van Kemenade, Ans. 1987. Syntactic Case and Morphological Case in the History of English. Foris, Dordrecht.

Kroch, Anthony and Ann Taylor. 1997. Verb movement in Old and Middle English: Dialect variation and language contact. In Ans van Kemenade and Nigel Vincent, editors, Parameters of Morphosyntactic Change, pp. 297–325. Cambridge University Press, Cambridge.

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM, first edition, URL: http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/.

Kroch, Anthony and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, CD-ROM, second edition (first edition 1994). URL http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/.

Kroch, Anthony, Ann Taylor, and Donald Ringe. 2000. The Middle English verb-second constraint: a case study in language contact and language change. In Susan C. Herring, Pieter van Reenen, and Lene Schoesler, editors, Textual Parameters in Older Languages, pp. 353–391. John Benjamins, Amsterdam/Philadelphia.

Pintzuk, Susan. 1991. Phrase structures in competition: Variation and change in Old English word order. PhD thesis, University of Pennsylvania.

Pintzuk, Susan. 1993. Verb-seconding in Old English: Verb movement to Infl. The Linguistic Review, 10:5–35.

Pintzuk, Susan. 1999. Phrase Structures in Competition. Garland, New York/London.

Ries, John. 1907. Die Wortstellung im Beowulf. Verlag von Max Niemeyer, Halle a.S.

Speyer, Augustin. 2004. Topicalization and the trochaic requirement. Penn Working Papers in Linguistics, 10.2, pp. 243–256.

Speyer, Augustin. 2005. A phonological factor for the decline in topicalization in english. In Stephan Kepser and Marga Reis, editors, Linguistic Evidence, Studies in Generative Grammar 85, pp. 485–506. Mouton de Gruyter.

Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen, 2006. Parsed Corpus of Early English Correspondence. Oxford Text Archive. URL http://www-users.york.ac.uk/~lang22/PCEEC-manual/corpus_description/index.htm.

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. York-Toronto-Helsinki Parsed Corpus of Old English Prose. Oxford Text Archive, first edition. URL http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Trips, Carola. 1999. Scandinavian characteristics in the Ormulum - evidence for Scandinavian influence on word order change in Early Middle English. Ms., University of Stuttgart. Paper presented at ConSOLE 8, Vienna.

Trips, Carola. 2002. From OV to VO in Early Middle English. John Benjamins, Amsterdam/Philadelphia.

Karin Donhauser

Zur informationsstrukturellen Annotation sprachhistorischer Texte

Abstract English

The article demonstrates a procedure for annotating information structure in texts from historical corpora which was developed by Collaborate Research Group SFB 632 Information Structure (SFB) at Humboldt University, Berlin. Here, the annotation is designed in a way allowing to retrace each decision step made throughout the annotation process. Therefore, annotation is conducted incrementally distinguishing the following three layers of Information Structure: (1) cognitive status, (2) predication structure and (3) informational relevance. This cumulative approach has delivered optimal results in practice, among all in working out an information-structural cartography of the left and right sentence periphery in Old High German. This is exemplified here with respect to focus positions in subordinate clauses.

Abstract Deutsch

Der Artikel stellt ein Verfahren zur informationsstrukturellen Annotation von sprachhistorischen Texten vor, das im Rahmen des SFB 632 Informationsstruktur an der Humboldt-Universität zu Berlin erarbeitet wurde. Dabei werden Annotationen so angelegt, dass Entscheidungswege immer nachvollziehbar bleiben. Die Annotation erfolgt deshalb sequentiell unter strikter Trennung folgender informationsstruktureller Ebenen: (1) kognitiver Status, (2) prädikative Strukturierung und (3) informationsstrukturelle Gewichtung. Dieses kumulative Vorgehen hat sich im Projekt hervorragend bewährt, u.a. bei der informationsstrukturellen Kartographie des linken und rechten Satzrandes im Althochdeutschen, deren Ergebnisse hier an einem Beispiel (Fokuspositionen im Nebensatz) illustriert werden.

1. Einführung

Der korpuslinguistische Ansatz, den wir im Rahmen der DDD-Initiative in den Jahren 2003-2005 für den Aufbau eines historischen Referenzkorpus für das Deutsche entwickelt haben (zur Beschreibung dieses Ansatzes vgl. Lüdeling/Poschenrieder/Faulstich 2005), sieht eine Mehrebenenarchitektur vor, die kumulatives Arbeiten unterstützt und es deshalb gut möglich macht, auch speziellere Forschungsinteressen zu verfolgen.

Wir haben uns diesen Ansatz auch in dem 2004 gegründeten Sonderforschungsbereich 632 Informationsstruktur (Potsdam/ HU Berlin) zueigen gemacht. Dabei wurde in

Sprache und Datenverarbeitung 1-2/2007: S. 39–45

Projekt B4 (Die Rolle der Informationsstruktur bei der Entwicklung der Wortstellungsregularitäten in den germanischen Sprachen) unter der Leitung von Roland Hinterhölzl und mir ein Verfahren zur informationsstrukturellen Annotation von historischen Texten entwickelt, das ich hier im Folgenden vorstellen möchte.

2. Problemstellung und Lösungsansatz

Die informationsstrukturelle Bewertung von historischen Texten ist aus verschiedenen Gründen alles andere als trivial. Problematisch sind insbesondere die folgenden zwei Defizite, die die Zuweisung informationsstruktureller Werte systematisch erschweren:

- Das Fehlen prosodischer Informationen: Im Deutschen wie auch in den verwandten germanischen Sprachen werden informationsstrukturelle Kategorien wie Fokus formal vorrangig mit prosodischen Mitteln (Akzent) ausgewiesen. Da unser Schriftsystem keine systematische Kennzeichnung prosodischer Eigenschaften vorsieht, stehen diese Merkmale für die Bestimmung von Informationsstruktur in historischen Texten nicht zur Verfügung (Für einen Überblick über potentielle Quellen zur Ermittlung prosodischer Information in den ältesten althochdeutschen Sprachzeugnissen siehe Fleischer (eingereicht)).
- Das Fehlen muttersprachlicher Intuitionen: Da wir bei Analyse von historischen Sprachen bzw. Sprachstufen nicht auf die Intuitionen muttersprachlicher Sprecher zurückgreifen können, müssen die informationsstrukturellen Funktionen ausschließlich aus dem Kontext erschlossen werden. Dies hat zur Folge, dass nicht alle Zuweisungen informationsstruktureller Werte in der gleichen Weise gesichert sind. Bei inhaltlich schwierigen Textstellen ist die informationsstrukturelle Bewertung manchmal sogar unmöglich.

Eine weitere Schwierigkeit, die mit der speziellen Zielsetzung unseres Projektes verbunden ist, ergibt sich aus der Überlieferungslage des Althochdeutschen.

- Die Mehrheit der Texte, die uns aus der Frühzeit des Deutschen überliefert sind, sind mehr oder weniger wortgetreue Übersetzungen aus dem Lateinischen bzw. haben zumindest lateinische Vorlagen. Die Zahl der autochthonen althochdeutschen Texte ist gering. Der einzige größere autochthone Text, Otfrids Evangelienbuch ist (end)reimgebunden und deshalb für Untersuchungen zur Wortstellung nur bedingt geeignet (dazu Fleischer 2006).

Der Lösungsansatz, den wir für diese Probleme entwickelt haben, ist deshalb komplex. Er beruht auf drei Grundentscheidungen:

1. Wir passen unser Vorgehen der Überlieferungslage an und modellieren die althochdeutschen Verhältnisse in mehreren Schritten, ausgehend von den Belegen, die direkt gegen das Lateinische stehen und damit eine hohe Aussagekraft haben. Wir starten deshalb unsere Auswertung mit dem althochdeutschen Tatian, also mit dem althochdeutschen Text, der die engste Bindung an das Lateinische aufweist, und konzentrieren uns dabei zunächst nur auf die Differenzbelege, aus denen wir ein erstes – vorläufiges – Modell der althochdeutschen Verhältnisse generieren (Fleischer/Hinterhölzl/Solf eingereicht). Dieses Modell wird dann in einem zweiten Schritt durch Abgleich der Korrespondenzbelege für den Tatian validiert und schließlich an Belegen aus anderen Textüberlieferungen einer abschließenden Evaluation unterzogen, die wir den autochthonen Textüberlieferungen entnommen haben.
2. Wir verzichten auf die vollständige Analyse von Texten und beschränken uns auf die Sätze bzw. Textabschnitte, die so kontextualisiert sind, dass sie sich mit hinreichender Sicherheit auch informationsstrukturell interpretieren lassen. Das althochdeutsche Korpus, das wir auf diese Weise zusammengestellt haben, umfasst derzeit fast 2000 Sätze. Diese stammen überwiegend aus narrativen Textpassagen, die sich insgesamt deutlich besser interpretieren lassen, als dies bei argumentativen Sequenzen der Fall ist.
3. Wir haben die informationsstrukturelle Annotation gezielt so angelegt, dass nicht nur die zugewiesenen Werte, sondern auch die Entscheidungswege, die zu der Vergabe eines informationsstrukturellen Wertes führen, im Rahmen der Annotation verzeichnet werden. Die Annotationsentscheidungen sind damit maximal transparent und können im Modellierungsprozess jederzeit überprüft und gegebenenfalls auch revidiert werden. Den kumulativen Annotationsansatz, den wir zu diesem Zweck entwickelt haben, möchte ich im Folgenden nun etwas genauer erläutern (eine ausführlichere Beschreibung des Annotationsansatzes findet sich in Petrova und Solf (eingereicht)).

3. Das Annotationsverfahren

Ebenso wie dies für das von uns geplante historische Referenzkorpus des Deutschen (DDD) vorgesehen ist, erhält auch in unserem Korpus jeder Satz eine morphologische Annotation, die sich hier auf die Angabe der Wortartenkategorie beschränkt. Dazu kommen – jeweils auf getrennten Ebenen – ausgewählte phonologische und syntaktische Informationen (Silbenzahl, Satztyp, Form und Funktion von Phrasen), so dass wir im Gesamten über ein nahezu vollständiges grammatisches Profil der einzelnen Sätze verfügen.

Darauf aufbauend erfolgt nun die informationsstrukturelle Annotation des jeweiligen Satzes. In Abstimmung mit den anderen Projekten des SFBs berücksichtigen wir dabei drei Ebenen der informationsstrukturellen Gliederung, denen auch jeweils eigene Annotationsebenen entsprechen: i) kognitiver Status (gegeben vs. neu), ii) prädikationelle Gliederung (Topik vs. Kommentar) und iii) informationelle Gewichtung (Fokus vs. Hintergrund), dazu Krifka (erscheint). Die Annotation erfolgt in folgenden Schritten:

In einer ersten Stufe bestimmen wir den kognitiven Status der einzelnen Phrasen. D.h. wir ermitteln, ob es sich bei der in den Phrasen kodierten Information um neue Information (‚new') oder um bereits bekannte Information handelt. Im letzteren Fall

unterscheiden wir zwei Konstellationen: Information, die im vorangehenden Kontext explizit vorerwähnt ist (‚given') und solche, die zwar nicht unmittelbar vorgegeben ist, aber dennoch kontextuell erschließbar bzw. allgemein zugänglich ist (‚accessible').

In einem zweiten Interpretationsdurchlauf bestimmen wir die Topik-Kommentar-Gliederung im jeweiligen Satz (sofern Sätze eine solche aufweisen). Dabei muss man sich zwischen den jeweiligen Ausprägungen des Topikbegriffs in der modernen informationsstrukturellen Literatur entscheiden. Um dabei möglichst theorieneutral zu bleiben, weisen wir den jeweiligen Topik-Kandidaten nicht unmittelbar den Topikstatus zu, sondern nehmen konstitutive Eigenschaften auf, die in der einschlägigen Literatur zu Topikalität diskutiert werden, darunter neben dem kognitiven Status (Ebene 1) die positionelle Realisierung, Definitheit und Referentialität.

In einer dritten Stufe analysieren wir die informationsstrukturelle Gewichtung innerhalb des Satzes, was mit der Realisierung der Fokus-Hintergrund-Gliederung zusammenhängt. Als Fokus gelten dann jene Teile des Satzes, die für seine Rolle im jeweiligen Kontext als besonders wichtig oder relevant ausgewiesen werden können. Weniger wichtige Information, z.B. bekannte oder erschließbare Information, bildet dabei den Hintergrund der Äußerung (background = ‚bgr'), neue oder kontrastiv hervorgehobene Information macht den Fokus aus, der entweder als Neuinformationsfokus (‚nif') oder Kontrastfokus (‚cf') gekennzeichnet wird.

Dazu ein Beispiel:

Lat. ***F**uit in diebus herodis regis/ iudeę quidam sacerdos/ nominee zacharias /de uice abia.'/ & uxor illi de filiabus aaron/& nomen eius elisab&h/ **erant autem iusti ambo ante deum***

Ahd. uuar [sic!] In tagun herodes thes cuninges/ Iudeno sumer biscof/ namen zacharias/ fon themo uuehsale abiases/ Inti quena Imo fon aarones tohterun/ Inti ira namo uuas elisab&h/ ***siu uuarun rehtiu beidu fora gote*** (T 25, 29–26, 3)

Lat.		erant	autem	iusti	ambo	ante	deum
OHG	siu	uuarun		rehtiu	beidu	fora	gote
cognitive status	given						acc
topic-comment structure	top	comment					
referentiality	ref						
definiteness	def						
position	init						
focus-background structure	bgr	focus					
new-information focus		nif					
contrastive focus							

Erläuterung: Im ahd. Satz wird das Subjektspronomen *siu*-3pl.n. gegen das lat. Original eingefügt und an die satzinitiale Position, vor die finite Kopula, gesetzt. Die hinzugefügte

Konstituente nimmt eine bereits eingeführte Entität wieder auf, nämlich Zacharias und dessen Frau Elisabeth, was die Vergabe des Merkmals ‚given' motiviert. Auf das Subjektspronomen treffen neben der kontextuellen Vorerwähntheit alle übrigen Merkmale von Topiks zu: Der Ausdruck ist definit, hat referentielle Eigenschaften und ist an einer frühen Position im Satz realisiert. Darüber hinaus kann das Subjektspronomen an die Stelle von X im Topik-Test *A says about X that X...* eingesetzt werden, womit sich eine adäquate Interpretation dieser Äußerung im jeweiligen Kontext ergibt. Deshalb wird auf das Vorliegen einer Topik-Kommentar-Gliederung geschlossen bzw. der Ausdruck *siu* als Aboutness-Topik der Aussage identifiziert. Der Fokus der Aussage entspricht der VP, die die neue Information über Zacharias und Elisabeth übermittelt. Dieser Teil des Satzes ist auch unter Anwendung eines angepassten Questio-Ansatzes als kongruente Antwort auf die implizite Frage „Was ist mit Zacharias und Elisabeth?" denkbar, d.h. er entspricht dem relevantesten Ausschnitt aus dem Satz, der die Fortführung der Erzählung sicherstellt.

Die Eingabe dieser Annotationen erfolgt manuell mit Hilfe des Partitureditors EXMARaLDA. In der Folge werden diese Daten in PAULA (=Potsdamer Austauschformat für Linguistische Annotationen), ein im SFB entwickeltes Format, konvertiert und in die Datenbank des SFBs ANNIS (=a linguistic database for Annotated Information Structure) eingespeist. Suche und Auswertung der Daten in ANNIS erfolgt via Netz.

4. Die Ergebnisse: ein Beispiel

Auf der Basis eines so annotierten Korpus haben wir in der ersten Bewilligungsphase des SFBs u.a. eine umfassende informationsstrukturelle Kartographie des linken und rechten Satzrandes im Althochdeutschen vorgenommen, deren Ergebnisse ich hier an einem Beispiel veranschaulichen möchte. Dabei geht es um die positionelle Realisierung von fokussierten Phrasen in Nebensätzen des Althochdeutschen. Unser Basiskorpus von Differenzbelegen liefert dazu die folgenden Auskünfte:

1. Phrasen, die bekannte Information enthalten und dem Äußerungshintergrund angehören, werden im Althochdeutschen häufig gegen das Lateinische oder auch unabhängig vom Lateinischen in die Position zwischen der nebensatzeinleitenden Konjunktion und dem finiten Verb gesetzt. Solche Verschiebungen oder Einfügungen sind in unserem Korpus 228 Mal belegt, und zwar für Pronomina ebenso wie für volle Nominalphrasen.

(1) ahd. thanne ***Ir*** ***Iz*** find& (T 40,4)
wenn ihr es findet
lat. cum Inuerneritis

(2) ahd. so hér ***thén buoh*** int&a (T 53,21)
als er dieses Buch aufmachte
lat. & ut reuoluit **librum**

2. Phrasen, die neue Information enthalten und im Fokusbereich der Äußerung stehen (Neuinformationsfokus), werden gegen das Lateinische oder auch unabhängig vom Lateinischen in der Position nach dem finiten Verb realisiert. Dabei wird wie in (3) die Verbendstellung des Lateinischen aufgelöst. In anderen Fällen wie in (4) hat die althochdeutsche Konstruktion im Lateinischen keine direkte Entsprechung.

(3) ahd. Inti thie thár hab&un ***diuual*** (T 59,1)
und die hatten [den] Teufel
lat. & qui **demonia** habebant

(4) ahd. soso thie lihhazara **sint** gitruobte (T 68, 23)
so wie die Heuchler sind betrübt
lat. sicut hypocrite tristes

Veränderungen dieser Art sind in unserem Korpus insgesamt nur zwölf Mal belegt. Allerdings werden Fokusphrasen diesen Typs häufig bereits im Lateinischen in postverbaler Position realisiert. Die Serialisierung des Lateinischen bleibt dann im Althochdeutschen unverändert.

(5) ahd. thane thú tuos ***elimosinam*** (T 66,27)
wenn du gibst Almosen
lat. Cum ergo facies **elimosinam**

3. Phrasen, die kontrastive Information enthalten und im Fokusbereich der Äußerung stehen (Kontrastfokus), werden im Althochdeutschen gegen das Lateinische in die Position unmittelbar vor dem finiten Verb verlagert. Solche Verschiebungen sind in unserem Korpus 36 Mal belegt.

(6) ahd. niuuizze íz thin uuinistra/ uuaz ***thin zesuua*** tuo (T 67,4-5)
nicht wisse deine Linke was deine Rechte tue
lat. nesciat sinistra tua/ quid faciat **dextra** **tua**

In vielen Fällen, in denen die Verschiebung unterbleibt, liegt zwischen dem Verb und der fokussierten Nominalphrase der Zeilenumbruch, der bei der interlinear aufgebauten Übersetzung des Tatian meist erhalten bleibt.

Auf der Basis dieser Beobachtungen lassen sich die informationsstrukturellen Verhältnisse im Nebensatz des Althochdeutschen mit Hinterhölzl 2004 in folgender Weise modellieren:

$$[C \quad \text{Backgr} \quad [_{\text{FocP}}\text{ContrF V} \; [_{\text{AgrP}}\,\text{PresF} \quad [_{\text{VP}}\, t_V]]]]$$

Das finite Verb rückt in eine satzmediale Position vor, deren Spezifikator kontrastiv fokussierten Phrasen vorbehalten ist. Konstituenten, die neue Information enthalten, bleiben im Skopus dieser Fokusphrase im postverbalen Bereich. Konstituenten, die bekannte Information enthalten und zum Äußerungshintergrund gehören, werden aus dem Skopusbereich des Fokus bewegt. Dieses Modell wurde im Folgenden an breiteren Datensätzen des Althochdeutschen evaluiert. Es konnte dabei weiter validiert werden.

Literatur

Fleischer, J. (2006): Zur Methodologie althochdeutscher Syntaxforschung. Beiträge zur Geschichte der deutschen Sprache und Literatur 128, 25-69.

Fleischer, J. (eingereicht): Paleographic clues to Old High German prosody? – Accents, word separation and related phenomena in Old High German manuscripts. In: New Approaches to Word Order Variation and Change in the Germanic Languages, eds. R. Hinterhölzl and S. Petrova. Berlin: Mouton de Gruyter.

Fleischer, J., Hinterhölzl, R. und Solf, M. (eingereicht): Zum Quellenwert des AHD-Tatian für die Syntaxforschung: Überlegungen auf der Basis von Wortstellungsphänomenen. Zeitschrift für germanistische Linguistik.

Hinterhölzl, R. (2004): Language Change versus Grammar Change: What diachronic data reveal about the distinction between core grammar and periphery. In: Diachronic Clues to Synchronic Grammar, eds. E. Fuß and C. Trips. Amsterdam and Philadelphia: John Benjamins, 131–160.

Krifka, M. (erscheint): Basic notions of Information Structure. In The notions of Information Structure, eds. Caroline Féry et al., Potsdam: Universitätsverlag, 13–56.

Lüdeling, A., Poschenrieder, Th. und Faulstich, L. C. (2005): DeutschDiachronDigital – Ein diachrones Korpus des Deutschen. Jahrbuch für Computerphilologie 2004, 119-136.

Petrova, S. and Solf, M. (eingereicht): On the Methods of Information-Structural Analysis of Texts from Historical Corpora. A Case Study on the OHG Tatian. In: New Approaches to Word Order Variation and Change in Germanic Languages, eds. R. Hinterhölzl and S. Petrova. Berlin: Mouton de Gruyter.

Sprache und Datenverarbeitung

International Journal for Language Data Processing

31. Jahrgang 2007 Heft 1-2

Begründet durch
Winfried Lenders und Harald Zimmermann

Herausgegeben durch das
Institut für Kommunikationswissenschaften
der Universität Bonn
Abteilung Sprache und Kommunikation
Poppelsdorfer Allee 47
53115 Bonn

von:
Hermann Cölfen, Essen
Annelie Rothkegel, Chemnitz
Ulrich Schmitz, Essen
Bernhard Schröder, Essen

Schriftleitung:
Ulrich Schmitz
Universität Duisburg-Essen
Fachbereich Geisteswissenschaften
Universitätsstraße 12
45117 Essen
E-Mail: ulrich.schmitz@uni-due.de

Layout:
Sabine Walther, Duisburg

Titelillustration:
Michael Hüter, Bochum

Sprache und Datenverarbeitung im Internet: http://www.linse.uni-due.de