*Thomas Schmidt*[1]

# Korpus „Skandinavische Semikommunikation“ – ein mehrsprachiges Diskurskorpus auf XML-Basis

## 1 Datenbank Mehrsprachigkeit und EXMARaLDA

Der Sonderforschungsbereich 538 „Mehrsprachigkeit“ an der Universität Hamburg vereinigt in seinen dreizehn Teilprojekten eine Vielzahl von Forschern, die sich unter verschiedenen Herangehensweisen mit dem Thema der Mehrsprachigkeit auseinandersetzen. Die weitaus größte Zahl der Projekte arbeitet dabei empirisch auf der Grundlage von Aufnahmen gesprochener Sprache, die durch eine computergestützte Transkription der Analyse zugänglich gemacht werden.

Das Projekt „Datenbank Mehrsprachigkeit“ hat sich zum Ziel gesetzt, ein gemeinsames Dach für diese Transkriptionsdaten zu schaffen, unter dem es möglich sein soll, Transkriptionen zwischen einzelnen Teilprojekten auszutauschen oder Daten aus verschiedenen Projekten gemeinsam zu analysieren. Das in diesem Beitrag vorgestellte Korpus „Skandinavische Semikommunikation“ vereinigt die im Teilprojekt K5 „Semikommunikation und rezeptive Mehrsprachigkeit im heutigen Skandinavien“ erstellten Transkriptionsdaten und gilt als Testfall und Prototyp für die Datenbank Mehrsprachigkeit.

Wie z.B. in Schmidt (2001, 2002a, 2002b, i.V.) dargestellt, fungiert dabei die XML-Anwendung EXMARaLDA als zentrale Architekturkomponente. Sie übernimmt erstens die Rolle einer Interlingua zwischen den vorhandenen älteren Datenformaten. Zweitens dient sie als Ziel- bzw. Ausgangsformat für mehrere Eingabe- und Ausgabemethoden und bildet mit diesen zusammen ein eigenständiges System zur computergestützten Transkription. Drittens schließlich ist sie Grundlage der Datenbank Mehrsprachigkeit und der von ihr zur Verfügung zu stellenden Analyseinstrumente. Abbildung 1 auf der folgenden Seite veranschaulicht dies.

Auch für das hier vorgestellte Korpus „Skandinavische Semikommunikation“ ist daher das EXMARaLDA-System von zentraler Bedeutung. Wie genau bei seiner Erstellung von den verschiedenen Eingabemethoden für Transkriptionsdaten Gebrauch gemacht wird, und welche Ausgabemethoden und Analysewerkzeuge auf die Daten angewandt werden können, wird im nächsten Abschnitt ausgeführt werden.

[1] Mit Dank an die MitarbeiterInnen des Projekts K5 (Frank Stinner, Kurt Braunmüller, Gerard Doetjes, Peer Warter, Franziska Watzke und Ludger Zeevaert) für die Unterstützung.

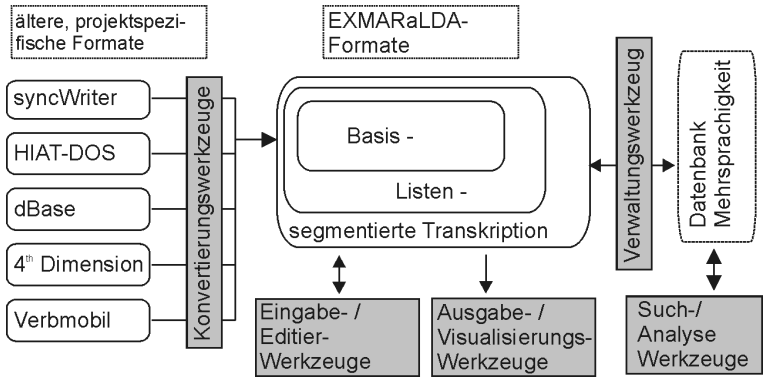


Abb. 1: Systemarchitektur.

## 2 Das Korpus „Skandinavische Semikommunikation“

Das Korpus „Skandinavische Semikommunikation“ besteht aus Transkriptionen von Radiosendungen, Interviews, Unterrichtsdiskursen, Gruppendiskursen und Vorträgen, in denen Sprecher des Dänischen, Norwegischen oder Schwedischen in ihrer jeweiligen Muttersprache miteinander kommunizieren und dabei ausnutzen, dass ihre Gegenüber mindestens rezeptive Kompetenz in diesen Sprachen besitzen. Das Korpus ist die empirische Analyse-Grundlage des Projektes K5, in dem solche Formen der **Semikommunikation** (Braunmüller 2001) – die von einer rezeptiven **Mehrsprachigkeit** zu unterscheiden sind – nach diskursanalytischen Gesichtspunkten untersucht werden.

### 2.1 Eingabemethoden

In der ersten Projektphase (1999-2002) wurden die Transkriptionen des Projektes mit Hilfe des Programms HIAT-DOS (Ehlich 1993) erstellt. Diese Daten wurden mittels eines Importfilters in das EXMARaLDA-Format überführt, allerdings führte diese Konvertierung nur mit einer anschließenden manuellen Nachbearbeitung zu einem befriedigenden Ergebnis. Nicht nur aus diesem Grund, sondern auch, weil sich HIAT-DOS bzgl. des Bedienungskomforts (z.B. keine Mausunterstützung) und speziell mehrsprachiger Belange (z.B. stehen keine skandinavischen Sonderzeichen wie *Å*, *Æ*, *Ø* zur Verfügung) als nicht mehr zeitgemäßes Werkzeug erwiesen hat, wurde der Eingabeprozess für die zweite Phase (seit 2002) grundlegend geändert.

Transkriptionen werden jetzt zunächst mit Praat (siehe Praat-Homepage) angefertigt. Diese Software erlaubt die direkte Synchronisierung von Teilen der Aufnahme mit zugehörigen Transkriptionsausschnitten und stellt darüber hinaus phonetische Visualisierungsverfahren wie Spektrogramme, Oszillogramme etc. zur Verfügung, die als zusätzliche Hilfe beim Transkribieren dienen können.

Das Ergebnis dieser Transkription in Praat wird in den EXMARaLDA-Partitur-Editor geladen. Dies geschieht wiederum mittels eines Importfilters, in diesem Falle ist eine manuelle Nachbearbeitung aber nicht notwendig. Im Partitur-Editor können die Transkriptionen dann gemäß den HIAT-Konventionen (Ehlich/Rehbein 1976), nach denen das Projekt verfährt, weiterbearbeitet werden. Beispielsweise können in getrennten Spuren äußerungsweise Übersetzungen hinzugefügt oder bestimmte Phänomene, die für den Untersuchungsgegenstand relevant sind, annotiert werden. In der untenstehenden Abbildung wurden z.B. in einer Spur Stellen annotiert, an denen Sprecher „code-switchen“, also innerhalb einer Äußerung von einer Sprache in eine andere wechseln. In einer weiteren Spur wurden die Äußerungen ins Englische übersetzt.

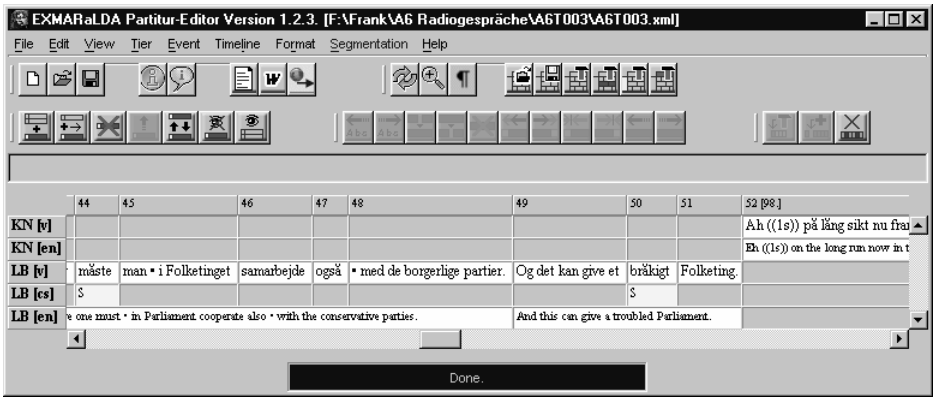


Abb. 2: EXMARaLDA-Partitur-Editor.

Das Konzept von EXMARaLDA stellt dabei sicher, dass solche Annotationen **nach Bedarf** hinzugefügt werden können, d.h. insbesondere, dass auch solche Kategorien, die sich erst während der Analyse als relevant erweisen, flexibel in das System aufgenommen werden können.

## 2.2 Ausgabemethoden

Unerlässliche Grundlage jeder diskursanalytischen Untersuchung sind **Transkripte**, d.h. geschlossene Visualisierungen der transkribierten Phänomene, mittels derer sich der Diskurs in seinem zeitlichen Verlauf auf Bildschirm oder Papier nachvollziehen lässt. Dabei können – abhängig vom anvisierten Analysezweck – für ein und dasselbe Transkriptionsdatum unterschiedliche grafische Organisationsformen für die Visualisierung gewählt werden (vgl. Edwards 1993). Weil auch am SFB „Mehrsprachigkeit" mehrere solcher Verfahren (Partiturnotation vs. Zeilennotation vs. Spaltennotation) miteinander „konkurrieren", stellt das EXMARaLDA-System unterschiedliche Methoden zur Verfügung, um aus der in den XML-Dateien kodierten logischen Struktur der Transkriptionen eine grafische Präsentation zu generieren. Ausgangspunkt kann dabei wiederum der Partitur-Editor sein, der es z.B. erlaubt, eine Transkription als (entsprechend umgebrochene) Partitur auf dem Drucker oder als RTF-Datei auszugeben oder sie als Äußerungsliste gemäß der Zeilennotation in eine HMTL-Datei zu exportieren:

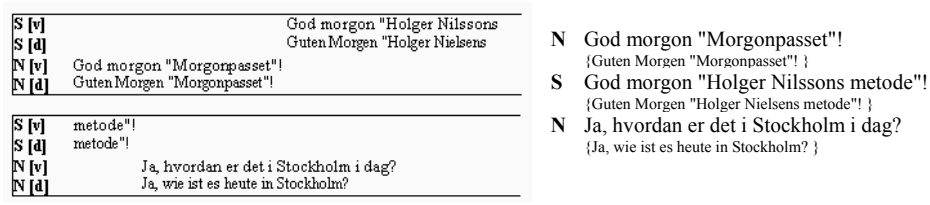


Abb. 3: Ausgabe in Partitur- und Zeilennotation.

Zusätzlich erlaubt EXMARaLDA die Verknüpfung ausgewählter Stellen der Transkription mit Bild-, Ton- oder Videomaterial. Bei der Ausgabe nach HTML können diese Verknüpfungen dann mit jedem gängigen Internet-Browser angezeigt werde. Da dies naturgemäß in einem Aufsatz in Papierform nicht illustriert werden kann, sei für Beispiele auf die EXMARaLDA-Homepage verwiesen.

## 2.3 Analysewerkzeuge

Die typische Anforderung an ein Analysewerkzeug für Transkriptionsdaten besteht in der Unterstützung beim Auffinden bestimmter Wortvorkommen im Transkript. EXMARaLDA stellt zunächst eine einfache Methode zur Verfügung, um ein solches Auffinden in **einzelnen Transkripten** zu erleichtern: Das Transkript und eine zugehörige (alphabetisch geordnete) Wortliste werden in zwei Frames eines HTML-Dokuments ausgegeben und über Hyperlinks miteinander verknüpft. Das Klicken auf

ein Wort in der Wortliste lässt dann die entsprechende Stelle im Transkript in den Bildschirmausschnitt rollen:

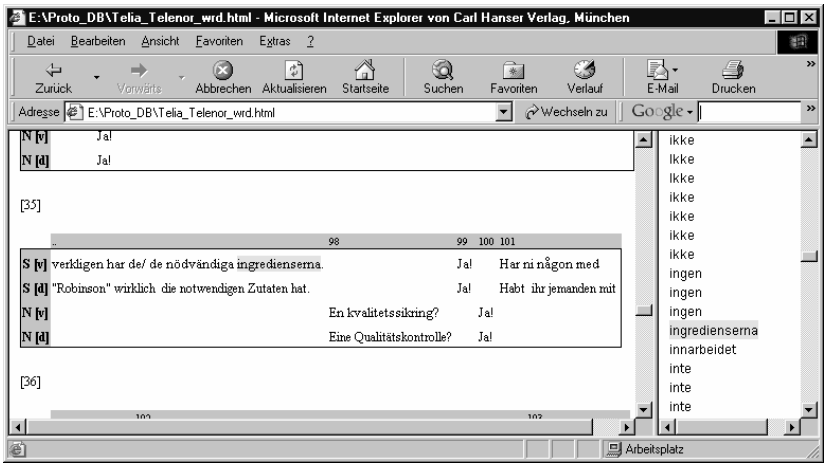


Abb. 4: Verknüpfung von Transkript und Wortliste.

Unerlässlich wird die Computerunterstützung bei der Analyse aber vor allem dann, wenn nicht einzelne Transkripte sondern **das gesamte Korpus** nach bestimmten Phänomenen untersucht werden sollen. Zu diesem Zweck werden die EXMARaLDA-kodierten Transkriptionen des Korpus „Skandinavische Semikommunikation“ in eine relationale Datenbank überführt. Von dort aus können dann sämtliche Suchmechanismen relationaler Datenbankmanagementsysteme genutzt werden, um Transkripte nach verschiedenen transkribierten und annotierten Phänomen zu durchsuchen, z.B.:

| **Suche nach Wort „mycket“** | |
|---|---|
| Transkription | Äußerung |
| **A6T005** | Och/ och sen det sista är/ • • vi hade ett problem här då första säsongen att de fick alldeles för mycket ris. |
| **A6T003** | • Så vi risikerer da • at få et • Folketing som er mycket • afhængig af • konflikt, som er mycket afhængig af • at man kan søge et kompromis, men jeg • tvivler på at det kan blive så mycket. |
| **A6T003** | Jeg tror ikke at det betyder • så forfærdelig mycket • for regeringsbildningen. |
| **A6T003** | Långsiktigt så ka/ får vi den konsekvensen • at det vil være mycket æh besværligt/ • bråkigt for den mindretalsregering at föra en økonomisk politik. |
| **A6T003** | Æh det vil være mycket afhængigt af • om SF er villige • at gå på kompromis i dette frågan. |

| **Suche nach Code-Switches** | | | |
|---|---|---|---|
| Transkription | Sprache | Wort | Äußerung |
| **A6T003** | S | frågan | Det er det centrale frågan • i alle • kampvalg. |
| **A6T003** | S | bilda | Og det vil blive Poul Nyrup Rasmussen som vil bilda en mindretalsregering, men han skal have • stöd • fra Socialistisk Folkeparti. |
| **A6T003** | S | stöd | Og det vil blive Poul Nyrup Rasmussen som vil bilda en mindretalsregering, men han skal have • stöd • fra Socialistisk Folkeparti. |
| **A6T003** | S | risk | Men • • der er ingen • risk for at Centrumdemokraterne • og Radikale Venstre vil föra en økonomisk politik som er afhængig af et/ Socialistisk Folkeparti. |
| **A6T003** | S | föra | Men • • der er ingen • risk for at Centrumdemokraterne • og Radikale Venstre vil föra en økonomisk politik som er afhængig af et/ Socialistisk Folkeparti. |
| **A6T003** | S | bråkigt | Og det kan give et bråkigt Folketing. |

Wichtig ist dabei, dass diese Phänomene zwar isoliert gesucht werden können, dass aber dennoch jederzeit der Rückgriff auf die entsprechende Stelle im Transkript gewährleistet bleibt, um so – wie es einer diskursanalytischen Verfahrensweise entspricht – den zugehörigen **Gesprächskontext** betrachten zu können. Z.Z. erfolgt dieser Rückbezug auf das Transkript noch **statisch** mittels eines einfachen Hyperlinks in ein vorab generiertes HTML-Dokument. In Zukunft sollen aber an dieser Stelle aber entsprechende EXMARaLDA-Werkzeuge die jeweils relevanten Transkriptstellen **dynamisch** aus der Datenbankrecherche heraus erzeugen und anzeigen.

Auch aus den Daten der anderen Teilprojekte des SFB 538 werden Korpora erstellt werden, die dem hier vorgestellten vergleichbar sind. Die Gesamtheit dieser Korpora, zusammen mit zugehörigen Ein- und Ausgabewerkzeugen, werden dann die „Datenbank Mehrsprachigkeit“ bilden. Dadurch, dass die Datenbank mit EXMARaLDA auf einer XML-Anwendung basieren, ist nicht nur die Möglichkeit einer flexiblen Weiterbearbeitung der Daten gewährleistet, sondern es wird darüber hinaus auch deren langfristige Archivierbarkeit sichergestellt.

## Literatur

Braunmüller, Kurt (2001): Semicommunication and Accommodation: Observations from the Linguistic Situation in Scandinavia. *Arbeiten zur Mehrsprachigkeit, Serie B* 17.

Edwards, Jane (1993): Principles and Contrasting Systems of Discourse Transcription. Edwards, Jane/Lampert, Martin (Hrg.): Talking Data – Transcription and Coding in Discourse Research. Hillsdale: Erlbaum.

Ehlich, Konrad (1992): Computergestütztes Transkribieren - das Verfahren HIAT-DOS. Richter, Günther (Hrg.) Methodische Grundfragen der Erforschung gesprochener Sprache, Frankfurt a.M.: P. Lang, 47-59

Ehlich, Konrad/Rehbein, Jochen (1976): Halbinterpretative Arbeitstranskriptionen (HIAT). Linguistische Berichte 45, 21-41.

Schmidt, Thomas (2001): The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. Proceedings of the IRCS Workshop on Linguistic Databases, 219-227.

– (2002a): Gesprächstranskription auf dem Computer: das System EXMARaLDA. *Gesprächsforschung* 2. 1-23.

– (2002b): EXMARaLDA - ein System zur Diskurstranskription auf dem Computer. *Arbeiten zur Mehrsprachigkeit, Serie B* 34.

– (i.V.) EXMARaLDA - ein System zur computergestützten Diskurstranskription. Mehler, Alexander/Lobin, Henning (Hrg.): Werkzeuge zur automatischen Analyse und Verarbeitung von Texten: Formate, Tools, Software-Systeme.

### URLs

EXMARALDA-HOMEPAGE : http://www.rrz.uni-hamburg.de/exmaralda

PRAAT-HOMEPAGE: http://www.praat.org