehler machen als Muttersprachler. Wenn sie hingegen als Fehler gewertet werden, könnte man schließen, dass Lerner mit Korrelatkonstruk-

<sup>2</sup> Die Beispiele stammen aus dem Lernerkorpus Falko, das in Abschnitt 3 eingeführt wird. Die Beispiele enthalten außer den Korrelatfehlern noch andere Fehler, die aber hier nicht betrachtet werden.

<sup>3</sup> Dies gilt letztendlich auch für die sogenannten korpusgesteuerten (corpus driven) Ansätze (Xiao, erscheint) oder auch für Arbeiten wie die von Golcher (2006), in denen ohne explizites Lexikon oder andere sprachspezifische Interpretationen Muster in einer Sprache gefunden werden.

tionen Probleme haben. (auf die Schwierigkeiten der Fehlerkategorisierung komme ich in Abschnitt 4 zurück).

Dass die Kategorisierung bei kleinen Korpora die Ergebnisse wesentlich beeinflussen kann, ist einleuchtend. Kleine spezielle Korpora werden meistens von Hand vorverarbeitet. Wenn die Verarbeitungskriterien und -verfahren gut dokumentiert sind, sind die Entscheidungen nachvollziehbar.

Aber auch bei großen Korpora (hier meine ich zum Beispiel die über das IDS oder über die Berlin-Brandenburgische Akademie verfügbaren Korpora, aber auch Korpora, die in der Computerlinguistik eine Rolle spielen) spielt die Kategorisierung eine entscheidende Rolle. Große Korpora werden in der Regel automatisch vorverarbeitet, d. h. tokenisiert, getaggt, lemmatisiert oder auf anderen Ebenen annotiert. In den meisten Fällen geschieht die Verarbeitung von Korpora sequentiell. Zunächst wird tokenisiert, dann werden die so gefundenen Tokens lemmatisiert und mit Wortarten getaggt, danach werden größere Einheiten (wie Chunks oder Sätze) ausgezeichnet (Manning & Schütze 1999; Carstensen et al. 2004). Bereits das automatische Tokenisieren kann problematisch sein. In den meisten europäischen Sprachen werden graphische Wörter als Tokens angesehen.<sup>4</sup> Dies ist in vielen Fällen problematisch, wenn man ‚Token‘ als ‚Wort‘ ansehen möchte (Evert & Fitschen 2001, Haß-Zumkehr 2002): Einerseits hat man den Fall dass mehrere Tokens als eine Einheit verarbeitet werden sollten (es gibt keinen linguistischen Grund, Namen mit Spatium wie *New York* oder *Weil der Stadt* anders zu behandeln als *Hamburg*; es gibt eine Reihe von linguistischen Gründen, nicht kompositionelle Sequenzen mit Spatium wie engl.: *pigeon hole* ‚Postfach‘ als Einheiten zu sehen). Andererseits gibt es Tokens, die als mehrere lexikalische Einheiten angesehen werden sollten (dazu gehören zum Beispiel Verschmelzungen wie *siehste*). In einer sequentiellen Architektur bauen alle weiteren Verarbeitungsschritte auf dieser Tokenisierung auf, d. h., alle Probleme werden ‚mitgeschleppt‘. Die Wortartzuweisung und die Lemmatisierung müssen irgendwie mit merkwürdigen Tokens umgehen. Die Strategien sind hier unterschiedlich. Ganz generell ist es aber so, dass bekannte Probleme (wie *New York* oder *zum*) einheitlich behandelt werden können, während seltenere Fälle nicht abgefangen werden können, wie die folgenden Beispiele<sup>5</sup> zeigen.

<sup>4</sup> Dies ist anders für viele nicht-europäische Sprachen, aber auch für historische Texte mit scripta continua (siehe Lüdeling, Poschenrieder & Faulstich 2005 oder Ostler, erscheint).

<sup>5</sup> Alle hier verwendeten Korpora wurden mit dem DecisionTreeTagger (Schmid 1994, <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>) annotiert. Hier wird ein Tabellenformat verwendet (siehe Abschnitt 5). Jedes Token ist mit einem Wortarttag und einem Lemma versehen. Die Wortarttags entstammen dem Stuttgart-Tübingen Tagset, <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>. Die hier vorkommenden Tags sind ADJA für attributives Adjektiv, KOUS für unterordnende Konjunktion, ART für Artikel, NN für Nomen, NE für Name. Die Korpora sind unter <http://www2.hu-berlin.de/korpling/korpora/index.php> verfügbar (kostenlose Registrierung erforderlich).

(2a) *Ach!ITJ!ach,/\$,/,siehstelADJA!<unknown>!/\$.!!*

(aus dem Parlamentsredenkorpus)

(2b) *Rolf Sannwald, Vorstandsmitglied der 'Calwer Decken- und Tuchfabriken AG, Calw, und AR-Vorsitzender der Wolldeckenfabrik WeillKOUS!weil der!ART!d Stadt!NN!Stadt AG [...] ist mit dem Bundesverdienstkreuz 1. Klasse ausgezeichnet worden.*

(aus dem Bonner Zeitungskorpus; zur besseren Lesbarkeit wurden Wortart-Tags und Lemmanamen für alle Tokens außer den hier relevanten entfernt).

Die besten statistischen Wortarttagger für das Deutsche gehen bei Zeitungstexten von einer Korrektheit von zwischen 96 und 98 % aus (Carstensen et al. 2004). Allerdings sind die Fehler nicht gleichmäßig über alle Wortarten verteilt, sondern oft fast ‚strukturell‘. Lüdeling (2000) diskutiert zum Beispiel die Konsequenzen, die sich daraus ergeben, dass getrennt stehende Verbpartikeln von statistischen Taggern wegen ihrer Homographie zu anderen Kategorien unzuverlässig getaggt werden. Bosch, Rozario & Zhao (2003) zeigen, dass das Demonstrativpronomen *der* bei einem statistischen Tagger fast durchgängig falsch getaggt wird, weil es deutlich seltener vorkommt als das Relativpronomen *der* und daher in den Trainingsdaten kaum auftritt. Quantitative Analysen des Demonstrativpronomens sind also auf dieser Basis nicht möglich. Bei Ambiguitäten zwischen homographen Namen und Nomina kann es auch zu systematischen Effekten kommen, wie in (3) illustriert wird. Die Tagfolge CARD NN NN ist viel häufiger als die Tagfolge CARD NN NE, so dass ein Tagger bei dem mehrdeutigen *Kohl* immer auf diese Art entscheidet (man kann sich ja auch Situationen vorstellen, in denen man ziemlich lange das Gleiche essen muss).

(3) *Wir haben jetzt 16/CARD Jahre!NN Kohl!NN hinter uns!*

(aus dem Parlamentsredenkorpus)

Bei großen, weit verbreiteten Korpora kann man sich auf die jeweils vorliegende Vorverarbeitung stützen; wenn das Korpus nur über Internetabfragen zugänglich ist, hat man keine andere Wahl. Die Vorverarbeitungsprogramme sollten daher generell bekannt sein.

Hier unterscheiden sich kleine, spezialisierte Korpora, wie Dialektkorpora, historische Korpora, die meisten Korpora gesprochener Sprache und auch Lernerkorpora. Kann man mit spezialisierten Korpora genauso umgehen wie mit den ‚Standard‘korpora? Für solche Korpora gibt es bisher kaum Annotationsstandards. Je weniger ‚standardisiert‘ die Daten selbst sind, desto weniger kann man auf vorhandene Annotationsprogramme und -verfahren zurückgreifen, da diese ja meist statistische Komponenten enthalten und daher auf Regelmäßigkeiten angewiesen sind. Spezialisierte Korpora werden daher, wenn überhaupt, manuell annotiert. Da ergibt sich ein Problem: oft sind die Analysen selbst kontrovers. Daher stellen sich die folgenden Fragen: Wie geht man mit konfligierenden Analysen um? Wie kann die Korpusarchi-

tektur aussehen, so dass unterschiedliche Analysen repräsentiert werden können? Wie kann man bei konfligierenden Analysen zählen? Zählt jedes Vorkommen eines Phänomens gleich viel?

Ich möchte diese Fragen in Abschnitt 4 anhand eines Lernerkorpus diskutieren. Zunächst stelle ich das Korpus kurz vor.

### 3. Lernerkorpora

Lernerkorpora sind kontrollierte Sammlungen von Äußerungen von Lernern einer Fremdsprache.<sup>6</sup> Lernerkorpora können – zusammen mit gezielten Erhebungen wie zum Beispiel Elizitationstests – zur Erforschung von Erwerbsprozessen verwendet werden. Die Erkenntnisse können dann in der Vermittlung eingesetzt werden. Für das Englische als Fremdsprache werden Lernerkorpora schon relativ häufig verwendet (Granger 2002, erscheint; Mukherjee 2002; Pravec 2002; Tono 2003; Nesselhauf 2005; Römer 2006 und viele andere), für das Deutsche als Fremdsprache haben sie sich dagegen noch nicht so stark durchsetzen können. Obwohl es an vielen Orten mehr oder weniger private Sammlungen von Lerneräußerungen gibt (wie zum Beispiel das nicht frei zugängliche Korpus, das in Belz 2004 beschrieben wird), stehen bisher noch kaum (fehlerannotierte) Lernerkorpora frei und allgemein zur Verfügung (Ausnahmen sind die ESF-Korpora<sup>7</sup> mit gesprochenen Daten aus dem ungesteuerten Zweitspracherwerb und das tief prosodisch annotierte Leap-Korpus<sup>8</sup>).

Falko<sup>9</sup> versucht, hier eine Lücke zu schließen. Falko enthält Texte von fortgeschrittenen Lernern des Deutschen und besteht aus unterschiedlichen Subkorpora. In diesem Artikel beziehe ich mich nur auf Daten aus dem so genannten Zusammenfassungskorpus, das Texte enthält, die nicht-deutschsprachige Germanistikstudierende an der Freien Universität Berlin im Rahmen einer obligatorischen Sprachstandsüberprüfung im Zusammenhang mit der Zwischenprüfung verfasst haben (zur Zeit ca. 37.000 Tokens; das Korpus wächst ständig). Dies sind Zusammenfassungen eines linguistischen oder eines literaturwissenschaftlichen Fachtextes; die Aufgabenstellung war „Fassen Sie den Text mit eignen Worten zusammen“. Die Texte liegen zunächst

---

<sup>6</sup> Lernerkorpora unterscheiden sich von reinen Fehlersammlungen (wie in Heringer 1995) dadurch, dass sie ganze Texte enthalten, so dass zum Beispiel auch untersucht werden kann, welche Konstruktionen, Wörter etc. ein Lerner schon beherrscht und quantitative Untersuchungen möglich sind.

<sup>7</sup> [http://www.mpi.nl/world/data/esf\\_archive/html/esf.html](http://www.mpi.nl/world/data/esf_archive/html/esf.html) (Mai 2006).

<sup>8</sup> <http://www.phonetik.uni-freiburg.de/leap/corpus.html> (Mai 2006), siehe auch Milde & Gut (2002).

<sup>9</sup> Falko steht für fehlerannotiertes Lernerkorpus. Das Korpus wird in Kooperation der Humboldt-Universität und der Freien Universität in Berlin entwickelt und ist frei zugänglich. Informationen und Zugriff unter <http://www2.hu-berlin.de/korpling/projekte/falko/index.php>.

handschriftlich vor<sup>10</sup> und werden dann digitalisiert.<sup>11</sup> Des Weiteren gibt es ein vergleichbares Kontrollkorpus mit Muttersprachlerdaten (zurzeit ca. 13.000 Tokens).

#### 4. Lernerkorpora: Annotation und Auswertung

Ich möchte die in Abschnitt 2 diskutierten Fragen nun exemplarisch anhand von Falko-Daten diskutieren. Nach jedem Unterabschnitt werde ich kurz diskutieren, wie sich die hier besprochenen Probleme in anderen Korpusstypen auswirken.

Lernerkorpora werden auf zwei Arten ausgewertet, durch Fehleranalyse (Error Analysis, EA) und durch kontrastive Analyse (Contrastive Interlanguage Analysis, CIA). In der Fehleranalyse werden Fehler klassifiziert und annotiert. In der kontrastiven Analyse werden die Lernertexte statistisch mit anderen Texten (Muttersprachlertexte oder Texte anderer Lernergruppen) verglichen. So kann man zum Beispiel feststellen, ob Lerner ein bestimmtes Wort, eine Klasse von Wörtern oder eine Konstruktion im Vergleich zu häufig (Übergebrauch, *overuse*) oder zu selten (Mindergebrauch, *underuse*) anwenden, ob sich die durchschnittliche Satz- oder Wortlänge unterscheidet und wie sich die Fehleranzahl der Lerner im Vergleich zur Fehleranzahl der Muttersprachler verhält.

Sowohl für die Fehleranalyse als auch für die kontrastive Analyse müssen die Lernerdaten kategorisiert werden. Während sich für die Lemmatisierung und die Wortartzuweisung noch relativ leicht gemeinsame Kategorien finden lassen, muss man in der Arbeit mit vom Standard abweichenden Daten noch deutlich stärker interpretieren.

##### 4.1 Fehleranalyse

Zunächst muss beantwortet werden, was ein *Fehler* ist. Dies ist in der Literatur häufig und kontrovers diskutiert worden – ich kann hier nur kurz einige Grundzüge der Diskussion zusammenfassen (zur Diskussion siehe zum Beispiel Cherubim 1980; Corder 1981 oder Lennon 1991). Ein Fehler kann zum

<sup>10</sup> Dadurch wird verhindert, dass die Studierenden automatische Rechtschreibkorrekturen etc. nutzen können.

<sup>11</sup> Durch die Klausursituation ist die Erhebung stark kontrolliert: die Zeitdauer ist vorgegeben, die Studierenden haben keinen Zugriff auf Hilfsmittel. Die Studierenden werden als ‚fortgeschritten‘ eingestuft, da sie die DSH-Prüfung bestanden haben und ihr Grundstudium an der FU absolviert haben. Dies ist sicher zu diskutieren – man braucht jedoch ein externes Kriterium, um dann die sprachlichen Eigenschaften der jeweiligen Varietät zu beschreiben (siehe die Diskussion in Walter & Grommes 2006). Zusätzlich zu dem hier beschriebenen Zusammenfassungskorpus gibt es noch ein Korpus mit Longitudinaldaten, das an der Georgetown University in Washington erhoben wurde und einige weitere kleinere Korpora. Zur Zeit wird ein kontrolliertes Korpus mit Essays erhoben.

einen als ein Bruch einer Regel aufgefasst werden und zum anderen als eine Abweichung von einer Norm. Die Begriffe ‚Regel‘ und ‚Norm‘ werden unterschiedlich verwendet – man könnte sie zum Beispiel folgendermaßen sehen: Eine Regel ist von der Sprache vorgegeben (‚im Deutschen sind Subjekt und Verb numeruskongruent‘), der Bruch einer Regel führt zu ungrammatischen Äußerungen. Eine Norm hingegen ist von außen gesetzt (Orthographie oder Ausdruckskonventionen), die Abweichung einer Norm führt zu unakzeptablen Äußerungen. Corder (1973) spricht hier von ‚breaches of code‘.

Wie auch immer eine theoretische Einteilung von Fehlern oder Abweichungen aussieht, in konkreten Äußerungen ist es nicht immer möglich, sauber zu kategorisieren. Wenn man zum Beispiel Satz (4) betrachtet, kann man kaum sagen, wo hier die Fehler sind. *Zweispältig* ist sicher ein Lexikonfehler, allerdings ist das Wort nach den Regeln der Grammatik an sich korrekt gebildet. *Verhältnis in seinem Wort* ist sicher nicht korrekt, aber was genau ist hier der Fehler? Sollte man es ganz weglassen? Wie umformulieren? Gegen welche Regel oder Norm wird hier verstoßen?

(4) *Auf der anderen Seite hat Kunstmärchen ein zweispältiges Verhältnis in seinem Wort.*

(Falko L2, Text 46)

Dann muss man auch unterscheiden zwischen solchen Fehlern, die ein Lerner macht, weil er eine bestimmte Regeln/Norm noch nicht beherrscht (Kompetenzfehler, errors) und solchen, die er macht, weil er müde oder unkonzentriert ist, auch wenn er die relevanten Regeln schon kennt (Performanzfehler, mistakes, lapses). An der Oberfläche sind errors und mistakes allerdings nicht zu unterscheiden.

Man kann auch Übergebrauch und Mindergebrauch als ‚Fehler‘ auffassen, wenn man Lennons Fehlerdefinition zugrunde legt. Hier ist ein Fehler „a linguistic form, ... which, in the same context would in all likelihood not be produced by the learner’s native speaker counterparts.“ (Lennon 1991, S. 182). Um solche Fehler zu finden braucht man ein vergleichbares Referenzkorpus mit Muttersprachleräußerungen und muss operationalisieren, was ‚in the same context‘ und ‚in all likelihood‘ bedeuten würde. Selbst wenn man das könnte, ist es schwierig zu sagen, wo genau der Fehler liegt: bei Übergebrauch (zum Beispiel einer bestimmten Konjunktion) ist ja jede einzelne Verwendung der Konjunktion korrekt. Wie soll man entscheiden, welche nicht hätte gebraucht werden dürfen? Mindergebrauch ist noch schwieriger: An welcher Stelle in einem Text müssten Lerner denn ein (im Vergleich zu Muttersprachlern) zu selten gebrauchtes Wort einsetzen? Wo genau ist der Fehler?

Man kann also nicht systematisch Regel- von Normverstößen oder errors von mistakes unterscheiden. Außerdem ist eine Operationalisierung von Über- und Mindergebrauch kaum möglich.

Analoge Probleme ergeben sich in vielen anderen Korpusstypen. So ist es zum Beispiel nicht klar, welche Parameter getaggt werden sollen, wenn man Variation und kurzfristigen Wandel untersuchen möchte (Baumgarten et al. erscheint). Oder wie zum Beispiel Code-Switching zu behandeln wäre. Welche zugrundeliegenden Kategorien sollen annotiert werden, wenn man sich mit Informationsstruktur (Hinterhölzl, Petrova & Solf 2005) oder mit Diskursaufbau befasst?

## 4.2 Konfligierende Analysen: Zielhypothesen

Ich möchte noch auf ein weiteres Problem eingehen: um einen Fehler zu annotieren, muss man eine Hypothese darüber haben, wie die Äußerung hätte korrekt lauten müssen. In den meisten Lernerkorpora sind diese so genannten Zielhypothesen oder „reconstruction of those utterances in the target language“ (Ellis 1994, S. 54) implizit. Die folgenden Beispiele aus Weinberger (2002) sollen das verdeutlichen, wobei Weinbergers Arbeit stellvertretend für viele steht. In (5a) gibt es eine Kongruenzverletzung bei *diese Phänomen*. Weinberger nimmt für *diese* einen Genusfehler an. Ihre implizite Zielhypothese ist also *dieses Phänomen*. Wahrscheinlich ist diese Hypothese durch weiteren Kontext gestützt (der aber in der Arbeit nicht gegeben wird). Ohne Kontextinformation hätte man hier auch annehmen können, dass *diese Phänomene* gemeint wäre und ein Numerusfehler bei *Phänomene* auftritt. In (5b) ist das Problem schwieriger: Weinberger geht von einem phrasalen Ausdrucksfehler aus. In ihrer Arbeit gibt sie an, dass der Lerner hätte schreiben sollen *die Gesellschaft hat sich verändert*. Dies ist aber im Korpus nicht zu sehen – jemand, der nicht gleichzeitig die Arbeit vor sich hat und mit dem Korpus arbeitet, kann das nicht erschließen.<sup>12</sup>

- (5a) *die Erklärung für <MoArInGn> diese Phänomen ist*  
 (Fehlertags: Mo – morphology, Ar – article, In – inflection, Gn – gender)
- (5b) *<LxPhCh> Es gibt eine veränderte Gesellschaft und*  
 (Fehlertags: Lx – lexical, Ph – phrase, Ch – incorrect choice)

Hier sieht man, dass jede Fehlerannotation notwendigerweise eine Zielhypothese voraussetzt, egal ob diese implizit oder explizit (wie zum Beispiel bei Izumi, Uchimoto & Isahara 2005) angegeben ist. Zielhypothesen sind aber fast nie unumstritten, meistens sind mehrere Zielhypothesen für eine Lerneräußerung möglich. Je komplexer die betrachteten Sätze werden, desto unterschiedlicher sind die möglichen Zielhypothesen.

Unterschiedliche Zielhypothesen führen zu verschiedenen Fehlerzählungen. Betrachten wir Beispiel (6a) aus dem Falko-Korpus und zwei der möglichen

<sup>12</sup> Außerdem gibt es hier Probleme durch das Tabellenformat. Das Ende des phrasalen Fehlers ist nicht explizit kodiert (die Hervorhebung findet sich nicht im Korpus). Siehe dazu Lüdeling, Poschenrieder & Faulstich (2005).

Zielhypothesen (Fehler, die in beiden Zielhypothesen korrigiert werden wie *Jahrhunder* werden hier nicht betrachtet).

Für Zielhypothese 1 in (6b) wird das Komma ernst genommen und ein Relativsatz konstruiert. Das führt zu deutlichen Wortstellungsveränderungen und in einer Fehlerzählung zu mehreren Wortstellungsfehlern. Außerdem sind zwei Auslassungsfehler (*die* und *gegründet wurde*) festzustellen. Zielhypothese 2 in (6c) greift weit weniger in die Struktur ein und fügt ein adjektivisches Partizip ein. Hier sehen wir also einen Auslassungsfehler und keine Wortstellungsfehler, aber einen zusätzlichen Orthografiefehler (ein überflüssiges Komma). Allein an diesem Beispiel sieht man, wie stark sich Zählungen je nach Zielhypothese unterscheiden können (eine ausführlichere Auswertung des so genannten Inter-Annotator Agreement von Zielhypothesen und den daraus folgenden Fehlerzählungen findet sich in Lüdeling (2007)).

- (6a) Lerneräußerung: *Der Realismus ist eine im 19. Jahrhundert, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte.*  
(Falko L2, Text 25)
- (6b) Zielhypothese 1: *Der Realismus ist eine literarische Richtung, die im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründet wurde und sich ...*
- (6c) Zielhypothese 2: *Der Realismus ist eine im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründete literarische Richtung, die sich ...*

Wenn die Zielhypothese so entscheidend für die Fehlerzählung und alle sich daraus ergebenden Konsequenzen für die Beschreibung von Lernabläufen ist, sollte sie explizit angegeben und zugänglich sein. Idealerweise sollte es möglich sein, konkurrierende Zielhypothesen für denselben Text aufzustellen. Darauf komme ich in Abschnitt 5 zurück.

Auch dies ist nicht nur für Lernerkorpora problematisch, sondern auf andere Korpusarten übertragbar. Es gibt in jedem Korpus klar ungrammatische Äußerungen, die nicht einmal unbedingt selten sein müssen, wie man am Beispiel der berühmten Rede von Trappatoni sieht, die oft zitiert wird, oder an der Tatsache, dass bestimmte Tippfehler, wie *immer* mit drei *m* häufig vorkommen. Wie mit solchen ‚Fehlern‘ umgegangen wird, hängt sicher mit der Forschungsfrage zusammen, aber es gibt sicher viele Linguisten, die keine Grammatik des Deutschen schreiben wollen, die *ich habe fertig* lizenziert. Das bedeutet, dass das Korpus nach externen Maßstäben gefiltert wird – implizit werden hier Hypothesen über korrekte Strukturen angenommen.

Auch historische Korpora oder Korpora gesprochener Sprache werden interpretiert. Abkürzungen werden expandiert, Hypothesen über kaum lesbare Buchstaben oder unverständliche Äußerungen gebildet. In solchen Fällen wird im Korpus nur eine ‚Ziel‘hypothese angegeben, oft ohne zu markieren, dass die Hypothese unsicher ist.

### 4.3 Tagsets und Evaluation

Ein weiterer Bereich, der für die Zählung wichtig ist, ist die Kategorisierung der Fehler selbst. Dafür müssen Tagsets entwickelt werden, analog zur Entwicklung von Tagsets für Wortarten oder andere Annotationsebenen. Dabei müssen mehrere Entscheidungen getroffen werden: (a) Fehlerexponent (das Wort oder die Phrase, an der der Fehlertag hängt), (b) Granularität und (c) Beschreibung des Fehlers. Alle drei Entscheidungen haben Auswirkungen auf die Zählung.

Zu (a): Ein Fehlertag kann (abhängig von der Korpusarchitektur, siehe auch Abschnitt 5) an ein Wort oder eine Phrase ‚andocken‘. Bei Wortstellungsfehlern kann es schwierig sein zu entscheiden, wo genau der Fehler ist. Wo sollen fehlende Elemente kodiert werden? In Beispiel (7) fehlt ein *es* nach dem *und*. In den meisten Architekturen müssen Tags an vorhandene Tokens gebunden sein – soll man das *es* am *und* oder am *ist* markieren? Oder vielleicht am *ist klar*, dem ja das Subjekt fehlt? Eine mögliche Zielhypothese wäre *Wenn ... und es klar ist, ...* Hier wäre also auch eine Umstellung von *ist* und *klar* zu markieren. Wieder stellt sich die Frage, wo genau die Umstellung zu markieren wäre. Je nachdem, wie man entscheidet, hat man einen Verbstellungsfehler oder nicht.

(7) *Wenn man die Valenz eines Verbs kennt und ist klar, was die einzelnen Kasus bedeuten, [...].* (Falko L2, Text 60)

Zu (b) Dieser Punkt muss für jedes Tagset entschieden werden. Will man ein möglichst allgemeines Tagset, das zwar relativ grob, aber dann auch auf viele andere Korpora übertragbar ist, wie Degneaux et al. (1996) oder Izumi, Uchimoto & Isahara (2005)? Oder will man sehr spezifische, für ein kleines Problem oder auf eine spezielle theoretische Annahme aufbauende Tagsets, wie Lippert (2005)? Dass die Zählungen von der Granularität des Tagsets beeinflusst werden, liegt auf der Hand. Ein weiteres Problem ist die Tagging-zuverlässigkeit. Bei automatischer Annotation wird üblicherweise gegen einen so genannten Goldstandard evaluiert, Dazu wird ein Teil des Korpus manuell getaggt. Der Tagger soll diesen Teil dann möglichst gut reproduzieren. Die Grundannahme ist, dass der Goldstandard korrekt ist. Dies wäre nur dann der Fall, wenn (a) das Tagset so eindeutig ist, dass es für jedes Token nur genau eine Möglichkeit gibt und (b) die Annotatoren keine Fehler machen. Beide Grundannahmen sind problematisch.

Für manuelle Annotation gibt es ein anderes Verfahren: hier wird derselbe Text mit denselben Tags nach denselben Richtlinien von mehreren Annotatoren getaggt. Dann werden die Annotationen verglichen und ein so genannter Inter-Annotator-Agreement-Wert (manchmal auch Inter Rater Reliability oder Urteilerübereinstimmungswert genannt, siehe zum Beispiel Carletta 1996, Bortz, Lienert & Boehnke 2000) ermittelt. Mir sind keine Untersuchungen zum Urteilerübereinstimmungswert bei Fehlerannotationen bekannt. Es gibt je-

doch Papiere, die zeigen, dass der Wert bei Aufgaben wie der Lesartenzuweisung schlechter ist als Zufall (Veronis 2001). Ohne eine Angabe des Urteilerübereinstimmungswerts ist also – zumindest bei Kategorisierungen, die nicht völlig trivial sind – eine quantitative Analyse kaum aussagekräftig.

Zu (c) Die Erstellung des Tagsets ist wesentlich für die späteren Abfragemöglichkeiten: nur was kodiert ist, kann automatisch gesucht werden. Fehler können dabei unterschiedlich kategorisiert werden: man kann zum Beispiel nach formaler Art eines Fehlers (Einfügung, Auslassung, Vertauschung) kategorisieren oder die Hypothese über die Fehlerentstehung (zum Beispiel Interferenz der Muttersprache des Lerners oder die mangelnde Beherrschung einer Regel) in das Fehlertag mit einfließen lassen. Man kann die Wortart des Fehlerexponenten mit kodieren oder in bestimmten linguistischen Bereichen arbeiten. Ein Problem ist hier, dass in den vorliegenden Tagsets oft mehrere Informationen in einem Tag zusammengefasst werden. Ein Beispiel hierfür sind die Tags in (8). Problematisch ist dies, wenn die unterschiedlichen Teile des Tags unterschiedlich sicher sind. Nehmen wir an, wir wissen, dass Satz (8) von einem Lerner mit der Muttersprache Polnisch geschrieben wurde. Dann könnte man ein Fehlertag setzen, das einerseits den fehlenden Artikel enthält und andererseits eine Annahme, dass dieser Fehler durch eine Interferenz mit der Muttersprache hervorgerufen wurde. Für die Suchmöglichkeiten wäre dies sicher vorteilhaft. Während sich allerdings wahrscheinlich die meisten Annotatoren sicher sind, dass hier ein Artikel fehlt, ist die Annahme über die Ursache des Fehlers nur eine Hypothese. Unterschiedlich sichere Informationen sollten nicht in einem Tag zusammengefasst werden.

(8) *In Grammatik wird es aber undenkbar dass [...].* (Falko L2, Text 63)

#### 4.4 Korpusstruktur: Unterschiedliche Gewichtung von Vorkommen

Ich möchte an dieser Stelle noch ein weiteres Problem für die Zählung von Fehlern ansprechen, das zwar sehr speziell für das Lernerkorpus Falko erscheint, aber auch eine grundsätzliche Komponente hat. Das Falko-Zusammenfassungskorpus enthält, wie oben beschrieben, Zusammenfassungen von linguistischen und literaturwissenschaftlichen Texten. Eine mögliche Strategie bei Zusammenfassungen ist es, bestimmte als wichtig erkannte Wörter und Phrasen direkt zu übernehmen. Diese Strategie wird von Lernern und Muttersprachlern angewendet, von Lernern allerdings deutlich stärker. Beispiel (9) zeigt alle Übernahmen von zwei oder mehr Wörtern eines Ausschnitts des Lernertextes 33, daneben der relevante Ausschnitt des Vorlagentexts.<sup>13</sup> Die Frage, die sich hier stellt, ist, wie die übernommenen Sequenzen gezählt werden können. Einige Übernahmen sind so, dass sie in einer Zusammenfassung des Ursprungstextes auftreten können, ohne dass man sie als Textübernahme

<sup>13</sup> Die Übernahmen wurden automatisch von dem Plagiatsentdeckungsprogramm WCopyfind 2.6 <http://plagiarism.phys.virginia.edu/Wsoftware.html> markiert.

werten muss (*des Realismus, tritt er, ...*) – der Lerner hätte sie wahrscheinlich auch selbständig so produziert. Bei anderen ist es unklar, ob der Lerner sie selbst hätte produzieren können (*sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts*). Das Problem ist nicht einfach lösbar. Mir ist hier nur wichtig, dass man die übernommen Textpassagen kenntlich machen muss, so dass sich die Auswertungen darauf beziehen können.

(9)

<p>Falko L2, 33, Ausschnitt, Übernahmen kursiv</p>	<p>Vorlage, Ausschnitt (aus: Sprengel, Peter (1998): III. Stile und Richtungen. 1. Realismus In: ders.: Geschichte der deutschsprachigen Literatur 1870–1900. Von der Reichsgründung bis zur Jahrhundertwende. München: Verlag C.H. Beck, S. 99–101)</p>
<p>Der Realismus ist eine überragende geistige und künstlerische Tendenz des 19. Jahrhunderts. Er erstreckt sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts. In der Literatur tritt er unter verschiedenen Namen auf. Die Literatur- und Kunstgeschichte einigte sich erst im nachhinein über die Grenzen des Realismus. Anfangs haben sich die Naturalisten hauptsächlich als Realisten verstanden. Hieraus erkennt man, dass die genauere Bezeichnung des Realismus nach dem damaligen Sprachgebrauch noch nicht vorhanden war.</p>	<p>Der Realismus ist die überragende geistige und künstlerische Tendenz des 19. Jahrhunderts. Noch in der ersten Hälfte des Jahrhunderts einsetzend, und dort vor allem als Gegenbewegung zu klassisch-romantischen Kunstauffassungen begründet, erstreckt er sich als international weit ausgreifende Epochenströmung bis gegen Ende des Jahrhunderts. Freilich tritt er dabei unter einer Reihe verschiedener Namen auf, die vielfach in Konkurrenz zu einem eigentlichen Realismus (im engeren Sinne) stehen, über dessen Grenzen sich Literatur- und Kunstgeschichte erst im nachhinein verständigt haben und z. T. auch heute noch uneins sind. So haben sich z. B. die Naturalisten zunächst und hauptsächlich als Realisten verstanden, was nicht nur einiges über das Selbstverständnis dieser Bewegung verrät, sondern auch beweist, daß der Begriff des Realismus nach dem damaligen Sprachgebrauch noch nicht besetzt, jedenfalls nicht mit der gleichen Verbindlichkeit wie heute als Bezeichnung für eine – im wesentlichen in Distanz zum Naturalismus verharrende – literatur- und kunstgeschichtliche Richtung (inner- oder unterhalb jenes oben umrissenen umfassenden Realismus) etabliert war.</p>

Dieses sicherlich sehr spezielle Problem deutet auf ein allgemeines hin: Kann man jedes Vorkommen eines Wortes/einer Kategorie etc. in einem Korpus zählen wie jedes andere auch? Diese Frage ist bisher in der Literatur eher stiefmütterlich behandelt worden, kann aber starke Auswirkungen auf die Analyse haben. Es ist bekannt, dass Wörter, die einmal in einen Diskurs ein-

geführt werden, oft aufgegriffen werden (der so genannte Clumpiness-Effekt). Quantitative Analysen der Lexik oder der morphologischen Produktivität (Baayen 2001) müssen darauf Bezug nehmen. Szmrecsanyi (2006) zeigt, dass Persistenz – also das Wiederholen von bereits im Diskurs vorgekommenen Strukturen – auch für morphosyntaktische Eigenschaften eine Rolle spielt.

In großen Korpora muss man damit rechnen, dass Texte oder Textteile mehrfach vorkommen, zum Beispiel Agenturmeldungen, die von verschiedenen Zeitungen wörtlich übernommen werden. Wenn in einer solchen Agenturmeldung ein Vorkommen der zu untersuchenden Kategorie (zum Beispiel eine seltene Wortbildung) auftaucht, das dann aber von fünf Zeitungen übernommen wird, zählt man es dann einmal oder fünfmal (siehe Clough & Gaizauskas, erscheint, für eine ausführliche Diskussion von Textwiederverwendung).

Mehrfachvorkommen des gleichen Textes ist für Webkorpora ein besonderes Problem – Verfahren zur duplicate detection spielen hier eine wichtige Rolle (Lüdeling, Evert & Baroni 2007; Baroni & Kilgarriff 2006).

Man kann das Problem sicher nicht generell lösen, da für unterschiedliche Fragestellungen und Phänomene unterschiedliche Verfahren eine Rolle spielen. Ein Desiderat wäre allerdings, dass Textübernahmen und Mehrfachvorkommen von Texten im Korpus gekennzeichnet sind, so dass für jede Analyse deutlich wird, was gezählt wird.

## 5. Architektur

In Abschnitt 3 sollte deutlich werden, dass alle Korpusauswertungen auf Interpretation der Daten beruhen. In 4 habe ich exemplarisch anhand des Lernerkorpus Falko dargestellt, dass die Faktoren (a) Zielhypothese/Interpretation, (b) Tagset/Vergaberichtlinien und (c) Korpusstruktur die Kategorisierung und daher auch alle darauf aufbauenden quantitativen Analysen beeinflussen.

In diesem Abschnitt möchte ich nun dafür plädieren, die Grundlagen für jede quantitative Analyse so transparent wie möglich darzustellen. Dazu gehört dass Tagsets, Richtlinien und Zuverlässigkeit veröffentlicht werden und, wenn möglich, das Korpus frei zur Verfügung steht. Dazu gehört aber auch, dass alternative und konfligierende Hypothesen und Kategorisierungen in einem Korpus dargestellt werden können. Dies ist in den ‚traditionellen‘ flachen Korpusarchitekturen nicht möglich, wohl aber in Mehrebenenarchitekturen.

In ‚flachen‘ Architekturen werden Kategorisierungen in einem ‚Tabellenformat‘ an Tokens gehängt (daher stammt auch der Begriff ‚positionelle Annotation‘; zu jeder Korpusposition wird eine festgelegte Anzahl von Annotationsebenen – oder Spalten zur Verfügung gestellt). Diese Architektur ist gut geeignet für sehr große, automatisch annotierte Korpora, da sie gut zu indizieren und schnell zu durchsuchen ist. In einer solchen Architektur können konfligierende Annotationen allerdings nicht dargestellt werden.

Außerdem ist es nicht einfach möglich, neue Ebenen einzufügen. Dies führt dann oft zu der oben erwähnten Zusammenfassung verschiedener Informationen in ein Tag.

In einer Tabellenarchitektur ist es auch nicht möglich, Informationen an eine Sequenz von Tags zu hängen. Wie oben diskutiert, beziehen sich viele Fehler (und andere Annotationen) aber auf Sequenzen und nicht auf einzelne Tokens. Daher werden heute oft Baummodelle (meist XML) verwendet, die die zu annotierenden Token durch Anfangs- und Endtags einschließen. Ein Beispiel aus einem englischen Lernerkorpus ist (10) (aus Izumi, Uchimoto & Isahara 2005).

(10) *I belong to two baseball <n\_num crr=„teams“>team</n\_num>*

In XML-Dateien können im Prinzip mehrere Tokens durch ein Tag eingeschlossen werden. Allerdings ist es auch in XML-Dateien schwer möglich, konkurrierende Analysen zu kodieren. Dies liegt zum einen an der Baumstruktur von XML, die konfligierende Hierarchien nicht zulässt (das Problem ist technisch lösbar) und zum anderen daran, dass Daten und Annotationen in derselben Datei gespeichert werden.

In den letzten Jahren wurden für multimodale Korpora Mehrebenenmodelle mit standoff Architektur entwickelt (Bird & Liberman 2001, Carletta et al. 2003, Dipper et al. 2004, Wörner et al. 2006, Wittenburg, erscheint), in denen alle Korpusebenen (Sprachsignal, Transkription, Gestik) und Annotationsebenen in getrennten Dateien gespeichert werden, die alle auf eine gemeinsame Referenzebene (timeline) verweisen.

Die Mehrebenenarchitektur ist ursprünglich nicht für die Modellierung von konfligierenden Analysen entwickelt worden, kann aber dafür gebraucht werden. Statt eines Zeitstrahls ist der Originaltext die Referenzebene. Alle anderen Ebenen sind dann unabhängig davon. Konfligierende Hypothesen und Tagsets können dargestellt werden. Ich möchte dies wieder am Beispiel von Falko erläutern (für eine technischere Darstellung siehe Lüdeling et al. 2005).

Falko wird zurzeit im Partitur-Editor EXMARaLDA kodiert.<sup>14</sup> Jedes Token eines Lernertextes bekommt eine Position in der Partitur – dadurch wird

<sup>14</sup> EXMARaLDA ist an der Universität Hamburg zur Darstellung von gesprochenen Korpora entwickelt worden und frei verfügbar <http://www1.uni-hamburg.de/exmaralda/>. EXMARaLDA hat zwar mehrere Ebenen, ist aber bisher nicht standoff. Bisher gibt es noch wenige frei verfügbare und gut zu bedienende Tools, mit denen echte standoff-Architekturen kodiert werden können. Solche befinden sich aber an vielen Orten in der Entwicklung (siehe z. B. Wörner et al. 2006 und die darin beschriebenen Tools für die linguistischen Sonderforschungsbereiche in Tübingen, Berlin/Potsdam und Hamburg und die Vorarbeiten für das DDD-Projekt in Dipper et al. 2004; Lüdeling, Poschenrieder & Faulstich 2005; Faulstich, Leser & Lüdeling 2005). Im Moment werden für solche Architekturen auch mächtige Such- und Darstellungstools entwickelt (Schmidt & Wörner 2005, Vitt 2005, Siemen, Lüdeling & Müller 2006).

der Zeitstrahl simuliert. Alle Annotationsebenen sind in getrennten Spuren gespeichert. Man kann so viele Annotationsebenen hinzufügen, wie man möchte. Das hat die Vorteile, dass man sich nicht von vorneherein auf die Menge und Granularität der Tagsets festlegen muss und dass man konfligierende Zielhypothesen darstellen kann. Alle Annotationsentscheidungen und Tags stehen im Netz zur Verfügung, jede quantitative Aussage kann daher überprüft werden.

Im Moment wird jeder Falko-Text zunächst automatisch mit dem Tree-Tagger mit Wortart und Lemma getaggt (dabei ist zu bedenken, dass die Qualität des Taggings gegenüber Zeitungstexten sinkt, weil durch die hohe Anzahl an Rechtschreibfehlern der Lexikonzugriff oft fehlschlägt und durch Wortstellungsfehler die statistische Komponente des Taggers nicht greifen kann, siehe dazu auch van Rooy & Schäfer 2003). Dann wird eine explizite Zielhypothese angegeben. Alle Fehler werden im Bezug auf diese Zielhypothese hin getaggt. Wir haben dann verschiedene Schichten für jeden Annotationsbereich (zurzeit Definitheit, Wortstellung, Kongruenz & Rektion; an weiteren Bereichen wird gearbeitet).

Beispiel (11) zeigt einen Lerneratz in der Tabellendarstellung (die Tagsets kann ich aus Platzgründen nicht im Einzelnen erläutern, eine Dokumentation findet sich unter <http://www2.hu-berlin.de/korpling/projekte/falko/index.php>). Die erste Zeile ([word]) ist die Referenzzeile. Jedes Token bekommt eine Korpusposition. Unter [target\_hypothesis] ist die Zielhypothese gespeichert. Dabei sind nur solche Tokensequenzen aufgeführt, die sich vom Originaltext unterscheiden.

Abweichende Zielhypothesen können jederzeit hinzugefügt werden, so dass unterschiedliche Analysen der Daten möglich sind. Im Unterschied zu vielen bisher vorliegenden Korpora sind alle Interpretationen sichtbar, die Ergebnisse werden so transparent und reproduzierbar. Auf eine ähnliche Art können andere spezialisierte Korpora kodiert werden.

Ein weiteres Vorteil einer Mehrebenenarchitektur besteht darin, dass dadurch das verteilte Arbeiten an denselben Daten möglich ist.

(11)

EXAMEN DA Partizip II con 1812 /home/peterr/Wortz/alkon\_output/003.xml

File Edit View Iter Event Timeline Format Segmentation Help

38| 39| 40| 41| 42| 43| 44| 45| 46| 47| 48| 49| 50| 51| 52| 53| 54| 55| 56| 57

[word]	Das	bedeutet	, dass	fur	unsere	Lexikon	notig	ist	,	die	Wörter	zu	erkennen	und	die	unterscheiden	Unterscheiden	bede			
[lemma]	d	bedeuten	, dass	fur	unsere	Lexikon	notig	sein	,	d	Wort	zu	erkennen	und	d	unterscheiden	Unterscheiden	bede			
[pos]	\$	PDS	\$	KOUS	APPR	PPOSAT	NN	ADJD	VAFIN	\$	ART	NN	PTK	ZU	VVINF	KON	PRELS	VVINF	VVFIN		
[target_hypothesis]	es für unser Lexikon nötig ist																				
[transcripior_comment]																					
[kongruenz_id]	x																				
[kongruenz_tag]	GsFem.Neut																				
[fktion_id]																					
[fktion_tag]																					
[matrix_satz]	x																				
[matrix_satz_felder_2]	VF_MS	LSK_MS	NF_MS	na																VF_MS	LSK_A
[konstituenten_satz_1]	x																				
[konstituenten_satz_1_felder]	LSK_KS MF_MS																				
[konstituenten_satz_1_felder_2]	f_kon_npr																				
[konstituenten_satz_2]	x																				
[konstituenten_satz_2_felder]	MF_KS	RSK_XS	MF_KS	RSK_KS																RSK_KS	f_kon_npr
[konstituenten_satz_2_felder_2]	MF_KS	RSK_XS	MF_KS	RSK_KS																RSK_KS	f_kon_npr
[konstituenten_satz_3]																					
[konstituenten_satz_3_felder]																					

## 6. Zusammenfassung

Jede quantitative Analyse von Korpusdaten stützt sich auf eine vorhergehende Kategorisierung. In diesem Artikel habe ich gezeigt, wie stark die Kategorisierung die Ergebnisse beeinflusst. Dabei spielen die Tagsets und die Richtlinien für deren Vergabe genauso eine Rolle wie die Zielhypothese und die Zusammensetzung des Korpus. Ich habe dies exemplarisch anhand des Lernerkorpus Falko und der Fehlerannotation diskutiert.

Da es nicht möglich ist, eine Interpretation der Daten zu vermeiden, habe ich dafür plädiert, die getroffenen Entscheidungen, wo immer möglich, explizit zu machen, so dass sie für die weitere Arbeit nachvollziehbar sind. Für spezialisierte Korpora ist eine standoff Mehrebenenarchitektur eine Möglichkeit, das zu erreichen. Ich habe an Falko gezeigt, wie eine solche Mehrebenenarchitektur, die eigentlich für multimodale Korpora entwickelt wurde, aussehen kann.

## 7. Danksagung

Falko würde es ohne die Mitarbeit von vielen Kolleginnen, Kollegen und Studierenden nicht geben. Maik Walter und Karin Schmidt sind verantwortlich für die Akquisition der Zusammenfassungstexte und die Gesamtkonzeption des Korpus. Peter Siemen hat das Suchprogramm erstellt und alles kompetent technisch unterstützt. Hanna Acke, Eva Lippert und Seanna Doolittle haben kommentiert und annotiert. Allen meinen herzlichen Dank!

## 8. Literatur

- Baayen, R. Harald (2001): *Word Frequency Distributions*. Dordrecht: Kluwer.
- Baroni, Marco/Kilgarriff, Adam (2006): *Large linguistically-processed web corpora for multiple languages*. Conference Companion of EACL 2006 (11<sup>th</sup> Conference of the European Chapter of the Association for Computational Linguistics), East Stroudsburg PA: ACL. S. 87–90.
- Baumgarten, Nicole/Herkenrath, Annette/Schmidt, Thomas/Wörner, Kai/Zeevaert, Ludger (erscheint 2006): *Studying Connectivity with the Help of Computer-Readable Corpora: Some Exemplary Analyses from Modern and Historical, Written and Spoken Corpora*. In: Hohenstein, Christiane/Pietsch, Lukas/Rehbein, Jochen (eds.): *Connectivity in Grammar and Discourse*. Amsterdam: Benjamins.
- 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, S. 577–591.
- Biber, Douglas (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Bird, Steven/Liberman, Mark (2001): *A formal framework for linguistic annotation*. In: *Speech Communication* 33(1,2), S. 23–60.
- Bloomfield, Leonard (1933): *Language*. London: Allen & Unwin.
- Bortz, Jürgen/Lienert, Gustav A./Boehnke, Klaus (2000): *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer-Verlag.

- Bosch, Peter/Rozario, Tom/Zhao, Yufan: Demonstrative Pronouns and Personal Pronouns. German der vs. er. Proceedings of the EACL 2003. Budapest. Workshop on The Computational Treatment of Anaphora.
- Carletta, Jean (1996): Assessing agreement on classification tasks: the kappa statistics. In: *Computational Linguistics*, 22(2), S. 249–254.
- Carletta, Jean/Evert, Stefan/Heid, Ulrich/Kilgour, Jonathan/Robertson, Judy/Voormann, Holger (2003): The NITE XML Toolkit: flexible annotation for multi-modal language data. In: *Behavior Research Methods, Instruments, and Computers* 35(3), S. 353–363.
- Carstensen, Kai-Uwe/Ebert, C./Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (eds.) (2004): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg/Berlin: Spektrum Akademischer Verlag.
- Cherubim, Dieter (Hg.) (1980): *Fehlerlinguistik. Beiträge zum Problem der sprachlichen Abweichung*. Tübingen: Niemeyer.
- Clough, Paul/Gaizauskas, Robert (erscheint 2007): Corpora and text re-use. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Corder, Stephen Pit (1973): *Introducing Applied Linguistics*. Harmondsworth: Penguin.
- Corder, Stephen Pit (1981): *Error Analysis and Interlanguage*. Oxford: Oxford University Press.
- Degneaux, Estelle/Denness, Sharon/Granger, Sylviane/Meunier, Fanny (1996): *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain.
- Dipper, Stefanie/Faulstich, Lukas/Leser, Ulf/Lüdeling, Anke (2004): Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In: *Workshop on XML-based richly annotated corpora*. Lisbon, Portugal.
- Ellis, Rod (1994): *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Evert, Stefan (2005): *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Online verfügbar unter <http://www.collocations.de/phd.html>.
- Evert, Stefan/Fitschen, Arne (2001): *Textkorpora*. In: Carstensen, Kai-Uwe/Ebert, C./Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (Hg.): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg/Berlin: Spektrum Akademischer Verlag. S. 369–376.
- Featherston, Sam (in diesem Band): Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie.
- Golcher, Felix (2006): *Statistical text segmentation with partial structure analysis*. In: *Proceedings of KONVENS 2006*. Konstanz.
- Granger, Sylviane (2002): A bird's-eye view of learner corpus research. In: Granger, Sylviane/Hung, Joseph/Petch-Tyson, Stephanie (Hg.): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. S. 3–33.
- Granger, Sylviane (erscheint): *Learner corpora*. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Gries, Stefan Th. (2003): *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London/New York: Continuum Press.
- Grommes, Patrick/Maik, Walter (2006): Fortgeschrittene Lernervarietäten. Vortrag auf der 28. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Bielefeld. Folien unter <http://userpage.fu-berlin.de/~maik/dgfs06.html>.

- Haß-Zumkehr, Ulrike (2002): Das Wort in der Korpuslinguistik. Chancen und Probleme empirischer Lexikologie. In: Ágel, Vilmos/Gardt, Andreas/Haß-Zumkehr, Ulrike/Roelcke, Thorsten (Hg.): Das Wort. Seine strukturelle und kulturelle Dimension. FS für Oskar Reichmann zum 65. Geburtstag. Tübingen: Max Niemeyer Verlag. S. 45–70.
- Heringer, Hans Jürgen (1995): Aus Fehlern lernen. CD-ROM für Win9x/NT. Augsburg.
- Himmelman, Nikolaus (1998): Documentary and descriptive linguistics. In: *Linguistics* 36, S. 161–195.
- Hinterhölzl, Roland/Petrova, Svetlana/Solf, Michael (2005): Diskurspragmatische Faktoren für Topikalität und Verbstellung in der althochdeutschen Tatianübersetzung (9. Jh.). In: Ishihara, S./Schmitz, M./Schwarz, A. (Hg.): Approaches and Findings in Oral, Written and Gestural Language, Interdisciplinary Studies on Information Structure (ISIS) 3. Potsdam: Universitätsverlag Potsdam. S. 143–182.
- Hirschmann, Hagen (2005): Platzhalterphrasen bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. Staatsexamensarbeit, Humboldt-Universität zu Berlin.
- Izumi, Emi/Uchimoto, Kiyotaka/Isahara, Hitoshi (2005): Error Annotation for a corpus of Japanese Learner English. In: Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005).
- Kepser, Stephan/Reis, Marga (2005) (Hg.): Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton de Gruyter.
- Lennon, Paul (1991): Error and the very advanced learner. In: *International Review of Applied Linguistics* 29, S. 31–44.
- Lippert, Eva (2005): Probleme von Nichtmuttersprachlern mit der Definitheit von Nominalphrasen. Magisterarbeit, Humboldt-Universität Berlin.
- Lüdeling, Anke (2000): Particle verbs in NLP lexicons In: Heid, Ulrich/Evert Stefan/Lehmann, Egbert/Rohrer, Christian (Hg.): Proceedings of the Ninth EURALEX International Congress, EURALEX 2000. IMS, Stuttgart. S. 625–630.
- Lüdeling, Anke (2007): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes, Patrick/Walter, Maik (Hg.): Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Niemeyer.
- Lüdeling, Anke/Evert, Stefan/Baroni, Marco (erscheint 2006): Using Web data for linguistic purposes. In: Hundt, Marianne/Biewer, Caroline/Nesselhauf, Nadja (Hg.): *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi.
- Lüdeling, Anke/Poschenrieder, Thorwald/Faulstich, Lukas C. (2005): DeutschDiachron-Digital – Ein diachrones Korpus des Deutschen. In: *Jahrbuch für Computerphilologie* 2004, S. 119–136. Online verfügbar unter <http://computerphilologie.tu-darmstadt.de/jahrbuch/jb6-content.html>.
- Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005): Multi-level error annotation in learner corpora. In: Proceedings of corpus linguistics 2005. Birmingham. Online verfügbar unter <http://www2.hu-berlin.de/korpling/projekte/falko/FALKO-CL2005.pdf>.
- Manning, Christopher/Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
- Milde, Jan-Torsten/Gut, Ulrike (2002): A prosodic corpus of non-native speech. In: Bel, B./Marlien, I. (Hg.): Proceedings of the Speech Prosody 2002 conference, 11–13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage. S. 503–506.
- Moisl, Hermann (erscheint 2007): Clustering techniques. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Müller, Stefan (in diesem Band): Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus.

- Mukherjee, Joybrato (2002): *Korpuslinguistik und Englischunterricht. Eine Einführung*. Frankfurt: Peter Lang.
- Nesselhauf, Nadja (2005): *Collocations in Learner Corpora*. Amsterdam: John Benjamins.
- Oakes, Michael (1998): *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Ostler, Nicholas (2007): *Corpora of less studied languages*. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Pravec, N. A. (2002): *Survey of learner corpora*. In: *ICAME Journal* 26, S. 81–114. Available at <http://nora.hd.uib.no/icame/ij26/>.
- Römer, Ute (2006): *Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for the Future*. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), S. 121–134.
- van Rooy, Bertus/Schäfer, Lande (2003): *Automatic POS tagging of a learner corpus: the influence of learner error on tagger accuracy*. In: Archer, D./Rayson, P./Wilson, A./McEnery T. (Hg.) *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster University: UCREL. S. 835–844.
- Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In: *Proceedings of International Conference on New Methods in Language Processing*. September 1994. Online verfügbar unter <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/DecisionTreeTagger.html>.
- Schmidt, Thomas/Wörner, Kai: *Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA*. *Gesprächsforschung Online* 2005 Online verfügbar unter <http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf>.
- Siemen, Peter/Lüdeling, Anke/Müller, Frank Henrik (2006): *FALKO – Fehler-Annotiertes LernerKORpus des Deutschen*. In: *Proceedings of Konvens 2006*. Konstanz
- Szmrecsanyi, Benedikt (2006): *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin/New York: Mouton de Gruyter.
- Tono, Yukio (2003): *Learner corpora: design, development, and applications*. In: *Pre-conference workshop on learner corpora, Corpus Linguistics 2003*, Lancaster.
- Veronis, Jean (2001): *Sense tagging: does it make sense?* Paper presented at the *Corpus Linguistics 2001 Conference*, Lancaster, U.K. 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*. MA thesis, Lancaster University.
- Wittenburg, Peter (erscheint 2007): *Preprocessing of multimodal corpora*. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. (Berlin: Mouton de Gruyter).
- Wörner, Kai/Witt, Andreas/Rehm, Georg/Dipper, Stefanie (2006): *Modelling Linguistic Data Structures*. In: *Proceedings of 'Extreme Markup Languages' 2006*, Montreal.
- Xiao, Richard (2007): *Theory-driven corpus research: using corpora to inform aspect theory*. In: Lüdeling, Anke/Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.