AG 8: Korpora und Grammatik nichtstandardisierter Sprache

Christiane Pankow	Anke Lüdeling
Universität Göteborg	Humboldt-Universität zu Berlin
Christiane.pankow@tyska.gu.se	anke.luedeling@rz.hu-berlin.de

Althochdeutsch: Die Rolle der Informationsstruktur bei der Herausbildung von Wortstellungsregularitäten im Germanischen

Roland Hinterhölzl und Svetlana Petrova
Humboldt-Universität zu Berlin, SFB 632 „Informationsstruktur“, B4
roland.hinterhoelzl@rz.hu-berlin.de; s.petrova@hu-berlin.de

Der Beitrag stellt die Methodik eines Forschungsprojekts vor, das die Rolle der Informationsstruktur bei der Entwicklung der Wortstellungsregularitäten in den germanischen Sprachen untersucht. Es ist Teil eines DFG-finanzierten Sonderforschungsbereichs, der die Kodierung informationsstruktureller Phänomene im diachronen sowie synchron-typologischen Plan erfassen und in ein Modell der Interaktion von Pragmatik und Kerngrammatik integrieren will. Die diachronen Untersuchungen basieren auf der Annahme, dass für die Zwecke der informationsstrukturellen Optimierung neuartige Konstruktionen und Satzmuster entstehen, die Varianz in das syntaktische System einer Sprache bringen und damit Voraussetzungen für Sprachwandel induzieren. Diese Ansicht eröffnet neue Perspektiven für die Erklärung von Wortstellungsvarianz und Syntaxwandel, wie sie sich insbesondere am Beispiel der historischen Syntax der germanischen Sprachen im Bereich der Verbstellung präsentieren.

Obwohl dieses Herangehen innovativ und vielversprechend ist, erfordert seine Umsetzung einige empirisch-philologische und theoretische Vorentscheidungen, die hier am Beispiel des Althochdeutschen erörtert werden. Aufgrund seiner Überlieferungssituation ist das Althochdeutsche eine in syntaktischer Sicht äußerst schwer zu untersuchende Sprachstufe. Der Großteil der Quellen besteht entweder aus eng am Original ausgerichteten Übersetzungen oder aus metrisch gebundenen Texten. In beiden Fällen ist damit zu rechnen, dass die bezeugten Wortfolgen nicht den autochthonen Sprachgebrauch abbilden, sondern aus der Originalsprache übernommen bzw. der Verstechnik geschuldet sind. Demnach soll zunächst das Problem der Sicherstellung verlässlicher, authentischer Satzmuster und Konstruktionstypen gelöst werden. Dazu wurde ein Modell erarbeitet, das sich auf die Erfassung von Differenzbelegen, d.h. vom Original abweichenden Satzmustern im ahd. Tatian (9. Jh.) stützt. Diese Belege werden hinsichtlich grammatischer und informationsstruktureller Merkmale mit dem Annotationstool Exmaralda annotiert. Das Ziel ist, durch Abfragen zu ermitteln, wie die pragmatischen Eigenschaften von Satzkonstituenten ihre Realisierung im Satz beeinflussen und welche Faktoren die Herausbildung neuer Muster bzw. die Generalisierung ursprünglich pragmatisch markierter Muster bedingen.

Daran knüpft die Frage nach der Identifikation informationsstruktureller Kategorien wie Topik und Fokus in Korpussprachen. Es handelt sich um kontroverse, stark umstrittene Begrifflichkeiten, die in den verfügbaren historischen Daten nicht in jedem Fall eindeutig zu identifizieren sind. Um möglichst theorieneutral zu bleiben, arbeiten wir kumulativ, d.h. wir ermitteln aus der bisherigen Forschung konstitutive Merkmale der informationsstrukturellen Hauptkategorien bei konsequenter Trennung der Ebenen ‘gegeben/neu’, ‘Topik/Kommentar’ und ‘Fokus/Hintergrund’. Ferner ist in unseren Daten der Zugang zur Prosodie, in der sich informationsstrukturelle Phänomene primär äußern, stark eingeschränkt. Das verlagert das

Gewicht der informationsstrukturellen Analyse auf die Interpretation kontextuell ableitbarer pragmatischer Eigenschaften, die Auskunft über den Status der jeweiligen Satzkonstituente im entsprechenden Satz erlauben.

Diesen Vorentscheidungen folgend, wurde ein eigenes Annotationsschema für die Erfassung grammatischer und informationsstruktureller Kategorien entwickelt und an ca. 1.600 Differenzbelegen aus dem Tatian erprobt. Die Umsetzung wird an ausgewählten Beispielen aus dem Althochdeutschen demonstriert.

Identifizierung der Satzgrenzen in der Newsgroupssprache: computer- und textlinguistische Probleme

Cristina Onesti
Università degli Studi di Torino
cristina_onesti@yahoo.it

Ein Korpus muss für die linguistische Analyse vorbereitet werden. Zusammen mit der Tokenisierung wird nach Satzgrenzen gesucht. Dabei ist der Punkt am Schluss des Satzes besonders problematisch, da er hochgradig ambig ist: Er kann am Satzende, in Daten oder Abkürzungen (*14. Okt.)*, in Zahlen (*3.100*), in Formeln etc. vorkommen.
Die automatische Zuweisung von Bedeutung zu jedem Satzzeichen ermöglicht *within sentence*-Queries, sie wird aber vom Kommunikationstyp erschwert. Wie schon Palmer/Hearst 1997 vermerken, hängt die Anzahl der Abkürzungen und ambigen Satzzeichen stark vom Textgenre ab. Die Satzgrenzenerkennung in Texten computervermittelter Kommunikation (CMC) weist zusätzlich zu den von Grefenstette/Tapanainen 1994 analysierten Disambiguierungsproblemen weitere Schwierigkeiten auf: Satzzeichen fehlen oft; dafür findet man häufiger Smileys, ASCII-Art und Signatures, die bisher im Tagging der standardisierten Schriftsprache außer Acht gelassen wurden. Selbst in einer asynchronen Kommunikationssituation, die die Möglichkeit zur Korrektur und zur Planung bietet, kann man eine nicht normkonforme Sprache feststellen, die eng an die gesprochene Sprache angelehnt ist. Schnelles Tippen hat dabei einen höheren Wert als die Orthographienorm.
Die Newsgroupsprache ist dabei wahrscheinlich problematischer als andere Varietäten: Termini der Fachsprachen kommen vor, wobei Benutzer in technischen Diskussionsforen auch Links, Anhänge, Programmcode, Akkorde usw. austauschen. Die Diskursübernahme (Turn-taking), wird nach einer bestimmten hierarchischen Textstruktur organisiert und Postings weisen das Quoting auf, d.h. die Gesamt- oder Teilwiederaufnahme voriger Textbeiträge, um darauf gezielt zu antworten. (Corino 2007)
In Erweiterung der bisherigen Forschungsergebnisse zur Annotation ist es deshalb wünschenswert, eben die genannten textuellen Eigenschaften zu beachten. Die ersten Ergebnisse der Arbeit zeigen die funktionelle Ausnutzung von in spitzen Klammern markierten (*quoted*) Textteilen. Daraus lassen sich Vorschläge für das Tagging des Quotings, des Turn-takings und der Leerzeile ableiten, die anhand eines Korpus und zusätzlichen Lexikons (z.B. von Smileys und Verabschiedungsformeln) ebenfalls skizziert werden.
Aus der Analyse des aus deutschen Newsgroupnachrichten bestehenden Korpus NUNC-De (Newsgroups UseNet Corpora - Deutsch, mehr als 300 Mill. Tokens) ergeben sich Eigenschaften, die die Verbesserung der Annotation erwarten lassen: aktuelle sprachliche Daten aus den Jahren 2002-2006, die grundlegend für die Analyse der aktuellen Tendenzen der Gebrauchssprache sind, sowie die für solche Online-Forentexte typische Spontaneität, die als eine interessante, der Mündlichkeit nahe kommende Schriftlichkeit betrachtet werden kann.

Corino, Elisa (2007) NUNC (Newsgroup UseNet Corpora). Questioni metodologiche ed aspetti della testualità. In: Manuel Barbera, Elisa Corino & Cristina Onesti (Hrsg.) (2007), *Corpora e linguistica in rete*, Guerra Edizioni, Perugia, 225-252.

Grefenstette, Gregory & Tapanainen, Pasi (1994) What is a Word, What is a Sentence? Problems of Tokenization. In: *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX '94)*, Research Institute for Linguistics - Hungarian Academy of Sciences, Budapest, 79-87.

Palmer, David D. & Hearst, Marti A. (1997) Adaptive Multilingual Sentence Boundary Disambiguation. In: *Computational Linguistics 23(2)*, 241-267.

Die Annotation von verstehenshemmenden Einheiten des gesprochenen Südtiroler Dialekts

Magdalena Putz
Europäische Akademie Bozen
Magdalena.Putz@eurac.edu

Aus dem in Südtirol (Alto Adige, Italien) vorherrschenden Spannungsfeld zwischen dem deutschen Dialekt und der deutschen Hochsprache ergeben sich Verstehensschwierigkeiten, die sich unter anderem auf die Kommunikation zwischen deutschen und italienischen MuttersprachlerInnen auswirken. (Sitta 1994: 14ff) Zum einen liegen die Verstehensschwierigkeiten an einer gewissen Unfähigkeit der Dialekt-SprecherInnen, sich in der Hochsprache privat zu unterhalten (Lanthaler 1990: 57-81), zum anderen haben Nicht-Dialekt-SprecherInnen oft große Mühe, den Dialekt zu verstehen.
Ziel des Dissertationsprojekts an der Universität Turin in Zusammenarbeit mit der Europäischen Akademie Bozen ist es, schwer verständliche dialektale Einheiten auszumachen, die in einem zweiten Schritt in Referenzmaterialien für Nicht-Dialekt-SprecherInnen eingearbeitet werden können.
Zu diesem Zweck wird ein Korpus von Visitengesprächen zwischen ÄrztInnen mit italienischer Muttersprache und meist Dialekt sprechenden PatientInnen erstellt. In diesem Korpus sollen jene Elemente annotiert werden, die für die nicht Dialekt sprechenden ÄrztInnen verstehenshemmend sind.
Die Aufnahmen werden im Partitureditor EXMARaLDA (Schmidt / Wörner 2005) nach den Richtlinien von HIAT (Ehlich / Rehbein 1976) transkribiert.
In diesem Beitrag sollen erste Analyseergebnisse vorgestellt sowie das angewandte Annotationsschema diskutiert werden.

Ehlich, Konrad / Rehbein, Jochen (1976): Halbinterpretative Arbeitstranskriptionen (HIAT). In: *Linguistische Berichte* 46. 21-41.

Lanthaler, Kurt (1990): Dialekt und Zweisprachigkeit. In: Lanthaler, Franz (Hg.) (1990): *Mehr als eine Sprache/Più di una lingua*. Meran: Alpha Beta. 57 – 81.

Schmidt, Thomas / Wörner, Kai (2005): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*. Ausgabe 6, 171-195 www.gespraechsforschung-ozs.de.

Sitta, Horst (1994): Im Dialekt leben. In: Lanthaler, Franz (Hg.) (1994): *Dialekt und Mehrsprachigkeit/Dialetto e plurilinguismo. Beiträge eines internationalen Symposiums*. Meran: Alpha Beta. 13 – 26.

Zur Syntax des mündlichen Erlebnisberichts
am Beispiel der deutschen Minderheit im Gebiet Omsk und Krasnojarsk

Valentina Djatlowa
Krasnojarsk (Russland)
dva-krsk@yandex.ru

Die heutigen deutschen Mundarten in Sibirien sind meist Mischmundarten mit vielen Varianten. Die hier besprochenen Dialekte sind Vertreter hochdeutscher Mundarten und zwar des Obersächsischen, des Oberhessischen (Gebiet Omsk), des Wolhynischen und des Schwäbischen (Gebiet Krasnojarsk). Die Sprachsituation vor Ort ist einzigartig und durch ihre Heterogenität charakteristisch.

Im Jahre 1979 lebten im Gebiet Omsk 120 806 Deutsche und im Gebiet Krasnojarsk 54 518 Deutsche. Die größte Welle der Einwanderung von Deutschen nach Sibirien ist mit dem Zweiten Weltkrieg verbunden. Die Russlanddeutschen wurden aus den Wolgadörfern Kind, Bettinger, Balzer, Kutter, Kukkus, Altwarenburg, Schönchen, Schilling, Katharinental, Dönhof, Gattung, Neudorf, Huk, Kano, Beideck, Wittmann, Straßburg, Jost u.a deportiert. Viele Deutsche haben inzwischen die zu erforschenden Gebiete verlassen und sind nach Deutschland ausgewandert, es gibt aber auch Rückkehrer.

Forschungsobjekte sind die deutschen Dialekte der Dörfer Nikolajewka (Gebiet Krasnojarsk) und Berjosowka (Gebiet Omsk). Ähnlichkeiten und Unterschiede des Dialektgebrauchs werden auf der Basis von Tonaufnahmen und von Abfragen zu den Wenker-Sätzen ermittelt und systematisiert. Es sei hier betont, dass zu Wenkers Fragebogen eine umfangreiche Ergänzung wünschenswert wäre.

Die Dialektsyntax des Russlanddeutschen bleibt bis heute eine Forschungslücke, leider gibt es auf dieser Ebene keine systematisch übergreifende Beschreibung der deutschen Dialekte in Russland.

Die syntaktischen Besonderheiten im Sprachgebrauch der Russlanddeutschen sind in drei Richtlinien zu analysieren: als Interferenzerscheinungen aus dem Russischen, als Besonderheiten der Umgangssprache und als Eigenständigkeit des deutschen syntaktischen Systems. Der Einfluss der deutschen Literatursprache ist in unserem Fall ausgeschlossen und das ist einer der Gründe, warum deutsche Dialekte fast in ihrer ursprünglichen Form erhalten geblieben sind.

In erster Linie fallen folgende syntaktische Phänomene auf: Rahmenverletzung, Reihenfolge der Satzglieder, Wortgruppenbildung, Infinitivgruppen, spezifische Satzanfänge, überwiegende Anzahl der Sätze mit parataktischen Konstruktionen (viele subordinierende Präpositionen fehlen), das Problem der „doppelten Negation“ usw. Auffallend sind die Lehnübersetzungen von einigen syntaktischen Konstruktionen des Russischen ins Deutsche. Man kann von einer engen Symbiose der deutschen und russischen Grammatik als Resultat der kontaktierenden Sprachvarietäten sprechen.

Alle Befragungen wurden unter soziolinguistischem Aspekt durchgeführt, d.h. auf Grund folgender Daten: Alter, Geschlecht, Ehe, Bildungsgrad, Religion, Sprachgebrauch in der Familie und anderen Sprechsituationen, Muttersprache, Dialekt, Mehrsprachigkeit.

Topologische Felder in einem Korpus der gesprochenen Sprache

Christiane Pankow
Universität Göteborg
christiane.pankow@tyska.gu.se

In der Syntax, die die Regeln zur Wortfolge modellartig beschreibt, geht man noch immer konzeptionell von der Schriftsprache aus. Beim gesteuerten Fremdsprachenerwerb orientieren sich daher auch die Lerner an syntaktischen Modellen, die von Strukturen und Eigenschaften der geschriebenen Sprache abgeleitet wurden. Ein Beispiel für solche schriftsprachlichen Kategorien ist der Satz, dessen Strukturen in der gesprochenen und geschriebenen Sprache große Unterschiede aufweisen. Das betrifft insbesondere die Wortfolge. In einem Korpus der gesprochenen Sprache wirft die Sequenz **ich würde dich ja mal noch mal was mit dir besprechen* eine Reihe von grammatischen Problemen auf (z.B. Satzgrenze, Wortfolge, Valenz des Verbs), die gegen ein syntaktisches Modell überprüft werden müssten.
Es ist bereits versucht worden, Besonderheiten der gesprochenen Sprache für den Fremdsprachenunterricht nutzbar zu machen. Selten ist jedoch bisher auf elektronische Korpora zurückgegriffen worden. Mich interessiert hierbei, wie solche Korpora für die Ausarbeitung einer Lernergrammatik nutzbar gemacht werden könnten. Welche grammatischen Kategorien sind für die gesprochene Sprache typischer als für die Schriftsprache? Sind bestimmte grammatische Kategorien häufiger anzutreffen usw.? Übergreifend stellt sich die Frage, wie Lernergrammatiken des Deutschen unter Einbeziehung der gesprochenen Sprache modifiziert werden müssten.
In dem Partitur-Editor EXMARaLDA[1], der besonders für die multimodale Annotation geeignet ist, wurde das Korpus der gesprochenen Sprache *Elizitierte Konfliktgespräche zwischen Müttern und jugendlichen Töchtern*[2] eingelesen, um mit schwedischen Germanistikstudierenden langsichtig ein linguistisches Korpus zu erarbeiten. Eine Zielstellung besteht darin, Korpusanalysen für Lernergrammatiken nutzbar zu machen.
Es sollen zunächst sowohl theoretisch einander ergänzende als auch korpusanalytisch voneinander abhängige Annotationsebenen und Annotationskategorien vorgestellt werden. Dabei soll diskutiert werden, wie im Korpus auftretende Abweichungen der Wortfolge im Kontrast zu gängigen Satzmodellen beschrieben d.h. annotiert werden könnten.
Die Annotierung dieses Korpus der gesprochenen Sprache trägt sowohl Prozess- als auch Resultatcharakter. Durch die Annotation und Annotationsproblematik können sich die Studierenden einerseits operational mit der Beschreibung spontan gesprochener Sprache auseinandersetzen, andererseits können durch linguistisch annotierte Ausschnitte der spontanen Rede systematische Einsichten in die Struktur der natürlichen Sprache gewonnen werden.

Pankow, Christiane (2007): Korpora der gesprochenen Sprache und Fremdsprachengrammatik. In: *Deutsche Sprache, deutsche Kultur und finnisch-deutsche Beziehungen. Festschrift für Ahti Jäntti zum 65. Geburtstag* (= Finnische Beiträge zur Germanistik 19). Frankfurt am Main: Peter Lang Verlag. 173-187.

[1] Frei zugänglich unter: http://www1.uni-hamburg.de/exmaralda/
[2] Das Korpus besteht aus 138 Tonaufnahmen und Transkripten (ca. 150 000 Tokens); die Erhebungen wurden zwischen 1988-1990 durchgeführt. Siehe Institut für Deutsche Sprache Mannheim: Archiv für Gesprochenes Deutsch: http://www.ids-mannheim.de/ksgd/agd/

Aufbau eines Lernerkorpus mit mündlichen und schriftlichen Beiträgen aus realen Sprachproduktionssituationen

Antonie Hornung
Università degli Studi di Modena e Reggio Emilia
hornung.antonie@unimore.it

Seit 1998 arbeiten wir an der Universität Modena mit dem Modell einer „immersiven Sprachdidaktik", d.h. die Unterrichtssprache, aber auch die Sprache in semiformellen und formellen Kommunikationssituationen (Sprechstunde, Mails, Begegnung außerhalb der Universität) ist Deutsch. Es heißt aber auch, dass im Unterricht selbst Methoden der deutschsprachigen Universitäten praktiziert werden, die sich z.T. erheblich von den Vermittlungsmethoden der italienischen Universität unterscheiden. Es werden also beispielsweise Protokolle und kleinere schriftliche Hausarbeiten verfasst, die Studierenden halten mündliche Referate, usw. Auch die Abschlussarbeit, sowohl am Ende des Bachelorstudiums als auch für den Master, wird von vielen Studierenden auf Deutsch verfasst. Diese schriftlichen Arbeiten geben Einblick in die unterschiedlichsten Lernsituationen und Sprachentwicklungsphasen. Es treten allen gemeinsame Lernprobleme auf, aber es lassen sich auch höchst individuelle Wege und Lernstrategien feststellen.
Ähnlich individuell ist der jeweilige Stand im mündlichen Deutschen. Dies lässt sich aufgrund der im Rahmen meines nationalen Forschungsprojekts zum Thema „Wie sich jugendliche Europäerinnen ihre mehrsprachige Identität aufbauen" durchgeführten narrativen Interviews deutlich erkennen. Die Interviews wurden mit EXMARaLDA nach den HIAT-Richtlinien transkribiert; das Material steht auch als Audiomaterial zur Verfügung.
Ich werde in meinem Beitrag einige Beispiele aus meinem Materialfundus bezüglich ihrer Sprachqualität zur Diskussion stellen, möchte aber auch über das Konzept für den Aufbau eines Lernerkorpus mit derart unterschiedlichen Materialien sprechen und damit die Frage aufwerfen, wie diese Art Lernerkorpus für die Didaktik fruchtbar gemacht werden könnte.

Lernersprache: syntaktische Annotation des Lernerkorpus Falko – das Problem der Zielhypothese

Anke Lüdeling
Humboldt-Universität zu Berlin
Anke.luedeling@rz.hu-berlin.de

In diesen Vortrag geht es um die Rolle der Interpretation von Äußerungen, die von einem wie auch immer gewählten sprachlichen Standard abweichen. Anhand der syntaktischen Annotation von Lerneräußerungen (als ein Beispiel für vom Standard abweichende Äußerungen) soll gezeigt werden, *wie* und *wie sehr* die Interpretation die Ergebnisse beeinflusst.
Sprachliche Äußerungen von Lernern einer Fremdsprache weichen oft so stark von der zu lernenden Zielsprache ab, dass sie nicht einfach mit den für die Zielsprache entwickelten Beschreibungsmitteln analysiert werden können.
Beschreibungen von Lernersprache sind daher oft mehr oder weniger detaillierte Beschreibungen von ‚Fehlern' oder ‚Abweichungen', entweder rein qualitativ oder (oft bezogen auf Fehlerannotationen in Lernerkorpora) qualitativ und quantitativ. Die Beschreibungskategorien (Fehlertags) unterscheiden sich dadurch von den syntaktischen Beschreibungskategorien, die für ‚kanonische' Äußerungen verwendet werden.

Der Fehlerbegriff ist hochproblematisch (das zeigt eine bereits lang andauernde Diskussion, siehe z.B. Corder 1981 oder Dagneaux et al. 1998): Es ist nicht einfach festzulegen, was ein Fehler ist und wie Fehler klassifiziert werden können. Es ist aber klar, dass jeder Fehler- oder Abweichungsbegriff implizit oder explizit eine korrekte Struktur oder Zielhypothese annehmen muss. Für viele abweichende Sätze kann es unterschiedliche Zielhypothesen geben. Weil Abweichungen notwendigerweise bezogen auf eine Zielhypothese interpretiert werden, hat die Setzung der Zielhypothese große Auswirkungen auf die qualitativen und quantitativen Ergebnisse einer Analyse (Lüdeling, erscheint).
Hirschmann et al. (2007) zeigen, wie die Zielhypothese eingesetzt werden kann, um Lernersprache mit den gleichen syntaktischen Kategorien zu beschreiben wie kanonische Sprache. In meinem Beitrag analysiere ich Texte fortgeschrittener Lerner des Deutschen als Fremdsprache aus dem Lernerkorpus Falko[1] syntaktisch und zeige, wie unterschiedliche Zielhypothesen und unterschiedliche syntaktische Modelle miteinander interagieren.

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

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

Hirschmann, Hagen; Seanna Doolittle & Anke Lüdeling (2007): Syntactic annotation of non-canonical linguistic structures. In: *Proceedings of Corpus Linguistics 2007*, Birmingham. Online verfügbar unter http://www2.hu-berlin.de/korpling/mitarbeiter/anke/HirschmannDoolittleLuedelingCL2007.pdf

Lüdeling, Anke (erscheint): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Patrick Grommes & Maik Walter (Hrsg.) *Fortgeschrittene Lernervarietäten*. Tübingen: Niemeyer.

[1] http://www2.hu-berlin.de/korpling/projekte/falko/