120. Computeranalyse/Computer Analysis

1. Korpusbasierte Sprachanalyse

Computerprogramme sind für bestimmte Datenformate bestimmt und nur diese Datenformate können von ihnen verarbeitet werden. *Sprachdaten* als allgemeine Bestimmung soziolinguistischer Rohdaten ist dabei zwar korrekt, aber nicht ausreichend präzise. In verschiedenen Phasen der linguistischen Verarbeitung der Rohdaten werden sie in verschiedene Formate umgewandelt und enthalten dadurch verschiedene Informationen. Zu den soziolinguistischen Daten gehören *per definitionem* auch die Metadaten, d.h. Daten über die Sprecher, ihre Sprachgemeinschaften, die Situationen, in denen die Daten gesammelt wurden. Auch diese Daten erfahren verschiedene Phasen der Verarbeitung und erfüllen verschiedene Funktionen als primäre oder sekundäre Daten. Detaillierte Informationen zu hier verwendeten methodologischen Begriffen wie *primäre und sekundäre Daten*, *Untersuchungsziel*, *Datenbereich*, *Datenkontext* etc. findet der Leser in Skiba (1998, 201ff.). Ihre Verwendung hier ist meistens kontextuell selbsterklärend. Daten können (je nach Untersuchungsziel) als Zusatzinformationen zu den sprachlichen Daten die Analyse unterstützen oder selbst als eigentliche *Daten* dienen.

In diesem Artikel werden nur computergestützte Verfahren der Datenanalyse dargestellt, in denen *primäre sprachliche Daten* als *Datenbereich* dienen. Andere Untersuchungen, die z.B. soziolinguistische Fragebögen zur Sprachbenutzung oder zur Einstellung von Sprechern auswerten, werden hier nicht berücksichtigt. Computergestützte Analyseverfahren für die Auswertung von Fragebögen o.ä. entsprechen den in der statistisch ausgerichteten Soziologie angewandten Methoden (vgl. dazu Art. 115). Auch Methoden der Computerlinguistik (z.B. *parsing*, *Lemmatisierung*) werden in der vorliegenden Darstellung nicht behandelt (vgl. dazu Art. 111).

In den folgenden Abschnitten werden die Phasen der (sozio-) linguistischen Datenverarbeitung dargestellt.

1) die Datensammlung (Rohdaten)
2) die Datenstrukturierung (Untersuchungsdesign und Metabeschreibung)
3) die Annotation (Transkription, Kodierung etc.)
4) die Auswertung
5) der Korpusaufbau

Besonders soziolinguistische Untersuchungen vergrößern ihre Aussagekraft deutlich, wenn sie auf großen Datenmengen basieren. Die korpusbasierte Sprachanalyse, die besonders viele computergestützte Verfahren benutzt, ist mit der neuesten technischen Entwicklung ein zukunftsgerichtetes Fach. Ihre neuere Entwicklung, die auf den Traditionen der Korpuslinguistik und der empirischen Sprachforschung aufbaut, bietet der Forschung viele Vorteile: Sie definiert technische und konzeptuelle Standards für Datenaufbewahrung, Datenstrukturierung, Datenkodierung und Auswertung. Sie bewahrt die Daten und die Analyseergebnisse vor dem physikalischen Verfall und stellt sie geordnet und (meta-) beschrieben anderen Forschergruppen und -generationen zur Verfügung.

2. Datensammlung

Die Qualität und die Beschaffenheit der Rohdaten sind entscheidend für die späteren Auswertungsmöglichkeiten.

Audio- und Videoaufnahmen von sprachlichen Ereignissen bilden die primären Daten, und somit die Grundlage soziolinguistischer Untersuchungen. Wird eine computergestützte Weiterverarbeitung angestrebt, so sind digitale (aber nicht komprimierte) Datenformate zu empfehlen. Komprimierte Datenformate haben zwar den Vorteil des geringen Speichervolumens, sind aber für phonetische Analysen nicht geeignet. Minidisk-Formate sind ein Beispiel für komprimierte Datenformate. Auch einige DV-Videogeräte zeichnen den Ton in einer komprimierten Form auf. So ist darauf zu achten, dass abhängig von den Analysezielen das richtige Format gewählt wird (vgl. dazu die gute, auch für Laien zugängliche Darstellung in Dittmar 2002).

Analyseziele haben außerdem Einfluss auf die Menge und die zeitliche Anordnung der Aufnahmen (Anzahl der Sprecher, Länge der einzelnen Aufnahmen und ihre Häufigkeit). Eine zumindest grobe Beschreibung des Analysedesigns ist deshalb bereits vor Beginn der

Datensammlung notwendig (vgl. weiter unten, 2.1).

Die Datensammlung beinhaltet außerdem schriftliche oder auf dem Band aufgezeichnete Protokolle, d.h. Daten über die Informanten, die Aufnahmesituation, etc. Diese Informationen sind für die Analysen und den Korpusaufbau enorm wichtig (vgl. weiter unten 2.2).

2.1. Das Datendesign und die Datenstrukturierung

Das Datendesign, d.h. die sinnvolle Anordnung, Beschaffenheit und Menge von Daten in Abhängigkeit von den Untersuchungszielen erhöht die Überzeugungskraft wissenschaftlicher Argumentationen (vgl. dazu Skiba 1998, 131ff.). Querschittdaten scheinen für andere Ziele geeignet als Längsschnittdaten (z.B. Untersuchung der Sprachvarietät vs. Untersuchung des Spracherwerbs).

Eine gute Planung ist dabei ausschlaggebend. Dazu gehören u.a. die Festlegung der Aufnahmemengen und Aufnahmezeitpunkte (für diesen Beitrag nicht relevant) und die Sammlung von relevanten Informationen zu den Rohdaten. Auch alle wichtigen Zusatzinformationen zu Aufnahmen sind zu sammeln. Die Protokolldaten bilden die Grundlage für eine weitere, strukturierte Beschreibung der primären Daten. Sie dienen in einer Untersuchung als Datenkontext, z.B. in Form von Bemerkungen zum Geschehen bzw. als Grundlage für Meta-Beschreibungen.

2.2. Festlegung der Analyseeinheiten und ihre Meta-Beschreibung

Aus der linguistischen Praxis stammt die Idee, strukturierte Daten, d.h. meistens relativ kleine, abgeschlossene Einheiten zu analysieren. Diese Vorgehensweise schafft mehr Überblick und erhöht die Vergleichbarkeit und Argumentationskraft in wissenschaftlichen Argumentationen.

Die technologische Entwicklung (Internet, digitale Aufnahmemöglichkeiten, Digitalisierungsverfahren für Medien) und die Notwendigkeit, Daten im Internet auffindbar und zugänglich zu machen, haben die Idee gefördert, Standards für die Struktur von linguistischen Daten zu etablieren.

Eine internationale Initiative, in der Linguisten und Sprachingenieure zusammenarbeiten, und die sich diese Aufgabe zum Ziel gesetzt hat ist ISLE (**I**nternational **S**tandard for **L**anguage **E**ngeneering). Das Max-Planck-Institut für Psycholinguistik in Nijmegen in den Niederlanden (*http://www.mpi.nl*) beherbergt den Metadatensatz und sorgt für seine Publizierung im Internet sowie für die Stabilität der angegebenen Webadressen (*http://www.mpi.nl/ISLE/*). Ein sogenanntes IMDI-set (**I**sle **M**eta **D**ata **I**nitiative) entwickelte sich zum Standard für die Meta-Beschreibung von sprachlichen Primärdaten. OLAC (**O**pen **L**anguage **A**rchives **C**ommunity, *http://www.language-archives.org/*) – eine andere internationale Initiative – verfolgt die gleichen Ziele. Die dort entwickelte Metadatenstruktur unterscheidet sich teilweise von IMDI. Beide sind ineinander konvertierbar (vgl. *http://www.mpi.nl/ISLE/*).

Im IMDI – Standard werden linguistische Analyseeinheiten *sessions* genannt (Broeder et al. 2001, 48ff.). Eine *session* ist eine abgeschlossene Dateneinheit, z.B. ein Interview, eine Nacherzählung oder ein thematisch abgeschlossener Teil von einem Gespräch. Um diese Dateneinheit zu einer computerkompatiblen Analyseeinheit zu machen, muß aus den Rohdaten ein der *session* entsprechendes digitalisiertes Stück herausgeschnitten werden. Diese Mediendatei kann mit dem Computer weiter verarbeitet werden.

Einen wichtigen Teil des Datenverarbeitungsprozesses bildet die Meta-Beschreibung von *sessions.* Sie ist notwendig, um Informationen über die Daten mit den Medien zu verbinden.

Die Meta-Beschreibung enthält (xml-kodierte) Daten über:

- die Sprache(n) und den Aufnahmeort der *session*.
- den Inhalt der *session* (unter anderem Thema, Genre, Datenart).
- die Beteiligten (z.B. Sprecher, Projekt).
- die Originalaufnahmen (Name und Aufbewahrungsort der Rohdaten) und die Medieneinheiten (Ton-, Bild-, Video- und Textdateien), die mit der *session* assoziiert werden.

Alle Metadaten sind fein gegliedert (vgl. *http://www.mpi.nl/ISLE/schemas/schemas_frame.html*) und können jederzeit mit neuen Informationen (z.B. Publikationsangaben, Annotationsdateien etc.) ergänzt werden. Es existiert auch ein Computerprogramm (frei erhältlich unter *http://www.mpi.nl/tools*), mit dem die Daten eingegeben, validiert und gespeichert werden können. Zu sich wiederholenden Metadaten (z.B. zu Informanten, die

in verschiedenen *sessions* auftreten oder zur Projektbeschreibung) können Daten in assoziierten Dateien separat gespeichert, und damit wieder verwendbar gemacht werden.

Eine Meta-Beschreibung fasst also alle soziolinguistisch und technisch relevanten Informationen, die zu einer *session* gehören, zusammen und kann deshalb zum Aufbau eines gut beschriebenen Sprachkorpus verwendet werden. Die Korpusdaten sind auf diese Art leicht auffindbar (vgl. mehr dazu unter 5.).

Traditionelle Verfahren der Speicherung von Metadaten (z. B. als separat kodierte Informationen in Transkriptionsköpfen) haben den Nachteil, dass dadurch primäre Daten zusammen mit Metadaten gespeichert werden, und dadurch nur umständlich auffindbar sind. Fachlich spezialisierte Metadaten (vgl. z. B. das auf den Erstspracherwerb zugeschnittene Metadatenset von McWhiney 1995, 12ff.) sind für andere Forschungsrichtungen nicht ohne Anpassungen anwendbar. IMDI Metadaten sind innerhalb der vorgegebenen Struktur durch die Definition eigener Schlüsselbegriffe ohne Nachteile erweiterbar und einfach in neue Strukturen übersetzbar.

2.3. Datendigitalisierung

Unabhängig davon in welchen Formaten die Daten aufgenommen wurden, sollen sie vor einer weiteren Verarbeitung digitalisiert und als Computerdateien gespeichert werden. Die Qualität der digitalen Daten ist von der Qualität der Aufnahmen abhängig. Nur digital aufgenommene Daten können ohne Verluste in Computerdateien umgewandelt werden. Die Digitalisierung von Medien ist inzwischen auch auf einem PC möglich. Es ist dabei darauf zu achten, dass bestimmte Standards für Datenformate eingehalten werden. Wünschenswert wären dabei internationale Standards für Sprachdaten. Vorläufig geben größere internationale Projekte dazu in der Praxis erarbeitete Empfehlungen (*best practice*) heraus (vgl. z. B. *http://www.mpi.nl/DOBES/*).

Besondere Vorsicht ist bei der Digitalisierung von Tondaten zu empfehlen. Alle Komprimierungen von Tondaten (vgl auch oben 2.), obwohl sie auf die Verständlichkeit (Hörbarkeit) der Daten keinen Einfluss haben, machen eine computergestützte phonetische Analyse unmöglich. Eine einmal in Abhängigkeit von den eigenen Untersuchungszielen getroffene Entscheidung kann die Daten gegebenenfalls für künftige Untersuchungen unbrauchbar machen.

3. Annotationen

Englische Termini, wie *annotation*, *transcription*, *tagging*, *glossing*, *coding*, *encoding*, *comment*, *main tier*, *dependent tier*, die in der Literatur zu diesem Thema gefunden werden können, werden nicht immer einheitlich verwendet (vgl dazu Bird/Liberman 1999, 1ff.). Die deutsche Terminologie ist ebenfalls uneindeutig.

Um die Funktionalität von Computerprogrammen im Bereich der Annotationen präzise beschreiben zu können, ist daher eine terminologische Klärung notwendig.

3.1. Terminologie

Eine digitalisierte Medieneinheit (Audio- oder Videodatei als Teil einer *session*) ist der Hauptbezugspunkt meiner Darstellung. Sie wird hier *Datenbereich* genannt (zur hier verwendeten methodologischen Terminologie vgl. Skiba 1998, 201ff.). Ein Datenbereich kann die gesamte Datei oder nur Teile davon umfassen. Von dieser Perspektive aus sind alle strukturierten Beschreibungen dazu als *Annotationen* zu dieser zeitlich begrenzten Einheit (bzw. zu ihren Teilen) zu betrachten.

Jede Art der symbolischen Beschreibung eines *per definitionem* zeitlich begrenzten Datenbereiches wird Bird/Liberman (1999, 1ff.) folgend eine *Annotation* genannt. Somit sind nicht nur linguistische Kodierungen und Kommentare mit allen ihren Subelementen, sondern auch Transkripte als Annotationen zu betrachten.

Ein *Transkript* ist eine schriftliche (symbolische) Wiedergabe des Gesprochenen. Dabei können verschiedene Aspekte des zugrundeliegenden Datenbereichs für die Wiedergabe ausgewählt werden, und die Beschreibung kann in verschiedene Ebenen aufgeteilt werden. Diese Ebenen der Beschreibung sind von der linguistischen Theorie und von den Untersuchungszielen abhängig. So kann entschieden werden, dass ein Transkript zwei konzeptuell unterschiedliche Ebenen, (z. B. eine phonetische und eine phonologische Wiedergabe der Datenbereiche) enthält, oder nur eine grobe, an den orthographischen Regeln der jeweiligen Sprache orientierte Ebene der Wiedergabe benutzt (ein sogenanntes literarisches Transkript).

Die Wiedergabeebenen werden in Transkripten als vertikal abgetrennte (durch besondere Symbole gekennzeichnete) Zeilen realisiert. Diese Ebenen (*tiers*) werden hier *Annotationszeilen* (*AZ*) genannt. Eine theo-

retische Begründung für eine *AZ* wird hier *Schicht* genannt. Eine *Schicht* ist eine Interpretation des Datenbereichs. Der Begriff umfasst jede Art der Interpretation von Datenbereichen (so kann es z. B. eine syntaktische Schicht geben, die einer bestimmten Syntaxtheorie verpflichtet ist oder eine Schicht, die die gleichen Datenbereiche mit dem Hintergrund einer anderen Theorie zu beschreiben versucht, vgl. dazu Skiba 1998, 25ff.).Solche *Schichten* werden in Transkripten und anderen Annotationen als verschiedene *AZ* (*tiers*) realisiert. Eine *AZ* ist also eine symbolische, physikalisch erkennbare Fixierung einer *Schicht*.

Schichten und *AZ* sind wie bereits erwähnt nicht auf *Transkripte* beschränkt. Sie sind als konzeptuelle bzw. physikalische Einheiten in allen Arten von Annotationen zu finden. Schichten kann man eher allgemein, z. B. als morphologische oder syntaktische Schichten (mit ihren physikalischen Pendants, den entsprechenden *AZ*) festlegen oder detailliert z. B. als morphologische Schicht zur Analyse der Wortbildung im Nominalbereich. Es hängt von den Untersuchungszielen ab, für welches Abstraktionsniveau man sich entscheidet.

Innerhalb einer *AZ* werden Symbole benutzt, die den Datenbereich im Sinne der jeweiligen *Schicht* kodieren. Die einzelnen Symbole (Computerzeichen) setzen sich zu einem *Annotationskode* (*tag*) zusammen.

Ein *Annotationskode* ist ein Zeichen (bzw. eine Zeichenkette) innerhalb einer *AZ*, das den Wert eines horizontal (zeitlich) angeordneten primären bzw. sekundären Datenbereichs kodiert. Ein System von *Annotationskodes*, das mit der Konzeption der jeweiligen *Schicht* kompatibel sein muss, wird hier als *Kodierungssystem* bezeichnet. Ein phonetisches Kodierungssystem wie z. B. *IPA-Kiel* oder ein morphologisches Kodierungssystem, wie z. B. *Codes For Grammatical Morphemes* in CHAT (MacWhinney 1995, 113ff.) sind Beispiele für *Kodierungssysteme.*

Der Markt der linguistischen Computerprogramme verändert sich zwar nicht so schnell wie der kommerzielle Software- und Hardwaremarkt, ist aber doch ständig in Bewegung. Eine vollständige Darstellung der aktuell erhältlichen Produkte ist deshalb kaum realisierbar. Im Folgenden werden computergestützte Annotationsprogramme in einer Perspektive vorgestellt, die es dem Leser erlaubt, sich im Hinblick auf eigene Untersuchungsziele selbst ein Urteil über existierende und zukünftige Programme zu bilden.

Folgende Beurteilungskriterien werden dabei verwendet:

Einbindung primärer Daten.

Flexibilität bei der Definition von *AZ* und *Annotationskodes* in Hinblick auf bestimmte Untersuchungsziele.

Auswertungsmöglichkeiten der Programme in Hinblick auf die definierten *AZ* und *Annotationskodes.*

Möglichkeiten des Datenaustauschs (Datenformate und Zeichensätze der *Annotationen*).

3.2. Computergestützte Transkripte und ihre Auswertung

Die schriftliche Wiedergabe eines zuvor definierten (ggf. digitalisierten) Datenbereichs kann je nach wissenschaftlichem Anspruch, Datenbeschaffenheit (z. B. ein oder mehrere Sprecher) und Untersuchungszielen auf verschiedene Arten in *AZ* aufgeteilt werden. Leitend ist dabei die Schicht, d. h. das theoretische Pendant zur *AZ*. Sowohl methodologische als auch programmtechnische Gesichtspunkte sind dabei relevant. Zunächst soll die Frage der Umsetzung linguistischer Konzepte in computergerechte Strukturen behandelt werden.

3.2.1. Methodologische Überlegungen

Transkripte (vgl. dazu auch Art. 114) haben folgende Funktionen:

1) Wiedergabe der zeitlichen Dimension des zugrundeliegenden Datenbereiches
2) Wiedergabe (von Aspekten) des Gesprochenen
3) Wiedergabe des Geschehenen
4) Wiedergabe der Kontextinformationen

Zu 1):

Die zeitliche Ausdehnung des zugrundeliegenden Datenbereiches war früher (implizit) durch die Linearität der schriftlichen Wiedergabe abgebildet. Die neuesten Computergenerationen und Programme erlauben eine explizite Bindung von Datenbereichen an *AZ*, die aus *Annotationskodes* mit Zeitangaben bestehen (sog. *time linking*). Die Funktion dieser Anbindung ist die Aufteilung der Daten in relevante *Datenbereiche.* Je nach Untersuchungsziel, können bestimmte Teilbereiche der Daten, wie z. B. die Äußerungen nur eines Sprechers, oder nur Fragesätze als Daten relevant sein. Die Entscheidung über relevante Datenbereiche wirkt sich auf die Möglichkeiten von Auswertungsprogrammen aus:

Wenn z. B. jede Äußerung in der zugrundeliegenden Datei zeitlich mit dem Transkript verbunden ist (*time linking*), kann das Ausgabeprogramm den definierten Datenbereich zusammen mit den *Annotationskodes* der zugrunde liegenden Äußerung präsentieren, wenn zuvor nur eine grobe Textgliederung durchgeführt worden ist, wird das Ausgabeprogramm den gesamten zeitlich verbundenen *Datenbereichs* präsentieren, innerhalb dessen sich die gewünschten Äußerung befindet. In zeitlich grob gegliederten Dateien werden deshalb (oft unnötig) große *Datenbereiche* mit ausgegeben, auch wenn die darin enthaltenen *Annotationskodes* sich auf kleinere *Datenbereiche* beziehen.

Zu 2):

Bei der Entscheidung darüber, wie viele und welche *AZ* man für die Sprachwiedergabe benutzt, ist folgendes zu beachten. Traditionell enthalten *Sprachwiedergabeannotationszeilen* (*SAZ*) recht unterschiedliche Informationen: (a) Sprechername (eine Information zum Kontext), (b) Sprecherwechsel (eine an die Zeitachse gebundene Information), (c) *Annotationskodes*, die sich auf die Verarbeitung des Transkripts beziehen, z. B. eine alternative Interpretation, z. B. eine zweite Version von schwer verständlichen Passagen. Manchmal auch (d) Hilfsinformationen zum Kontext oder (e) zur Verständlichkeit. Auch *ad hoc* Kommentare zur Prosodie oder *Annotationskodes* zur Morphemkodierung etc. sind nicht selten in dieser *AZ* enthalten. Diese Vorgehensweise hat den Nachteil, dass das Kodierungssystem einer einzigen *AZ* heterogene (oft *ad hoc* entwickelte) *Kodierungssysteme* enthält (z. B. ein System zur Lautwiedergabe, eines zur Kodierung von Handlungen, eines zur Kodierung der Überlappung der Rede von verschiedenen Sprechern, etc.).

Die *SAZ* ist für die weitere Verarbeitung von großer Bedeutung (deshalb die verbreitete Bezeichnung *main tier*). Die Heterogenität der Informationen (Vermischung von Informationen aus mehreren *Schichten*) kann Nachteile für Datenauswertung nach sich ziehen.

Die *SAZ* wird öfter für eine grobe Orientierung über den Formenbestand der Daten benutzt: Programme zur Wortlistenerstellung (*type-token* Listen) benutzen diese Zeile als Datenquelle. Für eine bessere Übersichtlichkeit und Flexibilität bei der Datenauswertung empfiehlt es sich, die SAZ nur für die Kodierung der Wortformen zu benutzen, und andere Informationen auf andere homogen definierte *AZ* zu verteilen.

Die SAZ kann direkt mit der Zeitachse verbunden werden (*time linking*), um Überlappungen anzuzeigen. Die Kennzeichnung von Sprechern muss eindeutig sein. Wenn Programme die Option der Sprecheridentifikation nicht anbieten, sollte für jeden Sprecher eine neue *AZ* definiert werden.

Die Wahl des Kodierungssystems dieser Schicht (IPA-Konventionen, Orthographieregeln, etc.) ist für die Computerauswertung zweitrangig. Das Kodierungssystem muss allerdings gut beschrieben und seine technische Realisierung innerhalb eines Annotationsprogramms (Eingabehilfen für z. B. IPA-Zeichen oder andere nicht lateinische Zeichen) muss gewährleistet sein.

Zu 3) und 4):

Wie aus den Ausführungen oben leicht zu folgern ist, ist es empfehlenswert, für die Kodierung außersprachlicher Merkmale des zugrundeliegenden Datenbereiches separate *AZ* zu definieren. Die Anzahl und inhaltliche Ausrichtung dieser *AZ* hängt wiederum von den Forschungszielen ab. Gestik-Forscher werden sicherlich mehr und feiner gegliederte außersprachliche Informationen kodieren wollen als andere Sprachforscher, für die eine Kommentar-*AZ* mit relativ intuitivem (oft natürlich-sprachlichen) Kodierungssystem (zum Kodieren von z. B. lachend oder leise gesprochenen Passagen) ausreichend scheint.

3.2.2. Relevante Ein- und Ausgabeoptionen der Transkriptionsprogramme

A) Flexible Definition von *AZ*: Um verschiedene Konzepte in verschiedenen *AZ* kodieren zu können, muss der Benutzer die Möglichkeit haben, beliebig viele Zeilen zu definieren. Außerdem muss das Programm klar definierte und exportierbare Dateie-Formate haben, die die Struktur der *AZ* korrekt markieren. Dadurch können Daten in andere Datenformate (z. B. Datenbankformate oder Textformate) konvertiert werden. XML-basierte Datenformate haben den Vorteil, dass sie, auch wenn die Programme keine Exportfunktionen anbieten, eindeutig und einfach in neue Daten-Formate übertragbar sind.

B) Integration von Mediendateien (Audio- und Videodateien in verbreiteten Datenformaten) in die Transkripte, sowie Vorhandensein einer eigenen *AZ* zum Kodieren der zeitlichen Bezüge zur Mediendatei. Die Vorteile dieser Option liegen darin, dass das Transkribieren

selbst schneller durchgeführt werden kann als bei Benutzung traditioneller Medien. Wenn die Zeitangabe an bestimmte, für die Analyse relevante Phänomene (z. B. Sprecherwechsel oder Äußerungsanfang und -ende) gebunden ist, kann diese Information zur weiteren Verarbeitung (z. B. weitere Annotationen) genutzt werden. Auch für das Korrigieren von Transkripten sind die Zeitangaben von unschätzbarem Wert. Eine graphische Darstellung der Tonwellen erleichtert die Orientierung in den Daten zusätzlich. Eine erwünschte Option wären Spezifizierungen (Bezeichnungen) der Zeitbezüge. Damit sollte es beispielweise möglich sein, Korrektur-Zeitangaben von analyserelevanten Zeitangaben, wie Sätze oder Textteile zu unterscheiden.

C) Definition von verwandten *AZ*: Um z. B. mehrere *SAZ* für mehrere Sprecher zugleich als gleichartig (einer Schicht zugehörig) und doch verschieden zu markieren, sollten Programme die Möglichkeit anbieten, *AZ* als verwandt zu definieren (*parent-parent relation*), bzw. dies im Kommentarteil festzulegen. Die Zeilen können dadurch bei der Ausgabe unterschiedlich behandelt werden, beispielsweise um die Ausgabe auf bestimmte Sprecher zu limitieren. Wünschenswert wäre auch eine Eingabehilfe für Korrekturen bzw. alternative Kodierungen. Dazu könnte vom Programm eine der *SAZ* untergeordnete (*parent-child relation*) *AZ* angeboten werden, auf der der *Annotationskode* der einen Zeile mit dem *Annotationskode* der anderen Zeile programmintern verbunden ist. Die Verwaltung des Korrektur-*Annotationskodes* auf der untergeordneten Zeile, sowie seine Verbindung zur Hauptzeile müsste vom Programm derartig verwaltet werden, dass beispielweise die Funktion der Wortlistenerstellung diesen alternativen Annotationskode berücksichtigt. Ist diese Option nicht vorhanden, so ist es zu empfehlen, eine nur für Korrekturen bestimmte *AZ* selbst zu definieren und die dort gebrauchten *Annotationskodes* gesondert zu markieren.

D) Benutzung verschiedener Zeichensysteme: Um die traditionell zum Transkribieren benutzten Zeichensätze (z. B. IPA-Zeichensätze oder auf nicht lateinischen Buchstaben basierte Schriftsysteme) eingeben zu können, müssen die Transkriptionsprogramme Eingabehilfen (z. B. Veränderung des Tastaturlayouts) anbieten. Das Mischen verschiedener Zeichensätze innerhalb einer AZ sollte ebenfalls möglich sein. Die benutzten Zeichenkodierungen (*encoding*) müssen vom Programm unabhängig sein (d. h. keine programminternen Zeichendarstellungen), damit Datenexport möglich ist. UNICODE-basierte Kodierung garantieren die Austauschbarkeit von Daten mit anderen Programmen und die Darstellbarkeit in verschiedenen Systemen. Auf die nicht triviale Problematik der Kompatibilität von unterschiedlichen Betriebssystemen (Mac, Unix, PC) kann an dieser Stelle nicht eingegangen werden. Meine Darstellung beschränkt sich auf PC-basierte Programme.

Die Funktionalität von allen vier oben genannten Punkten wird (bis auf die unter (C) genannte Unterstützung der Verwaltung von Korrekturzeilen) meines Wissens nur bei einem einzigen Programm angeboten: ELAN (vgl. Brugman/Wittenburg 2001, 69 ff.) ist frei im Internet verfügbar (*http://www.mpi.nl/tools/*). Die Vielfalt und Funktionalität der auf dem Markt verfügbaren auf Transkription spezialisierten Annotationsprogramme ist zuletzt in Dittmar (2002) beschrieben. Eine zusätzliche Leistung für diesen Bereich bieten Programme an, die durch integriertes „linguistisches Wissen" *Datenbereiche* bzw. *AZ* (semi-) automatisch analysieren. Nützlich für bestimmte Fragestellungen ist die Möglichkeit, phonetische Analysen von ausgewählten Bereichen von Daten automatisch durchführen zu können. Das bekannteste Programm für phonetische Analysen ist Praat (*http://www.fon.hum.uva.nl/praat/*). Wünschenswert wäre auch, dass solche Programmmodule in Transkriptionsprogramme integriert werden.

Ein weiterer Bereich der auf linguistischem Wissen basierenden Verarbeitung von Transkripten ist die automatische Lemmatisierung der Annotationskodes aus der SAZ. Automatische Syntax-Parser gehören ebenfalls dazu. Diese Programme, die innerhalb der Corpuslinguistik entwickelt wurden, sind auf Schriftsprache spezialisiert und funktionieren meistens nur auf Großrechnern. Sie stehen an der Grenze zu Annotationsprogrammen, in denen linguistisches Wissen von den Forschern selbst annotiert wird. Die Eingabe der Annotationsprogramme muss es in diesem Fall ermöglichen, dieses Wissen flexibel

zu kodieren. Die Ausgabe der Annotationsprogramme muss gewährleisten, dass das Wissen zur weiteren Auswertung zur Verfügung gestellt wird. – Die Darstellung der weiteren Arten von Annotationen wird nach Kriterien der traditionellen, linguistisch relevanten Forschungsziele unterteilt. Der Hauptunterschied zu den zuvor charakterisierten Annotationen liegt darin, dass die Transkripte die *primären Datenbereiche* annotieren, die an dieser Stelle beschriebenen Annotationen dagegen charakterisieren diese nur mittelbar; direkt beziehen sie sich auf die in der *SAZ* enthaltenen *Annotationskodes*.

3.3. Annotationen, die sich an Lexik-und Paradigmenanalyse orientieren

3.3.1. Methodologische Überlegungen

Die Forschung in dem an dieser Stelle behandelten Bereich hat folgende Ziele: Erstellung von relativ kleinen Lexika (auch vorläufige Listen), die zu einzelnen Lemmata (Grundformen) semantische Charakterisierungen (Übersetzung in andere Sprachen, semantische Varianten der Lexeme), phonetische und morphologische Informationen (Aussprache, Zugehörigkeit zu Paradigmen) zusammenfassen und aus einem Korpus stammende Beispiele ihrer Verwendung enthalten. Gemeint sind z. B. im Entstehen begriffene Lexika zu relativ wenig beschriebenen Sprachen, wie sie in der anthropologisch orientierten Linguistik vorkommen, oder Lexemsammlungen, die der Charakterisierung einer bestimmten Sprachvariätet (z. B. einer Fachsprache oder einer Entwicklungsstufe im Spracherwerb) dienen. Größere lexikographische Projekte, die ebenfalls computergestützt auf Großrechnern arbeiten, können an dieser Stelle nicht präsentiert werden. Die in dieser Forschung entstehenden Lexika (Ansammlungen von Grundformen mit Zusatzinformationen) können eine der primären Funktionen von Lexika erfüllen (z. B. Übersetzungshilfe). Sie können auch ihren praktischen Nutzen haben, indem sie bei der Beschreibung der Grammatik einer bestimmten Varietät Informationen über paradigmatische Verhältnisse (z. B. die Vielfalt von *types*) in einem Korpus liefern. Die doppelte Funktion dieser *AZ* ist vielleicht vom methodologischen Standpunkt aus zu beklagen, hat sich aber als durchaus praktisch erwiesen. Die Endergebnisse dieser Forschung (ein Lexikon oder eine Grammatik) fallen zwar in unterschiedliche linguistische Bereiche, in der Phase ihres Entstehens (Beschreibung und Erforschung einer Sprachvariätet) scheint eine gemeinsame Kodierung jedoch von Vorteil zu sein. Die Grundform (Lemma bzw. Paradigmavertreter) kann in der Endphase korrekt definiert werden, bei der Annotierung der Daten dient sie als gemeinsame Bezeichnung für beides.

Je nach Kenntnisstand über die Grammatik der Varietät, Forschungsziel und Methode kann entschieden werden, in wie viele verschiedene *AZ* die morphologische Charakterisierung der Wortformen aufgeteilt wird. Die Alternativen dafür bedingen einander: (A) Bei nur wenigen verwendeten *AZ* (z. B. eine Zeile für die morphologische Kodierung von Formen aller Wortarten) wird das *Kodierungssystem* ausreichend breit und dadurch heterogen sein müssen. (B) Bei Verteilung der gleichen Information auf mehrere *AZ* (z. B. eine separate Zeile für jede relevante Wortart) bleibt das *Kodierungssystem* übersichtlich und methodologisch homogen. Für linguistisch gut beschriebene Sprachen, für die eine bestehende Grammatik Hilfe bei der Definition liefern kann, sind beide Alternativen möglich, (B) ist jedoch vorteilhafter. Für in der Entwicklung befindliche Varietäten (z. B. Lernersprachen) oder Sprachen, deren grammatische Beschreibung erst im Entstehen ist, bleibt für den Anfang der Annotation nur die Alternative (A). Eine feinere Gliederung in *AZ* kann Vorteile bei der Ausgabe haben: So kann z. B. morphologisch markierte Wortbildungsinformation in den Aufbau des Lexikons eingehen, während die anderen Informationen in die grammatische Beschreibung integriert werden. Bei der Eingabe ist die Aufteilung in mehrere *AZ* oft aufwändiger, besonders, wenn die Programme keine Option anbieten, ungebrauchte Zeilen unsichtbar zu machen.

3.3.2. Ein- und Ausgabeoptionen

Die Möglichkeit, separate *AZ* festzulegen, wird von den meisten Programmen angeboten. Die Funktion dieser Zeilen kann vom Benutzer selbst oder vom Programm verwaltet werden. Beide Aspekte werden im Folgenden berücksichtigt.

1) *AZ* für die Grundform: Die *Annotationskodes* in dieser Zeile sind Grundformen zu einem aus der SAZ stammendem *Annotationskode* (z. B. *gehen* als Lemma bzw. Paradigmengrundform zu *geht*). Es muss vom Programm gewährleistet sein, dass diese Relation mit gespeichert wird.

Eine wünschenswerte Eingabehilfe für diese Zeile wäre eine interaktive (semiautomatische) Lemmatisierung, die z. B. eine Datenbank in Verbindung mit *token*-Kodierungen verwendet, um Vorschläge für Grundformkodierung zu liefern. In diesem Modus könnte für eine bereits lemmatisierte Wortform (z. B. *geht*) vom Programm ein Vorschlag angeboten werden, der die entsprechende Grundform im Grundformenbestand (in diesem Fall *gehen*) sucht. Eine Hilfe zur Disambiguierung von Wortformen wäre auch von großem praktischen Vorteil. Um z. B. die Wortform *sein* als Infinitiv und als Possessivpronomen zu unterscheiden, könnten in diesem Fall zwei Grundformen angeboten werden.

2) *AZ* für semantische Kodierung: Damit Wortübersetzungen, Verweise auf andere Lemmata, Wortartencharakteristika von Grundformen etc. annotiert werden können, müssen diese Zeilen eine Verbindung (unterordnende *parent-child relation*) sowohl zu der *AZ* für Grundformen und als auch zu den darin enthaltenen *Annotationskodes* aufweisen (siehe oben 1). Bei bestimmten Forschungszielen sollte diese Information in mehrere verwandte *AZ* aufgeteilt werden. Dies erlaubt eine spezifischere Steuerung der Ausgabe. Eine separate Annotionszeile für Wortart ist auf jedem Fall von Vorteil.
3) Morphologische Informationen zu *Annotationskodes* in der SAZ: Zum Aufbau von Paradigmen einer Varietät, aber auch als Hilfe zur syntaktischen Annotierung, ist eine Charakterisierung der Bestandteile von Wortformen notwendig. Die dafür geeigneten *AZ* müssen die *Annotationskodes* in dieser Zeile (z. B. *3.sg*) mit den aus der SAZ stammenden *Annotationskodes* für Wortformen (z. B. *geht*) verbinden. Eine weitere sinnvolle Relation wäre auch die Verbindung zu der Information über die Wortart (z. B. *Verb*) und der Grundform (z. B. *gehen)*. Diese Art der Annotationen erlaubt dem Benutzer bei der Ausgabe zusammenhängende Informationen abzufragen. Wünschenswerte Eingabehilfen für morphologische Annotierung sind: Erkennung von bereits annotierten Zeichenketten mit interaktiver Eingabe, Erkennung von ähnlichen Zeichenketten mit interaktiver Eingabe, die Möglichkeit „linguistisches Wissen" zu definieren (z. B. Stamm-Endung Relationen für bestimmte Wortarten), um automatisch (aber mit Korrekturmöglichkeit) ganze bereits transkribierte Texte zu annotieren.
4) Exportmöglichkeiten: Die unter 1 bis 3 beschriebenen Annotationsarten sind Vorstufen zur Erstellung von Endprodukten (Grammatiken bzw. Lexika), die auch wenn sie einen nur vorläufigen Charakter haben, der Präsentation und Verarbeitung in anderen Programmen dienen. Nur wenige Annotationsprogramme bieten z. B. die Möglichkeit, formatierte Lexika zu erstellen. Ein Export der Annotationsdaten in Datenbank- oder Textformate ist eine Möglichkeit, diese weiter auszuwerten bzw. zu publizieren. Der Export zu Datenbanken bietet mehr Möglichkeiten der Weiterverarbeitung als der Textexport. Dabei ist es wünschenswert, dass nicht nur Informationen über die Struktur (*AZ* und ihre Beziehungen zueinander) und ihre Werte (*Annotationskodes*) exportiert werden können, sondern auch die relevanten primären Datenbereiche. Multimediale Datenbanken haben mehr Verarbeitungsmöglichkeiten von Daten als Annotationsprogramme und eignen sich (ab einer bestimmten Verarbeitungsstufe) besser dazu, weitere Verarbeitungsschritte (z. B. statistische Analysen) zu übernehmen.

3.4. Syntax-orientierte Annotationen

Syntaxkodierung ist in größerem Maße als andere Annotationsarten theorieabhängig. An dieser Stelle wird versucht, die Annotierungsmöglichkeiten theorieneutral zu besprechen. Jede Syntaxbeschreibung muss sowohl der Linearität der Sprache (Reihenfolge der Wortformen) Rechnung tragen als auch verschiedene Relationen zwischen den Wortformen bzw. zwischen den syntaktischen Funktionen darstellen. Im Folgenden sollen diese Relationsarten und die Möglichkeiten zur Annotierung dargestellt werden.

3.4.1. Methodologische Überlegungen

1) *AZ* zur Bezeichnung von syntaktischen Funktionen: Die Kodierung von Funktionen wie Subjekt, Objekt etc. muss zumindest die Information über die Verbindung des Annotationsskodes zu der Wortform (d. h. zu dem Annotationskode aus der Sprechwiedergabezeile) mitenthalten, die damit charakterisiert wird. Dabei muss es möglich sein, zeitlich voneinander getrennte Formen als Einheit zu

betrachten und gemeinsam zu kodieren. Wenn Programme diese Option nicht anbieten, ist es erforderlich dies im Kodierungssystem (z. B. durch Definition von eindeutigen Indizes) zu berücksichtigen. Alle Arten von Funktionen (z. B. semantische Rollen) können auf diese Art in separaten Zeilen kodiert werden. Diese AZ enthalten implizit oder explizit auch Wortfolgeinformationen und Information über die Grenzen der annotierten Einheit (z. B. Satzgrenzen).

2) *AZ* für die Kodierung von Relationen zwischen Funktionen: Wenn in der Theorie notwendig, muss es möglich sein, Abhängigkeitsverhältnisse zwischen den *Annotationsskodes* (unter 1) zu kodieren. Wenn Programme keine eigene Kodierung (z. B. Interpretation von Klammerungen) dafür anbieten, müssen *Kodierungssysteme* diese Funktion ersetzen. Wünschenswert wäre eine automatische Informationsübertragung über die Reihenfolge der Syntagmen (unter 1). Ggf. muss dies im Kodierungssystem berücksichtigt werden.

3) Andere Relationen: Um die Flexibilität der Anwendung auf mehrere Syntaxtheorien zu gewährleisten, ist es wünschenswert eine Option anzubieten, die es erlaubt, mehrere horizontale und vertikale Relationen zu definieren, zum Beispiel die Merkmalvererbung innerhalb einer syntaktischen Einheit (*head-complement* Relation).

4) Exportmöglichkeiten: Datenexport zur Weiterverarbeitung in Datenbanken und zur Publikation (z. B. Beispielsätze) sollte selbstverständlich möglich sein.

3.5. Konversationsanalytisch orientierte Annotationen

Die Eigenarten dieser Forschungsrichtung (Konversationsanalyse, Kommunikationsforschung) liegen einerseits in der Verwendung von relativ neuen und wenig verbreiteten Kodierungsschemen, andererseits ist diese Forschung stärker als andere auf die Kodierung von prosodischen und lexikalischen Phänomenen angewiesen, die für die Textgliederung verantwortlich sind. Beide Forschungserfordernisse lassen sich im Rahmen der bisher beschriebenen Funktionalität von Annotationsprogrammen realisieren.

3.5.1. Ein und Ausgabeoptionen

1) Kodierung von prosodischen Informationen: Wie oben erläutert, ist durch die Medieneinbindung in Annotationen (*time linking*) die Möglichkeit entstanden, schnell auf analysierte Datenbereiche, die mit der Zeitachse verbunden sind, zuzugreifen, und diese ggf. zu exportieren bzw. automatische Intonationanalysen durchzuführen. Die Kodierung mit eigens dafür entwickelten Kodierungssystemen in dafür definierten *AZ* wird dadurch selbstverständlich nicht beeinträchtigt.

2) Kodierung von Textgliederungen: Um kommunikationstheoretische *Annotationsskodes* in Verbindung zu Datenbereichen zu bringen, kann die Option der Verbindung von *Annotationsskodes* zur Zeitachse genutzt werden. Die Option der Definition von abhängigen *AZ* (*parent-child relation*) und die Zeitmarkierung ermöglichen die Darstellung von Relationen zwischen den relevanten Textteilen.

3.6. Zusammenfassung

Semi-automatische Annotationshilfen sind oft auf bestimmte Untersuchungsziele ausgerichtet. Auch die Ausgabemöglichkeiten der Programme sind auf die Untersuchungsziele zugeschnitten. Die zugehörigen Handbücher erklären ihre Benutzung mit Beispielen aus der eigenen Forschungsrichtung. Eine Veränderung der Untersuchungsziele ist nur mit sehr guten Kenntnissen des jeweiligen Annotationssystems möglich.

Es ist deshalb nicht verwunderlich, dass bestimmte Forschergemeinschaften, sich auf bestimmte Annotationsprogramme und ihre Konventionen eingeschworen haben. Eine extrem getrennte Benutzung von Programmen mit vergleichbarer Funktionalität innerhalb verschiedener geschlossener Forschergemeinschaften ist z. B. bei der Benutzung von *clan* (MacWhinney 1995) und *shoebox* (*http://www.sil.org/*) feststellbar. Die Spracherwerbsforschung benutzt das dafür entwickelte *clan* im Gegensatz zur anthropologisch orientierten Linguistik, die ihre Daten meistens mit Hilfe von *shoebox* annotiert. Um die Flexibilität der Programme zu steigern, werden alternative Annotations- und Kodierungssysteme präsentiert. Barnet/Codó/Eppler et al. (2000, 131 ff.) machen beispielsweise einen Vorschlag zur Erweiterung der Möglichkeiten von CHILDES; Lieb /Drude (2001, 3 ff.) machen einen Vorschlag für die anthropologische Forschung.

Die Aussichten dafür, dass flexible Programme alle oben beschriebenen Optionen anbieten, stehen gut. Die technischen Entwicklungen und die Preise für Speichermedien führen dazu, dass solche Optionen auch auf lokalen Rechnern realisierbar sind.

Mit der Option *templates* für Annotations- und Kodierungsschemata definieren zu können, werden sie zu flexiblen Werkzeugen für spezialisierte Forschergruppen. Die dort erstellten, einheitlich formatierten Daten können als Teile zum Aufbau von größeren homogenen Korpora dienen. Die Entwicklung auf diesem Softwaremarkt ist nicht kommerziell, sondern wird von wissenschaftlichen Institutionen vorangetrieben. Große und finanzkräftige internationale Projekte, die den Aufbau von Korpora unterstützen (siehe weiter unten 4.) sorgen auch im Bereich der Softwareentwicklung für gute Qualität. Auf internationalen Konferenzen, vgl. z. B. [*http://morph.ldc.upenn.edu/*] wird über laufende Softwareprojekte in verschiedenen Entwicklungsphasen berichtet. Über den aktuellen Stand informiert man sich am besten im Internet.

4. Korpusaufbau

Um einfachen Datenaustausch zwischen Sprachforschern zu ermöglichen, ist eine Konzentration von Daten von Vorteil. Korpora sind Datenansammlungen, die nach bestimmten inhaltlichen und formalen Kriterien zusammen gesetzt sind (vgl. auch Art. 111).

Die wachsende Verbreitung des Internets erlaubt einen neuen Zugang zu Sprachdaten. Die ersten multimedialen Sprachkorpora, die im Internet zugänglich wurden, waren mit den üblichen Browsern (http-Protokolle) erreichbar. Zu erwähnen in diesem Zusammenhang CHILDES Datenbank (McWhinney 1995) für Erstspracherwerbsdaten. Neue internationale Projekte und Initiativen treiben eine neue Entwicklung voran. XML-interpretierende Browser bieten mehr Möglichkeiten an, Daten im Internet zugänglich zu machen. Z. B. DoBeS – **Do**kumentation **Be**drohter **S**prachen (*http://www.mpi.nl/DOBES/*), OLAC – Open Language Archives Community (*http://www.language-archives.org/*) bauen umfangreiche multimediale Sprachkorpora auf und wenden die neue Technologie an. (vgl dazu auch: Linguistic Data Consortium (*http://www.ldc.upenn.edu/*).

Das am weitesten vorangeschrittene Projekt in diesem Bereich ist das Konzept eines *Browsable Corpus* (Broeder et al. 2001, 48 ff.). Der praxisorientierte Teil des Projekts bestand unter anderem aus der Entwicklung eines spezialisierten Browsers für die Präsentation von linguistischen Daten. Die Daten werden in Form eines nach verschiedenen Kriterien (z. B. projekteinterne Klassifizierungen) strukturierten Datenbaums präsentiert. Die Metadaten zu einzelnen *sessions* (vgl. oben 2.2.) bilden die Hauptquelle für Informationen über die Datenbestände. Nach Metadaten kann gezielt gesucht werden, d. h. Informationen aus auswählbaren Metadatenbereichen (z. B. Alter des Sprechers, soziale Charakteristik etc.) können abgerufen werden. Solche Metadaten sind meistens frei zugänglich. Die zu *sessions* gehörigen und ggf. annotierten Medien werden ebenfalls integriert und können nach von den beitragenden Wissenschaftlern definierten Regeln (frei, eingeschränkt auf Medien oder durch Passwort geschützt) anderen Benutzern zugänglich gemacht werden.

Die durch solche Entwicklungen begünstigten, nicht-kommerziellen Austauschmöglichkeiten für linguistische Daten erlauben eine bessere Zusammenarbeit zwischen räumlich getrennten Forscherteams. Sie verhindern auch, dass die mit viel Aufwand gesammelten und verarbeiteten Daten als Datenfriedhöfe ihr Dasein beenden. Eine mehrmalige Benutzung der gleichen Daten ermöglicht Modifizierungen und führt zu einer vielfältigen und aspektreichen Datenanalyse in der empirisch orientierten Linguistik. Diese Entwicklung erfordert jedoch seitens der Wissenschaftler eine präzise und nachvollziehbare Annotation. Computerprogramme können nur vor formalen Ungenauigkeiten warnen.

5. Literatur (in Auswahl)

Barnet, Ruthana/Codó, Eva/Eppler, E. et al. (2000) „The LIDES coding manual: A document for preparing and analyzing language interaction data“, in: *The International Journal of Bilingualism* 4 (2), Special Issue.

Bird, Stephen/Liberman, Mark (2000) „A formal framework for linguistic annotation“, in: *Speech-Communication* 33 (1,2), 23–60. [ftp://ftp.cis.upenn.edu/pub/sb/papers/specom01/ specom01.pdf]

Broeder, Dan/Offenga, Freddy/Willems Dan/ Wittenburg, Peter (2001) „The IMDI Metadata set, its Tools and accessible Linguistic Databases“, in: *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, 48–55. [http://www.ldc.upenn.edu/annotation/database/papers/Broeder_etal/32.3. broeder.pdf]

Brugman, Hennie/ Wittenburg, Peter (2001) „The Application of annotation models for the construction of databases and tools – an overview and analysis of the MPI work since 1994“, in: *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, 65–73, [http://www.ldc.upenn.edu/

annotation/database/papers/Brugman_Wittenburg/20.2.brugman.pdf]

Dittmar, Norbert (2002) *Transkription. Ein Leitfaden mit Aufgaben für Studenten, Forscher und Laien*, (Qualitative Sozialforschung 10) Berlin.

Lieb, Hans-Heinrich/Drude, Sebastian, (2000) „Advanced Glossing: A Language Documentation Format", (unpublished Working Paper).

[http://www.mpi.nl/DOBES/applicants/dobes-ling-aspects-lang-doc.html].

MacWhinney, Brian (1995) *The CHILDES Project: Tools for Analyzing Talk*, 2nd ed., Hilsdale, New Jersey.

Shoebox – *http://www.sil.org/computing/shoebox/*

Skiba, Ronald (1998) *Fachsprachenforschung in wissenschaftstheoretischer Perspektive*, Tübingen.

Web Links:

http://morph.ldc.upenn.edu/
http://www.fon.hum.uva.nl/praat/
http://www.language-archives.org/
http://www.ldc.upenn.edu/
http://www.mpi.nl
http://www.mpi.nl/DOBES/
http://www.mpi.nl/ISLE/
http://www.mpi.nl/ISLE/schemas/schemas_frame.html
http://www.mpi.nl/tools/
http://www.sil.org/

Romuald Skiba, Nijmegen (Niederlande)

121. Ethnographic Description/Ethnographische Beschreibung

1. Introduction

The central aims of ethnographic description in sociolinguistic research are to document and to analyze specific aspects of the practices of talk as those practices are situated in the society in which they occur. The focus, then, is at once on social situations of use, on the ordinary and persistent habits of use, and on the specific linguistic and behavioral organization of the usage itself.

In the actual conduct of ethnographic research data collection and data analysis are mutually constitutive. Because of this, substantive perspectives that inform ethnographic analysis need to be discussed, as well as the processes of observation and the creation of data records upon which a descriptive account is based. Accordingly this article begins by considering the main substantive foci of ethnographic description.

First a definition of ethnography and a brief survey of its origins is presented. This is followed by discussion of four essential characteristics of ethnography: (1) its particularistic focus on the specifics of naturally occurring performance in speaking; (2) its general focus on social and cultural entities, considered and described as whole systems in comparison with other systems in other societies: (3) its focus on the social meaning of utterances in addition to their referential meaning; and (4) its focus on the meanings of naturally occurring social action from the points of view of the actors engaged in it.

The first two of these characteristics are especially distinctive of ethnography in contrast to other approaches to sociolinguistic research. The latter two characteristics are shared with some approaches in sociolinguistics but not with others. Correlational research in sociolinguistics is the sort of work that differs most from ethnographic description. Correlational studies have been of two main types. In the first type some aspect of language choice (e. g., code, dialect, register, or politeness formula) is considered a discrete variable that is correlated with one or more attributes of the social identity of individual speakers (e. g., income, educational level, or political affiliation). In the other type of study, the direction of correlation is reversed; one or more discrete features of social identity (e. g., gender, ethnicity, or class) are correlated with a discrete feature of language style. The data for such studies are typically collected by survey methods, and the aspects of language and discourse that are studied are considered in abstraction from their situations of use. In contrast, ethnography as a naturalistic approach to social inquiry proceeds by direct observation of concrete situ-