Erschienen in: Pusch, Claus D./Raible, Wolfgang (Hrsg.): Romanistische Korpuslingustik: Korpora und gesprochene Sprache. - Tübingen: Narr, 2002. S. 209-243. (ScriptOralia 126)

Franck Bodmer (Mannheim), Marcus L. Fach (Stuttgart),
Rudolf Schmidt, Wilfried Schütte (Mannheim)

Von der Tonbandaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für die Arbeit mit Gesprächskorpora

The development of tools for computer-assisted transcription and analysis of extensive speech corpora is one main issue at the Institute of German Language (IDS) and the Institute of Natural Language Processing (IMS). Corpora of natural spoken dialogue have been transcribed, and the analogue recordings of these discourses are digitized. An automatic segmentation system is employed which is based on Hidden Markov Models. The orthographic representation of the speech signal is transformed into a phonetic representation, the phonetic transcription is transformed into a system-internal representation, and the time alignment between text and speech signal follows. In this article, we also describe the retrieval software COSMAS II and its special features for searching discourse transcripts and playing time aligned passages.

1. Vorbemerkung

In diesem Beitrag berichten wir über die Technologie zur Erfassung und Bearbeitung von Korpora gesprochener Sprache am Institut für Deutsche Sprache Mannheim (IDS) und am Institut für Maschinelle Sprachverarbeitung an der Universität Stuttgart (IMS): Welche Anforderungen stellen Linguisten, insbesondere Gesprächsanalytiker, an die Bearbeitung von Gesprächsaufnahmen, und welche Werkzeuge sind dazu am IDS und IMS entwickelt worden?[1]

Nach einer Darstellung dieses Anforderungsprofils an korpuslinguistische Werkzeuge und einem Überblick über die Bearbeitung von Gesprächsaufnahmen und -korpora im IDS (Wilfried Schütte / IDS) beschreibt Rudolf Schmidt (IDS) die Werkzeuge DIDA (zur Erfassung, Pflege und Archivierung von Transkripten) und SPRAT (zur automatischen Text-Ton-Synchronisation und Spracherkennung). Marcus Fach (IMS) stellt das Stuttgarter automatische Segmentierungssystem ALPHONES und Lösungen für Probleme bei der automatischen Text-Ton-Synchronisation vor. Der Prozess des automatischen Text-Ton-*Alignments* als einer Synchronisation von Transkript und Gesprächsaufnahme wird von Marcus Fach als „automatische Segmentierung" bezeichnet. Franck Bodmer (IDS) stellt die Volltextdatenbank COSMAS II vor;

1 Auf der Begleit-CD-ROM zum vorliegenden Band finden sich ergänzende Materialien zu diesem Beitrag (Verzeichnis BODMER_FACH_SCHMIDT_SCHUETTE).

Transkripterfassung und -pflege mit DIDA und Text-Ton-Synchronisation sind Voraussetzungen, um in Diskurstranskripten mit COSMAS II recherchieren und auf Treffer in den Transkripten und in den zugehörigen Sprachaufnahmen zugreifen zu können. Wilfried Schütte liefert für Recherchen in COSMAS II Anwendungsbeispiele und schließt mit einem Ausblick ab.

2. Ein Anforderungsprofil an korpuslinguistische Werkzeuge aus der Sicht von Linguisten und Gesprächsforschern

Was brauchen Linguisten und welche Anforderungen gelten speziell im IDS für die Bearbeitung von Gesprächsaufnahmen? Generell ergeben sich für die Arbeit mit Gesprächskorpora zwei Postulate:

a. Standardisierte Erhebung von Daten (manuell und auditiv im Verlauf einer projektbezogenen Korpusarbeit), Archivierung mit optimalem Zugriff und definierte Zugriffsrechte auf Korpusbestandteile nach Abschluss der Korpusarbeit;
b. Anreicherung von manuell erhobenen Korpora (bestehend aus Aufnahmen und Transkripten) durch eine computergestützte Extraktion neuer Arten von Information.

2.1. Standardisierte Erhebung

Diese Anforderung zielt darauf ab, dass alle manuell und auditiv erhobenen Informationen (z.B. Transkription, Dokumentation, Interpretation), die dezentral gesammelt werden, zentral zugreifbar sind und auch in einer späteren archivierten Form für definierte Nutzerkreise bei Korpusrecherchen für das *Retrieval* zur Verfügung stehen. Dabei muss bei einem modularen Aufbau der Korpustechnologie der Informationsfluss gesichert sein.

Die Aufzeichnung authentischer Gespräche, ihre detaillierte Verschriftlichung (Transkription) und die Analyse dieser Transkripte und Aufnahmen unter aus dem Material entwickelten Fragestellungen bilden den Kern der gesprächsanalytischen Arbeitsweise (für Näheres vgl. Becker-Mrotzek / Meier 1999; Deppermann 1999). Besondere Probleme bereiten beim Umgang mit Sprachdaten:

- die Dokumentation und Archivierung der Aufnahmen,
- die langfristige Aufbewahrung und der Erhalt von Aufnahmen,
- das Auffinden bestimmter Stellen in Aufnahmen,
- das Ermitteln von Vergleichsstellen (innerhalb einer oder verschiedener Aufnahmen) und die Durchführung entsprechender Vergleiche.

Wenn nicht nur einzelne Gesprächsaufnahmen zu Transkripten verarbeitet, sondern viele derartige Aufnahmen nach Varianz- oder Ähnlichkeitsgesichtspunkten zusammengestellt werden, sprechen wir von Korpora. Gesprächskorpora können u.a. nach systematischen Gesichtspunkten entstehen, etwa um Gespräche eines bestimmten Interaktionstyps zu dokumentieren (z.B. „Beratungen") oder um Tendenzen der Sprachentwicklung aufzuzeigen, indem Gespräche nach ihrem Aufzeichnungsdatum chronologisch zusammengestellt werden. Solche Korpora können zu einem Archiv zusammengefügt werden.

Für eine computergestützte Verarbeitung der Gesprächsaufnahmen muss ein (digitales) Sprachsignal vorliegen; die alten analogen Feldaufnahmen müssen also zunächst digitalisiert werden.

Die Analyse authentischer Gespräche setzt ihre Aufzeichnung und ihre detaillierte Verschriftlichung (Transkription) nach einem standardisierten Inventar von Konventionen voraus. Bei der Transkription von Aufnahmen wird das gesprochene Wort mit seinen prosodischen Eigenschaften in einer relativ feinen Weise in einen schriftlichen Text umgesetzt. Die Transkribierweise und die Genauigkeit des Transkribierens sind freilich nicht nur von etablierten Konventionen und der individuellen Erfahrung, sondern vor allem auch von den jeweiligen Forschungsinteressen abhängig. Generell ergeben sich folgende Anforderungen an die Transkription von Gesprächen:

- *Unterscheidung zwischen Sprechtext und Annotationen*: Diese Unterscheidung muss schon bei der Eingabe deutlich sein,[2] denn sie soll bei der Lektüre von Transkripten eine schnelle Orientierung ermöglichen, sie ist aber vor allem im Hinblick auf die spätere automatische Verarbeitung relevant.[3]
- *Literarische Umschrift*: Sie stützt sich grundsätzlich auf das orthografische System der Schriftsprache und ergänzt dieses zur Präzisierung der Lautwiedergabe um eine Reihe von Sonderzeichen. Dadurch soll auch die im Einzelfall besondere dialektale oder umgangssprachliche Artikulation der Sprecher im Transkript wiedergegeben werden.[4]
- Durch das Transkript-Layout sollen interaktive Bezüge (das betrifft eine Zuordnung von Redebeiträgen zu Sprechern, Sprecherwechsel, Überlappungen / Simultanpassagen) auch optisch sinnfällig deutlich werden. Der Editor muss Simultanpassagen / Überlappungen verwalten können. Tran-

---

2 Darum wird der Sprechtext im IDS grundsätzlich in Kleinschreibung, sprecherbezogene oder globale Kommentare werden aber in Großbuchstaben notiert. Ebenso werden Beschreibungen des sprachlichen oder nichtsprachlichen Verhaltens eines Sprechers auf der Sprecherzeile durch Großbuchstaben kenntlich gemacht (z.B. „LACHT").

3 So sind Annotationen, etwa Beschreibungen der Sprechweise oder prosodische Informationen für die Text-Ton-Synchronisation, das sog. *Alignment*, irrelevant und müssen in der Vorverarbeitung aus dem Transkript ausgefiltert werden (vgl. 5.1.). Für eine Datenbank-Recherche, insbesondere in COSMAS II, dürfen Annotationen nicht als Teil des laufenden Sprechtextes gelten, sondern sind gesondert oder in Kombination mit Sprechtext recherchierbare Informationen (vgl. 7.2.).

4 Die Umschrift, wie wir sie im IDS verwenden, orientiert sich grundsätzlich an den Regeln der Standard-Orthografie, verzichtet aber auf die Großschreibung, die normale Interpunktion sowie die Trennung am Zeilenende.

skripte werden im IDS darum grundsätzlich in *Partiturschreibweise* angefertigt, d.h. für jeden am Diskurs beteiligten Sprecher existiert eine eigene Zeile, auf der seine Äußerungen verschriftlicht werden. Die Reihenfolge der Sprecher innerhalb des Partiturblocks bleibt über das gesamte Transkript hinweg konstant.

- Die *prosodische Notation* umfasst Grenzintonationsmuster an Stellen möglicher Redeübergabe (also möglichen Sprecherwechsels), Pausen, Wechsel in der Sprechweise (Lautstärke und Sprechgeschwindigkeit).
- Der Editor muss bequeme Möglichkeiten zur *Korrektur* (insbesondere in Form einer interlinearen Verwaltung mehrerer Sprecher) und *Annotation* haben. Solche Korrekturdurchgänge sind insbesondere bei technisch oder akustisch schlechten oder „turbulenten" Gesprächsaufnahmen notwendig; dabei ist bei einer Ersteingabe oft nur eine „löchrige" Eingabe des Wortlauts möglich; erst in mehreren Korrekturdurchgängen können unverständliche Stellen entschlüsselt und prosodische Eigenschaften notiert werden. Annotationen (etwa Kommentare zur Sprechweise oder für die Interpretation der Stelle notwendige Informationen zu nichtsprachlichen Vorgängen) sollen auf einen bestimmten Teil des Redebeitrags eines Sprechers oder auf einen Abschnitt des gesamten Gesprächsereignisses bezogen werden können.
- Der Editor muss *multilingual* sein, d.h. Sonderzeichen und fremdsprachige Zeichensätze verwalten können. Das betrifft fremdsprachige Aufnahmen, polyglotte Diskurse (bei denen die Beteiligten jeweils in ihrer Muttersprache oder in einer „lingua franca" sprechen, aber anderssprachige Redebeiträge ihrer Gesprächspartner verstehen können) oder Gespräche mit „code-switching", also einem Wechsel zwischen mehreren Sprachen innerhalb von Redebeiträgen.
- Traditionell wird mit analogen Aufnahmen (meist Audiocassetten als Kopien der Original-Aufnahme) und Diktiergeräten transkribiert. Diese Geräte sind robust, aber in der Tonqualität beschränkt. Da für die Weiterverarbeitung analoge Aufnahmen ohnehin digitalisiert werden müssen und zukünftig gleich digital aufgezeichnet werden, ist eine Transkription mit einem in den Transkript-Editor *integrierten Audio-Editor* wünschenswert, der mindestens die Funktionen etwa des Diktiergeräte-Fußschalters bietet (Abspielen, Stopp, Sprung zurück), darüber hinaus aber komfortable weitere Funktionen (nichtsequenzieller Zugriff, also direkter Sprung an eine bestimmte Stelle im Sprachsignal; „Loop"-Funktion, also wiederholtes Abspielen einer bestimmten Stelle).
- Für die Laufzeit eines Projekts sind Aufnahmen und Transkripte vor einem unkontrollierten Zugriff zu schützen. Personenbezogene Daten müssen für Zitate in Publikationen, Textbände und eine Weitergabe der Transkripte *anonymisiert* bzw. *maskiert* werden – das gilt für Personen-, Ortsnamen, Aktenzeichen u.ä.

Mit der Transkription und der Archivierung von Transkripten in Korpora entstehen weitere Anforderungen, nämlich:

- zu einem Transkriptstück den entsprechenden Ausschnitt in der Aufnahme zu finden und anhören zu können,
- alle Vorkommen eines Phänomens in den Transkripten aufzufinden (und parallel dazu die entsprechenden Tonausschnitte hören zu können),
- nicht nur in einem Transkript, sondern im archivierten Korpusbestand oder in passend zur aktuellen Fragestellung zusammengestellten Teilkorpora recherchieren zu können.

Auch an die Korpuspflege sind Anforderungen zu stellen: Die manuell-auditive Transkription von Gesprächen ist so aufwändig, dass sie auch nach Abschluss des Projekts, in dem Aufnahme und Transkript erstellt wurden, für Recherchen zur Verfügung stehen soll. Dazu müssen Aufnahmen und Transkripte gut dokumentiert und Transkripte bzw. Korpora von Transkripten, die nach Dokumentationsgesichtspunkten neu zusammengestellt wurden, recherchierbar sein.

Transkripte auf Papier informieren über viele signifikante Vorgänge mit ikonischen Zeichen oder mittels des Transkript-Layouts; diese Informationen sind nur im Zusammenhang mit einer Legende verständlich und trotz nachhaltig verfolgter Initiativen (vgl. Selting *et al.* 1998) in der gesprächsanalytischen Forschung noch nicht standardisiert.[5] Wenn unterschiedliche Korpora, auch aus verschiedenen Forschungseinrichtungen, in einem Datenbank-Archiv zusammengeführt werden, müssen die Transkripte in einem standardisierten Austauschformat vorliegen, das über bestimmte Transkriptionskonventionen, die Idiosynkrasien von Transkript-Editoren oder Betriebssysteme / Computer-Plattformen hinaus eine einheitliche Recherche ermöglicht.

Der Zugriff auf Transkripte, Korpusbestandteile und Korpora über das Erstellungsprojekt hinaus muss kontrolliert werden: Haben die Gesprächsbeteiligten ihre Zustimmung zur Aufnahme, wissenschaftlichen Analyse und Publikation gegeben, sind die Aufnahmen anonymisiert und Transkripte bei personenbezogenen Daten maskiert, sind die Diskurse mithin freigegeben?

Für Recherchetreffer soll schließlich ein bequemer Zugriff auf die entsprechende Stelle in der Audioaufnahme möglich sein.

---

5 Beispielsweise wird in DIDA eine steigende oder fallende Grenzintonation an Stellen möglicher Redeübergabe mit Pfeil-Zeichen (↑ bzw. ↓) notiert; in GAT wird diese „letzte Tonhöhenbewegung vor dem Einheitenende" als „mittel steigend" mit Komma und als „mittel fallend" mit Semikolon notiert. Bei diesen unterschiedlichen Darstellungen durch Zeichen auf der Textoberfläche der Transkripte dient eine explizite Kodierung in SGML (*Standard Generalized Markup Language*) als notations- und systemübergreifendes Austauschformat, in das die Transkript-„Dialekte" überführt und das im Sinne der unterschiedlichen Notationskonventionen interpretiert werden kann. In unseren SGML-Transkripten werden die steigende Intonation mit dem Tag „<shift feature=intonation here=steigend>" und die fallende Intonation mit dem Tag „<shift feature=intonation here=fallend>" kodiert.

2.2. Anreicherung von Korpora

Diese Anforderung besagt, dass Informationen nicht manuell oder durch auditive Analyse erhoben werden, sondern computergestützt durch Korpuswerkzeuge. Solche Werkzeuge sind insbesondere:

- *Text-Ton-Alignment* als Synchronisation von Gesprächsaufnahme und Transkript; diese Synchronisation muss bei großen Korpora computergestützt-automatisch durchgeführt werden; die Genauigkeit ist vom Verwendungszweck des Ergebnisses abhängig.
- *Werkzeuge für Phonetik und Prosodie*: Ausschnitte aus einem Sprachsignal, also der digitalen Version einer Gesprächsaufnahme, sollen mit einem Computerprogramm zur automatischen Phonetik- und Prosodie-Analyse bearbeitet werden können, um intersubjektiv nachprüfbar Grundfrequenzverläufe (F0-Kurven), Intensitäten und – vor allem für dialektologisch-phonetische Untersuchungen – Spektrogramme zu erstellen.
- *Varietäten / Aussprachevarianten* und *Lemmatisierung*: Für die Recherche in Gesprächstranskripten, insbesondere mit Datenbanken, stellt sich ein Dilemma: Je genauer die Artikulation im Einzelfall als Abweichung von der standardsprachlichen Orthografie notiert wurde, desto brauchbarer ist das Transkript, um Muster für umgangssprachliche Formulierungen zu belegen, desto zufälliger wird aber auch das Ergebnis der Datenbankrecherche. Anfragen müssten alle denkbaren Verschriftungsformen angeben, die man aber vor der Recherche nicht kennt. Für Transkripte gesprochener Sprache ist eine Standardisierung der Schreibweise wie bei der schriftsprachlichen Orthografie nahezu ausgeschlossen. Durch Konzepte wie „literarische Umschrift" und „Nichtberücksichtigung von Allophonen" (z.B. Auslautverhärtung vs. Palatalisierung, etwa bei „König") und durch die standardmäßige Mehrfachkorrektur von Transkripten durch unterschiedliche Personen wird zwar versucht, die Transkribierweise über Idiosynkrasien und projektspezifische Präferenzen hinaus stabil und einheitlich zu gestalten. Dennoch ist eine Transkription immer von vordefinierten Forschungsinteressen abhängig; z.B. kann dialektale Varianz unterschiedlich genau notiert werden. Für das *Retrieval* im Rahmen von Datenbank-Anfragen stellt diese Streuung bei der Transkribierweise ein Problem dar: Man kann nur dann nach Standard- oder Grundformen suchen, wenn in der Datenbank die lokale Realisierung (*token*) mit einem Lemma (*type*) verknüpft ist, wenn man also nach allen soziolektalen oder dialektalen Varianten eines Worts und darüber hinaus nach allen Flexionsformen mit einer einfachen Anfrage suchen kann, ohne diese Varianten und Flexionsformen kennen und in der Anfrage aufzählen zu müssen. Der Ausweg ist eine Lemmatisierung, also eine regelbasierte automatische Zuordnung aller Flexionsformen, umgangssprachlichen und dialektalen Varianten zu einer Grundform, einem „Lemma".

Von diesen Werkzeugen sind die beiden erstgenannten im Wesentlichen realisiert, werden allerdings z.Z. noch optimiert. Das letztgenannte ist in der konkreten Projektplanung.

3. Die Arbeit mit Gesprächsaufnahmen und -korpora im IDS

Die Gesprächskorpora am IDS umfassen u.a. Gespräche aus mehreren Kommunikationsdomänen, teils aus institutionellen Zusammenhängen, nämlich

- Beratungsgespräche unterschiedlicher Art;
- Schlichtungsgespräche, z.B. aus der außergerichtlichen Schlichtung von Nachbarschaftsstreitigkeiten vor einer Vergleichsbehörde oder aus Verfahren, mit denen Verbraucherreklamationen vor Schlichtungsstellen von Handwerkskammern behandelt werden;
- Aufnahmen aus dem Projekt „Stadtsprache Mannheim", in dem Formen sowie kommunikative und soziale Funktionen des „Monnemerischen" untersucht wurden;
- Gespräche im Fernsehen, also Talkshows und Diskussionen als medial inszenierte Diskurse;
- Ethnografische Interviews mit Fernsehredakteuren und -moderatoren;
- Interviews mit Beamten der europäischen Institutionen in Brüssel, den sog. „Eurokraten".

Diese natürlich noch nicht für alle Kommunikationsbereiche des gesprochenen Deutsch repräsentativen Korpora werden fortlaufend erweitert. Eine vollständige Auflistung des gegenwärtigen (z.T. auch für Servicezwecke zur Verfügung stehenden) Bestands findet man auf dem WWW-Server des IDS.[6]

Die auf der folgenden Seite wiedergegebene Abb. 1 stellt in Grundzügen den Ablauf dar, wie am IDS Gesprächsaufnahmen verarbeitet werden, so dass sie schließlich über eine Datenbank recherchierbar sind. Im Prinzip sind diese Arbeitsschritte für jedes Gesprächskorpus notwendig, gleichgültig in welcher Zusammensetzung und in welcher Notationsform, wenn man ein Text-Ton-*Alignment* durchführen und die Recherche in einer Datenbank ermöglichen will.

Endziel der Bearbeitung von Gesprächsaufnahmen ist die Recherche in der Gesprächsdatenbank mit COSMAS II. Dabei sollen Treffer sowohl im Transkript angezeigt als auch im Ton abgespielt werden können. Um dieses Ziel zu erreichen, sind folgende Arbeitsschritte (in Abb. 1 dunkelgrau unterlegt) notwendig, die Zwischenprodukte (hellgrau unterlegt) ergeben:

---

6 Weitere Informationen unter <http://www.ids-mannheim.de/dsav/> bei den Links „Korpora" und „Datenbankrecherche".

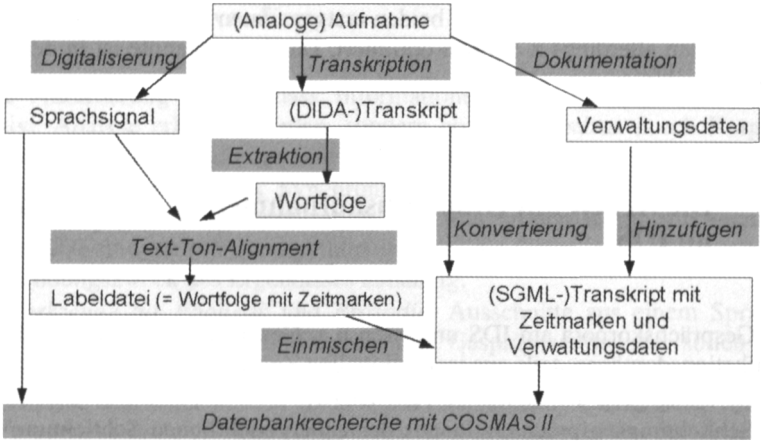

Abb. 1. Verarbeitung von Gesprächskorpora von der Aufnahme zur Datenbankrecherche

- *(Analoge) Aufnahme*: Die Grundlage bilden Aufnahmen von natürlichen Gesprächen, die – zumindest traditionell – auf Audio-, gelegentlich auch auf Videocassetten aufgenommen wurden. Natürlich ist die technische Qualität dieser Aufnahmen eingeschränkt und nicht mit Studioqualität zu vergleichen: Die älteren Aufnahmen sind zumeist nur in mono (was beim Transkribieren die Ortung und damit die Zuordnung von Redebeiträgen zu Sprechern erschwert), wurden mit möglichst kleinen Aufnahmegeräten erstellt (im Bestreben, die Gesprächssituation möglichst wenig zu stören) und weisen Hintergrundgeräusche auf.
- *Transkription*: Gespräche bzw. ihre Tonaufzeichnung sind für die empirische Gesprächsforschung die primären Daten; Untersuchungen dieser Gespräche ließen sich – wiederum traditionell gesehen – aber nur durchführen und intersubjektiv nachprüfbar machen, indem die Gespräche nach einem standardisierten Transkriptionssystem verschriftet worden waren. Am IDS wurden dazu ein Editor namens DIDA (= *DI*skurs-*DA*tenverarbeitung), mit dem man Gespräche in Partiturschreibweise aufzeichnen kann, und ein zugehöriges Notationssystem entwickelt, dessen Merkmale in Abschnitt 4 beschrieben werden.
- „*Dokumentation*": Zu den Gesprächen, den Aufnahmen und den Transkripten wurden Verwaltungsdaten erhoben, die z.B. auf Transkript-Deckblättern festgehalten werden: Aufnahmedaten, Datenträger, Sprecher (und zugehörige Siglen in Transkripten), Besonderheiten, Kommunikationsverlauf.
- Für eine computergestützte Verarbeitung der Gesprächsaufnahmen muss ein (*digitales*) *Sprachsignal* vorliegen; die alten analogen Feldaufnahmen müssen also zunächst digitalisiert werden. Dabei greifen wir aus archivarischen Gründen zu relativ großzügigen Parametern (WAV-Dateien mit

48 kHz Abtastrate in 16bit-Quantisierung, je nach analoger Vorlage mono oder stereo; die nach dem Nyquist-Theorem[7] dadurch erfassbaren Sprachsignale bis 24 kHz gehen zwar über die Obergrenze des durchschnittlichen menschlichen Gehörs von je nach Alter 15-20 kHz und auch über das, was mit älteren analogen Aufnahmegeräten aufgezeichnet werden konnte, hinaus, wir wollen aber auch für prosodische Analysen das Signal nicht beschneiden und verzichten darum auch auf Kompressionsformate wie etwa MP3).

- Die Volltextdatenbank COSMAS II greift auf dieses Sprachsignal und auf die Transkripte zurück, allerdings müssen die Transkripte dazu in ein besonderes Format konvertiert werden, nämlich in *TEI-konformes SGML* (TEI = *Text Encoding Initiative*[8]). In diesem Format werden alle Informationen, die das Transkript enthält, explizit kodiert. So können Transkripte aus einem beliebigen proprietären Format über Betriebssystem- bzw. Plattformgrenzen und einzelne Anwendungen, insbesondere Transkript-Editoren, hinaus in ein weit verbreitetes und auch zukunftssicheres Austauschformat gebracht werden. Das SGML-Transkript soll auch zu jedem Wort die Zeitmarken und zum Zweck einer Korpusauswahl zum gesamten Transkript die Verwaltungsdaten enthalten.
- Um die Zeitmarken zu gewinnen, muss nun ein *Text-Ton-Alignment* durchgeführt werden; Input für dieses Verfahren sind eine aus dem komplexen Transkript extrahierte einfache Wortfolge und das Sprachsignal (dafür ist aus Speicherplatz- und Performance-Gründen eine auf eine Abtastrate von 16 kHz reduzierte „Arbeitskopie" des Sprachsignals sinnvoll). Das Text-Ton-*Alignment* ergibt eine sogenannte „Labeldatei": eine um Zeitmarken angereicherte Wortfolge.
- Die *Konvertierung* besorgt ein im IDS entwickeltes Programm, das in einem Arbeitsschritt drei Aufgaben erledigt:
  a. Das DIDA-Transkript wird aus einem proprietären Format in TEI-konformes SGML konvertiert;
  b. die aus dem Alignment gewonnenen Zeitmarken werden den Wörtern des Transkripts zugeordnet, sozusagen in die SGML-Datei „eingemischt";
  c. die Verwaltungsdaten werden hinzugefügt.

Im Folgenden sollen zwei Instrumente vorgestellt werden, die für die Arbeit mit Gesprächskorpora unentbehrlich erscheinen. Es handelt sich dabei um

- DIDA, einem Werkzeug zur Erfassung, Pflege und Archivierung von Transkripten und
- SPRAT (*Speech Recognition and Alignment Tool*), einem Programm zur automatischen Text-Ton-Synchronisation und Spracherkennung.

---

7 „Das Nyquist-Theorem besagt, daß die Frequenz des zu digitalisierenden Signals höchstens halb so groß sein kann wie die Samplerate" (<http://www.speechdat.org/multi_hyper/MMAudio.html>). Zum Nyquist-Theorem vgl. auch <http://www.digital-recordings.com/publ/pubneq.html>.

8 Vgl. <http://www.tei-c.org>.

4. DIDA

DIDA dient der Verschriftlichung gesprochener Sprache und dem Darstellen von Kommunikation in Gesprächen. Dabei wird neben der Darstellung der Diskursteilnehmer und deren Äußerungen auch die Annotation weiterer kommunikativer Dimensionen ermöglicht.
Das DIDA-System besteht aus den Komponenten

- Partitureditor
- Audioeditor (optional)
- Projektdatenbank (optional)
- Exportprogramm
- Druckprogramm
- Netzkommunikation (optional)

Der Partitureditor spielt dabei eine zentrale Rolle, da von ihm aus alle anderen Komponenten angesprochen werden. Im Gegensatz zu einem herkömmlichen Textverarbeitungssystem erlaubt der Editor die Handhabung von Quasi-Endlos-Zeilen. Dabei wird jedem Sprecher eine Zeile zugeordnet, die die Äußerungen des Sprechers beinhaltet. Jeder dieser sogenannten Sprecherzeilen können beliebig viele kommunikative Dimensionen (kurz Kommentarzeilen genannt) hinzugefügt werden. In sie können z.B. Angaben zum Verhalten des Sprechers oder die Übersetzung des Gesprochenen aufgenommen werden. Die situative Beschreibung eines Diskurses geschieht in einer separaten Zeile, die keinem Sprecher zugeordnet ist. Die kommunikativen Dimensionen werden durch Markierung eines Referenzbereichs mit den transkribierten Daten synchronisiert, wobei Länge und Position eines Referenzbereichs individuell verändert werden können. Bei den Aktionen *Einfügen* und *Löschen* kann die Synchronität sowohl der Sprecher- wie der Kommentarzeilen wahlweise erhalten bleiben, was in der Regel wünschenswert ist. Darüber hinaus können fremdsprachliche Sonderzeichen (westeuropäisch, türkisch, kyrillisch, IPA) dargestellt werden.

Die Arbeit mit dem Audioeditor kann vom Partitureditor wahlweise über Funktionstasten oder Untermenüs gesteuert werden und gestattet beliebig genaue Positionierung, Parametrisierung der relativen Sprungintervalle sowie eine Parametrisierung der Wiedergabeintervalle. Damit kann das Arbeiten mit dem Tonmaterial individuell vom Transkribenten angepasst werden. Der Zeitaufwand für Suchen und Positionieren wird dadurch wesentlich reduziert.

Die *Online*-Version von DIDA arbeitet in einem heterogenen UNIX-Netzwerk. Die Daten werden zentral auf einem Server verwaltet. Der Zugriff auf sie erfolgt über eine Oracle-Datenbank, die die Art der Zugriffsberechtigung überprüft. Dazu werden Projekte definiert und Kennungen bestimmten Projekten zugeordnet. Einzelnen Kennungen kann ein Privileg vergeben werden (sogenannte Projektverwaltungskennungen), das erlaubt, alle zum Projekt gehörigen Transkripte bearbeiten zu können. Die übrigen Kennungen können nur die von ihnen erstellten Transkripte bearbeiten. Über Projektgrenzen hin-

weg können Transkripte nur durch eine Freigabe zum Lesen zur Verfügung gestellt werden. Mittels einer Markierung „entliehen" können Transkripte vollständig gesperrt werden, um z.B. *offline* daran weiter arbeiten zu können.

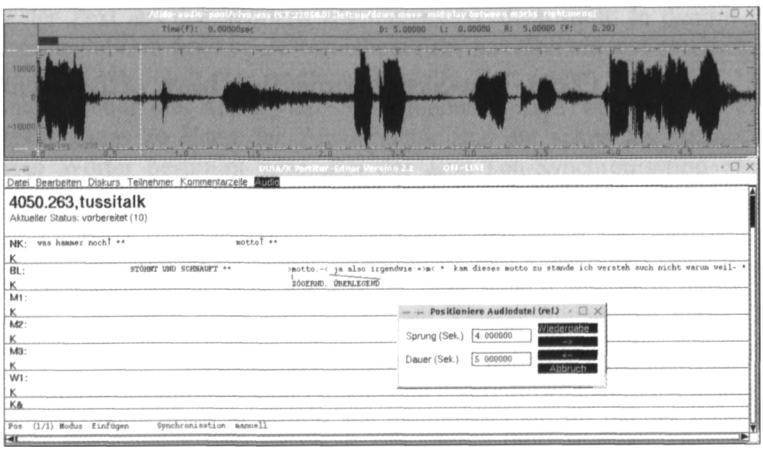

Abb. 2. Beispiel für die Darstellungsweise des DIDA-Partitureditors mit Audioeditor[9]

Bei der *Offline*-Version werden die Transkripte in einem bestimmten Verzeichnis abgelegt. Auf alle dort vorhandenen Transkripte kann zugegriffen werden. Lediglich beim Erstellen eines neuen Transkripts wird überprüft, ob schon ein gleichnamiges vorhanden ist, damit kein versehentliches Überschreiben stattfinden kann.

Um ein Transkript mit einem Textverarbeitungssystem weiterverarbeiten zu können, gibt es in DIDA eine Exportkomponente. Sie bereitet die in horizontaler Richtung Quasi-Endlos-Schreibweise in eine blattformatierte Schreibweise um und gibt sie als RTF-Datei aus. Optionen für verschiedene Schriftgrößen, Papierformat und weichen Umbruch, Zeilenlineal und -nummerierung bieten typografische Gestaltungsmöglichkeiten.

Die DIDA-Druckkomponente erzeugt eine um 90° gedrehte Endlosausgabe, die größtenteils der Bildschirmdarstellung entspricht und daher großen Papierbedarf hat. Auch hier gibt es Optionen für verschiedene Schriftgrößen, Papierformat und weichen Umbruch. Die Ausgabe erfolgt in Form einer Postscript-Datei, die ohne weitere Aufbereitung gedruckt werden kann.

---

9 Dargestellt ist der Anfang des Transkripts 4050.263,a.tussitalk (Ausschnitt aus der Sendung „Viva interaktiv" vom 21.4.1999). Das vollständige Transkript mit einer Liste der Transkriptionszeichen ist auf der beiliegenden CD-ROM unter BODMER_FACH_SCHMIDT_SCHUETTE\ANNEXES\ANNEX_1.PDF, das Video dazu unter BODMER_FACH_SCHMIDT_SCHUETTE\VIDEO\TUSSITALK.MPG zu finden.

Für die *Online*-Version wird eine verteilte Kommunikationskomponente benötigt, die die *Client-Server*-Kommunikation innerhalb eines heterogenen UNIX-Netzes (z.Z. Solaris, Irix, DG/UX und Linux) ermöglicht.

5. SPRAT (Text-Ton-Synchronisation bzw. *-Alignment*)

Die synchrone Verbindung von Text und Ton lässt sich computertechnisch mit Hilfe eines leicht modifizierten Spracherkenners (sogenannter *Aligner*) realisieren, wie er z.B. am IDS im Rahmen eines Projekts mit der Bundeswehr (SERGES) entwickelt worden ist. Grob skizziert besteht der *Aligner* aus folgenden Komponenten:

1. einem Inventar von Phonemen in Form von HMMs (*Hidden Markov*-Modellen, d.h. speziellen stochastischen Automaten),
2. einem Aussprachewörterbuch und einem Phonetisierungsprogramm,
3. einem Modell für die Darstellung der Wörter in Form einer Grammatik bzw. eines Netzwerks,
4. einem Mustererkennungsverfahren in Kombination mit einer Signalverarbeitungskomponente.

5.1. Architektur

Zu (1.): Bevor man einen Synchronisations- oder Erkennungsprozess starten kann, müssen die HMMs trainiert worden sein. Ein Training ist in der Regel eine zeitintensive Angelegenheit und nur zwingend bei der Erstellung oder der Verbesserung eines *Aligners*. Jedes HMM besteht aus miteinander verbundenen Zuständen, von denen von jedem mit einer bestimmten Übergangswahrscheinlichkeit in einen anderen oder denselben Zustand übergegangen werden kann. Zusätzlich wird in jedem dieser Zustände ein Symbol emittiert. Welches Symbol das sein kann, wird mit Hilfe von Emissionswahrscheinlichkeiten festgelegt, die für jeden Zustand verschieden sein können. Diese Symbole entsprechen Prototypen von Kurzzeitspektren, die zuvor mittels FFT (*Fast Fourier Transformation*) (und eventuell Vektorquantisierung) bestimmt worden sind. Die Topologie der HMMs wird durch den Entwickler festgelegt und die Wahrscheinlichkeiten werden beim Trainingslauf berechnet.

Zu (2.): Ein Aussprachewörterbuch hat den Vorteil, die Qualität der phonetischen Transkription in jedem einzelnen Fall überprüfen und modifizieren zu können. Das führt in der Regel zu einer akkurateren phonetischen Transkription, als sie ein maschinelles regelbasiertes Verfahren (Phonetisierer) liefert. Die Nachteile sind die zeit- und kostenintensive Erstellung und der begrenzte Wortschatz eines Wörterbuchs. Es liegt daher nahe, nur solche

Wortformen in ein Wörterbuch aufzunehmen, deren Aussprache unregelmäßig ist, und die restlichen mit Hilfe des Phonetisierungsprogramms zu erzeugen.

Zu (3.): Mit Hilfe der Grammatikkomponente wird die Regel, nach der sich der Text aus den Wörtern zusammensetzt, dargestellt. In diesem Punkt unterscheidet sich der *Aligner* von einem Erkenner. Bei einem Spracherkenner wird der Text durch eine endliche Schleife über ($wort_1$ | $wort_2$ | ... $wort_n$) erzeugt, wobei die Anzahl der Schleifendurchgänge während des Erkennungsvorgangs an Hand der Tonaufnahme und der schon als erkannt markierten Wörter bestimmt wird. Bei der Text-Ton-Synchronisation gestaltet sich dieser Punkt einfacher. Bei Vorliegen eines Textes von *m* Wörtern erfolgt ein einmaliger Durchlauf über ($wort_1$ $wort_2$ ... $wort_m$). Bei manchen Verschriftlichungen kann es jedoch vorkommen, dass Pausen und flüchtig geäußerte Hesitationsphänomene wie z.B. *äh*, *ähm*, *mhm* u. dgl. nicht transkribiert wurden. Um durch diese unvollständigen Verschriftlichungen keine Ungenauigkeit bei der Synchronisation zu verursachen, kann die Regel dahin gehend erweitert werden, dass sie zwischen den einzelnen Wörtern ein automatisches Einfügen von Pausen und Hesitationsphänomenen zulässt. Insbesondere das automatische Einfügen von Pausen führt zu einer Verbesserung der Synchronisation. Die Tatsache, dass Hesitationsphänomene gelegentlich mit Wortbestandteilen korrelieren, führt dazu, dass ihre Berücksichtigung meist nicht von Vorteil ist.

Zu (4.): Da die in Punkt 4 genannte Mustererkennungskomponente sowohl sprachen- als auch anwendungsunabhängig ist, soll hier nicht näher darauf eingegangen werden. In SPRAT werden dazu Standardprogramme aus dem HTK (*Hidden Markov Toolkit*) der Firma Entropics eingesetzt.

5. 2. Ergebnisse

Das Ergebnis einer Text-Ton-Sychronisation besteht darin, dass den einzelnen Wörtern (oder Phonemen) Zeitmarken zugeordnet werden, die auf den Zeitpunkt des Auftretens in der Audiodatei verweisen. Durch Anfangs- und Endzeitpunkt bzw. Anfangszeitpunkt und Dauer des Wortes wird bei gegebener Zuordnung von Text- und Audiodatei die Synchronisation festgelegt. Um sich davon zu überzeugen, bedarf es Werkzeuge, die alle gegebenen Informationen miteinander verknüpfen. Am IDS Mannheim gibt es dazu das Volltextbanksystem COSMAS II, das komplexe Rechercheanfragen erlaubt. Einfachere Werkzeuge, mit denen wahlweise Text und Ton manuell synchronisiert oder die Zuordnung von Text und Ton angezeigt werden können, sind z.B. die Audioeditoren *xwaves* (Fa. Entropics) und *Praat* (Universität Amsterdam, <http://www.praat.org>). Diese können mit Hilfe von Programmen so gesteuert werden, dass einfache Suchanfragen mit Anhören der Ergebnisse möglich sind.

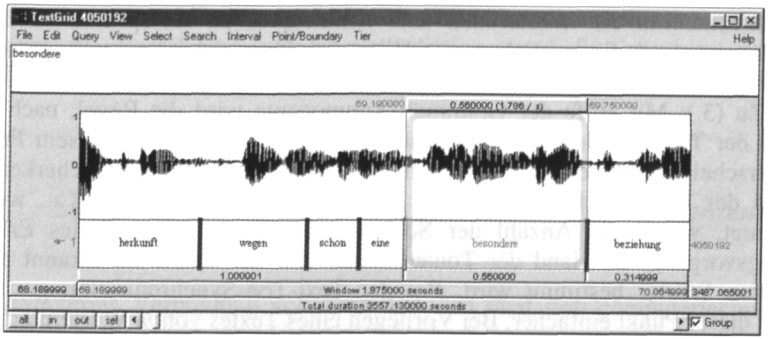

Abb. 3. Darstellung von Ergebnissen des *Aligners* mit Hilfe von *Praat*

Was die Qualität der Synchronisation betrifft, so ist diese nicht nur von den o.a. Komponenten eines *Aligners* abhängig, sondern auch von der Aufnahmequalität der Tonaufnahme und der Genauigkeit der Transkription. Starkes Rauschen, hallige Aufnahmen, leises undeutliches (verschliffenes) Sprechen wirken ebenso qualitätsmindernd wie Simultanpassagen (mehrere Sprecher gleichzeitig). Letzteres wirkt sich weniger stark aus, wenn für den betreffenden Zeitabschnitt nur der akustisch dominante Gesprächsteilnehmer transkribiert wurde (vgl. Schmidt / Neumann 1999).

Die Qualität der Synchronisation von mundartsprachlichen Aufnahmen ist meist auch dann noch akzeptabel, wenn sie in Standardsprache transkribiert wurden. Ausnahmen bestehen an den Stellen, an denen es Abweichungen von mehreren Silben gibt.

Eine Weiterentwicklung des oben geschilderten Lösungsansatzes zur Bereitstellung von mit Ton synchronisierten Texten geschieht derzeit in den Richtungen

- Robustifizierung des *Aligners* durch Training von nicht-sprachlichem Material (Musik, Klatschen, Lachen, Husten, Rauschen u.ä.),
- Behandlung von Simultanpassagen,
- Synchronisation von fremdsprachlichem Material,
- Erkennung von Aussprachevarianten.

6. Automatische Segmentierung nicht-sprachlicher Signalsequenzen in Korpora gesprochener Sprache

Die im Folgenden vorgestellten Arbeiten und Ergebnisse sind im Rahmen eines Kooperationsprojekts zwischen dem Institut für Deutsche Sprache (IDS) in Mannheim und dem Institut für Maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart entstanden. Das Ziel dieses Projekts lag in der

Verbesserung einer automatischen Segmentierung von Sprachsignalen, um eine maschinelle Repräsentation von Sprachsignalen zu ermöglichen und zu optimieren.[10]

Unter automatischer Segmentierung verstehen wir eine Ton-Text-Kopplung der Art, dass eine text-basierte Beschreibung des Sprachsignals entsteht, in welcher jede linguistische Einheit auf ein Textobjekt mit Zeitreferenzen zum Sprachsignal abgebildet wird. Es entsteht somit eine Einteilung des Sprachsignals nach bestimmten linguistischen Einheiten (z.B. Phonemen, Silben oder Wörtern). Dabei entspricht der Ton dem aufgenommen Sprachsignal und der Text einer Verschriftlichung dessen, was im Sprachsignal gesagt wurde. Somit ist die automatische Segmentierung eine Anreicherung der Verschriftlichung mit Zeitmarken aus dem Sprachsignal.

6.1. Automatische Segmentierung

6.1.1. Das Sprachsignal und dessen orthografische Beschreibung

Gesprächs- und Sprachaufnahmen sind die primären Daten für die empirische linguistische Forschung. Untersuchungen und Analysen solcher Aufnahmen sind aber erst dann systematisch durchführbar, wenn Daten nach standardisierten Kriterien konsistent verschriftlicht und segmentiert sind. Eine Verschriftlichung erfolgt manuell durch Anhören und Niederschreiben des aufgenommenen Sprachsignals. In Abb. 4 ist dieser Zusammenhang schematisch dargestellt.

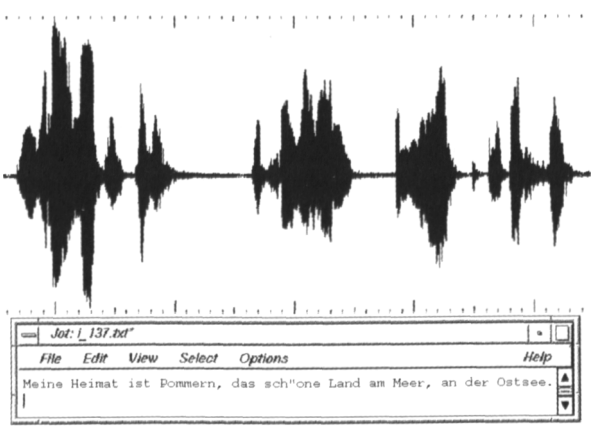

Abb. 4. Sprachsignal und Verschriftlichung

10 Das Kooperationsprojekt „Maschinelle Segmentierung von Korpora gesprochener Sprache des IDS (*Alignment*-Projekt)" wurde vom Land Baden-Württemberg finanziert und hatte eine Laufzeit von 1998 bis 2001.

Im oberen Bereich der Abb. 4 ist das aufgenommene Sprachsignal in einer Zeit-Frequenz-Darstellung gezeigt. Das darunterliegende „Fenster" zeigt den niedergeschriebenen Inhalt des Sprachsignals. Bei dieser Art der Verschriftlichung handelt es sich um die einfachste Art einer Annotation von Sprachsignalen: Es gibt keine Zeitmarkierung (Zeitreferenzen) zwischen den Wörtern der Verschriftlichung und dem Sprachsignal und es gibt über die Wörter hinaus keine weiteren Beschreibungen.

6.1.2. Segmentierung von Sprachsignalen

Mussten Forscher früher aufwändig und langwierig Tonbänder sequenziell abspielen und durchsuchen, um Analysen durchführen zu können und Belege für bestimmte Phänomene zu finden, so ist es heute dank moderner Spracherkennungstechnologie und Signalverarbeitungsmethoden möglich, Sprachsignal und Text so geeignet zu verbinden, dass Analysen auf Textebene und somit nicht-sequenziell durchgeführt werden können und durch Segmentierung jederzeit der direkte Zugriff auf das Sprachsignal möglich ist. Solche Segmentierungen können manuell und automatisch erstellt werden. Manuelle Segmentierverfahren werden in der Regel *Software*-basiert vorgenommen.[11] Die Bereitstellung von Signaleditor (Audioeditor) und synchronisiertem Texteditor ermöglicht dabei, die Segmentgrenzen linguistischer Einheiten von Hand einzutragen.

Bei den maschinellen Verfahren unterscheidet man zwischen verschiedenen Techniken von Spracherkennung, Merkmalserkennung und anderen mehr.[12] Vertreter von Segmentierung durch Spracherkennung wie etwa S. Greenberg (Greenberg 2000) postulieren hybride Systeme zur Klassifikation von merkmalsbasierten Systemen, die keine orthografische Repräsentation (Verschriftlichung) benötigten. Der Kern des Systems besteht aus neuronalen Netzen (zur Klassifizierung von artikulatorisch-akustischen Eigenschaften); die Abbildung der phonetischen Eigenschaften auf tatsächliche Phoneme und die Phonemsegmentgrenzen werden durch *Hidden Markov*-Modelle (welche als Basistechnologie der meisten herkömmlichen automatischen Systeme dienen) realisiert. Der Vorteil solcher Systeme liegt darin, dass keine zeitaufwändigen Verschriftlichungen benötigen werden, der Nachteil liegt in geringerer Genauigkeit der Segmentierung. Automatische Systeme wie MAUS (Kipp 1995) und ALPHONES (Rapp 1995; 1998) bieten dagegen eine sehr hohe Genauigkeit der Segmentierung, allerdings mit dem Nachteil, dass Verschriftlichungen benötigt werden. Beide Systeme werden durch spezielle Spracherkennungsverfahren realisiert. Fach (2000) bietet eine

---

11 Folgende Softwarepakete ermöglichen eine manuelle Segmentierung: die schon erwähnten Programme *Praat* und *xwaves* sowie *Transcriber* (<http://www.etca.fr/CTA/gip/Projets/Transcriber/>), wobei dies natürlich nur eine kleine Auswahl der verfügbaren Software-Werkzeuge ist.

12 Der interessierte Leser sei für weitergehende Informationen auf die Dissertation von Paul Hosom (2000) verwiesen, in der 32 verschiedene Verfahren zur automatischen Segmentierung verglichen und evaluiert werden.

detaillierte Diskussion des Zusammenhangs zwischen Spracherkennung und Segmentierung.

Für die hier vorgestellten Experimente ist das automatische Segmentierungssystem ALPHONES eingesetzt worden, dessen Architektur und Funktion im folgenden Abschnitt skizziert wird.

6.1.3. Das Stuttgarter automatische Segmentierungssystem ALPHONES

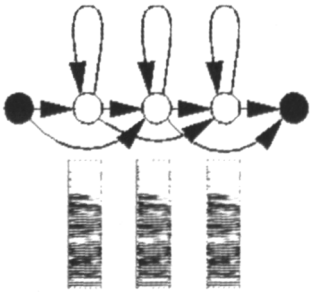

Abb. 5. Sprachsignal und Modell

Das automatische Segmentierungssystem ALPHONES (= *Al*ign *phones*) erzeugt Segmentierungen auf Phonem-, Silben- und Wortebene. Das System besteht aus einer Trainingsphase und einer Segmentierungsphase. Die Trainingsphase ist Teil der Entwicklung des Systems und mit der Fertigstellung des Systems abgeschlossen. In der Trainingsphase wird für jedes Symbol des Sprachinventars (z.B. Phonem einer Sprache) ein Modell angelegt und anhand der vorkommenden Muster trainiert. Das Ergebnis dieser Trainingsphase ist eine Modellmenge, in der die vorkommenden Modelle die spezifischen Eigenheiten (z.B. spektrale Eigenschaften) der jeweiligen Symbole der Sprache repräsentieren. In Abb. 5 ist die Zuordnung von Sprachsignal und Modell schematisch dargestellt. In der oberen Hälfte der Abbildung ist ein allgemeines Modell und in der unteren Bildhälfte sind Ausschnitte aus dem entsprechenden Sprachsignal gezeigt. Diese Ausschnitte sind sogenannte Vektoren, die das Sprachsignal über kurze Abstände hinweg darstellen. Jeder Zustand des Modells wird mit einem solchen Vektor in Zusammenhang gebracht.

Die Segmentierungsphase ist der normale Betrieb des Systems und durch eine Dreiteilung gekennzeichnet: die Textvorverarbeitung, die Sprachsignalvorverarbeitung und die Segmentierung. In Abb. 6 auf der folgenden Seite sind diese Funktionen durch die Architektur des Systems dargestellt.

Die Textvorverarbeitung besteht aus zwei Stufen. Die erste Stufe ist die Umsetzung von graphemischer in phonemische Repräsentation (Graphem-Phonem-Konversion). Die Graphem-Phonem-Konversion wiederum besteht aus zwei Lexikonzugriffen (Ausnahmelexikon und der deutsche Teil des CELEX-Lexikons) und einer regelbasierten Umsetzungsstufe für weitere im Eingabestrom verbliebene, nicht ersetzte Grapheme. Die zweite Stufe der Textvorverarbeitung überführt die phonemische Repräsentation in eine systeminterne Repräsentation („Grammatik"). Diese systeminterne Repräsentation ist eine Verkettung phonemischer Symbole, die durch ihre trainierten Modelle ersetzt wurden.

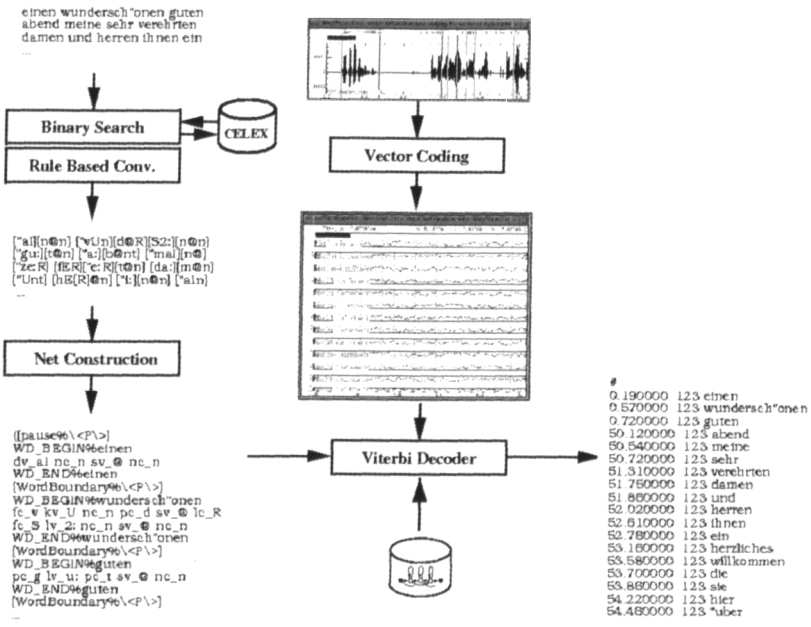

Abb. 6. Die Architektur von ALPHONES (nach Rapp 1995)

Das Sprachsignal unterliegt ebenfalls einer Vorverarbeitung; hier gelten die gleichen Prinzipien wie bei der automatischen Spracherkennung. Da die „waveform"-Darstellung des Sprachsignals nicht genug Informationen liefert, um akustische Ereignisse distinktiv beschreiben zu können, wird das Sprachsignal in eine andere Repräsentationsform überführt.

Die letztendliche Segmentierung erfolgt dann zwischen der Grammatik und dem parametrisierten Sprachsignal, in Abb. 6 durch die Box „Viterbi Decoder" dargestellt. Details zu diesem Verfahren können in Rapp (1995; 1998) nachgelesen werden.

6.1.4. Probleme bei der automatischen Segmentierung

Die Qualität automatischer Segmentierung ist abhängig von verschiedenen Faktoren. Ursachen für eine schlechte Segmentierungsqualität können sein: Art und Qualität der aufgenommen Sprachdaten (Einzelsprecheraufnahmen vs. Mehrsprecheraufnahmen und daraus resultierende Simultansprechsituationen, Dialektaufnahmen; Aufnahmen im Freien vs. Studioaufnahmen), nichtsprachliche Phänomene im Sprachsignal wie Lachen, Klatschen und Musik sowie spontan-sprachliche und dialog-sprachliche Phänomene wie *äh*, *ähm* und *mhm*. Während bei Problemen der Sprachsignalqualität oft nur eingeschränkte Optimierungsmöglichkeiten zur Verfügung stehen, gibt es bei Problemen auf Segmentebene Möglichkeiten zur Optimierung.

Die meisten Segmentierungsfehler von sprachlichen Einheiten entstehen im Zusammenhang mit der Textvorverarbeitung. Wie in Abschnitt 6.1.3. beschrieben, wird der Eingabetext in eine Kette von Einzelphonemen überführt, um damit die Segmentierung durchzuführen. Es können aber nur solche Signalabschnitte segmentiert werden, die ein akustisches Korrelat im Sprachsignal haben. Es entstehen Fehler, wenn Elemente in der Kette sind, denen kein akustisches Korrelat im Sprachsignal zugeordnet ist, und *vice versa*. Lange Signalabschnitte, die z.B. Klatschen oder Musik beinhalten, haben in der einfachen Texteingabe keine Entsprechung. Damit entstehen in der Modell-Kette sozusagen „Lücken", die ALPHONES durch gleichmäßiges Verteilen des angrenzenden Materials zu schließen versucht, was aber zu großen Fehlern führt. Als Lösung dafür bietet sich an, neue Modelle (Ganzwortmodelle) einzuführen, die an solchen Signalabschnitten trainiert worden sind.[13] Bei dieser Methode muss man allerdings dafür Sorge tragen, dass diese Ganzwortmodelle unverändert durch die Graphem-Phonem-Konversion geschleust werden. Würde die Graphem-Phonem-Konversion die in (1) dargestellte Ersetzung

(1)

aus MUSIK	wird	`[mu:]["zi:k]`
aus KLATSCHEN	wird	`["kla[tS]@n]`
aus mhm	wird	`[mm]`

durchführen, so wäre eine adäquate Segmentierung durch die Ganzwortmodelle nicht möglich, da in einem typischen 60 Sekunden langen KLATSCHEN-Signal normalerweise die Phonemfolge [kla[tS]@n] nicht vorkommt. Das gilt natürlich analog für dialog-sprachliche Äußerungen wie *mhm*. Ein *mhm*, das als zweigipfliges Rezeptionssignal realisiert wird, hat mit zwei isolierten [mm] nichts gemeinsam.

Für die hier angesprochenen Probleme und deren allgemeine Lösungsansätze werden in Abschnitt 6.3. experimentell gewonnene Ergebnisse vorgestellt.

6.2. Experimente

6.2.1. Experiment 1: Behandlung nicht-sprachlicher Signalanteile

In diesen Experimenten sollen Ganzwortmodelle erstellt, trainiert und optimiert werden, um Signalsequenzen abzudecken, die typisch sind für Sprachaufnahmen von Diskussionsrunden und Talkshows. Darüber hinaus sollen auch Ganzwortmodelle für Signalsequenzen erstellt werden, die unabhängig von der Domäne vorkommen, wie etwa Rauschen, Husten, Räuspern und sonstige Störgeräusche. Es werden insgesamt die in (2) gezeigten Ganzwortmodelle erstellt:

13 Dies setzt allerdings eine reichhaltigere Verschriftlichung des Textes voraus als eingangs beschrieben.

(2) ATMEN, HUSTEN, KLATSCHEN, LACHEN, RAUSCHEN, STOERUNG, MUSIK

In Abb. 7 ist die erfolgreiche Verwendung der neuen Ganzwortmodelle dargestellt. Die untere Segmentspur zeigt eine manuelle Referenzsegmentierung, die mittlere Spur ist die automatisch erzeugte Segmentierung ohne Ganzwortmodelle und die obere Spur die automatisch erzeugte Segmentierung mit den neuen Ganzwortmodellen. Hier sind die Ganzwortmodelle für Musik (MUSIK) und Klatschen (KLATSCHEN) eingesetzt. Im Falle von MUSIK wird eine minimale Abweichung von 0.3 Sekunden erzeugt, wobei in dem Intervall von [43..43,3] Sekunden noch Musik ist, aber auch schon Klatschen einsetzt, und im Falle von KLATSCHEN wird eine exakte Übereinstimmung erzeugt. Das bedeutet für das Modell KLATSCHEN in dem vorliegenden Fall eine perfekte Segmentierung. Durch die richtige Modellierung der Musik- und Klatschen-Sequenz sind auch die nachfolgenden Nutzsegmente *einen wunderschönen* ohne Abweichung zur Referenzsegmentierung zugewiesen worden, während bei der Segmentierung ohne Ganzwortmodelle erhebliche Fehler auftreten.

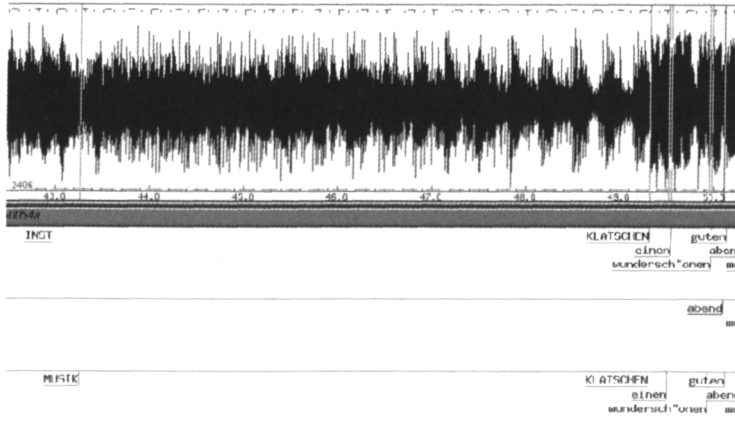

Abb. 7. Verbesserung der Segmentierung durch Ganzwortmodelle

6.2.2. Experiment 2: Behandlung von Hesitations- und Rezeptionssignalen

In diesen Experimenten sollen Ganzwortmodelle erstellt, trainiert und optimiert werden, die Signalsequenzen abdecken sollen, die unter dem Begriff ‚Hesitations- und Rezeptionsphänomene' subsumiert werden. Diese Klasse von Signalsequenzen sind domänen-unabhängig und kommen in spontan geäußerter Sprache sehr häufig vor. Es werden insgesamt die in (3) aufgezählten Ganzwortmodelle erstellt:

(3) EH, EHM, JA, JAJA, M, MH, MHM

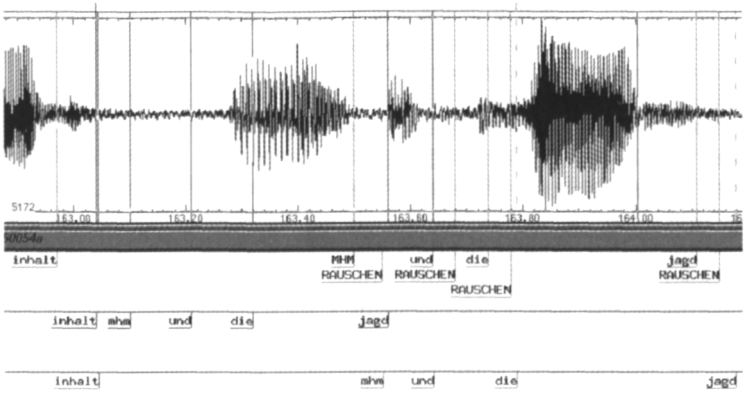

Abb. 8. Verbesserung der Segmentierung durch Ganzwortmodelle

Das Beispiel in Abb. 8 stellt die verbesserte automatische Segmentierung durch die Verwendung aller Ganzwortmodelle (mit den Ganzwortmodellen aus Experiment 1) in der oberen Spur dar. Im Unterschied zur mittleren Spur (ohne Ganzwortmodelle) zeigen die neuen Ganzwortmodelle die Synchronisation mit der Referenzspur. Das Rezeptionssegment *mhm* ist perfekt segmentiert. Alle nachfolgenden Nutzsegmente sind nun ebenfalls richtig segmentiert, während in der mittleren Spur große Abweichungen zur Referenzsegmentierung auftreten. Die eingesetzten RAUSCHEN erzeugen minimale, auditiv nicht wahrnehmbare Abweichungen.

Eine ausführliche Darstellung dieser Experimente ist in Fach (2001) wiedergegeben.

6.3. Ergebnisse der Experimente

Im Abschnitt 6.2. wurde die Erzeugung von Segmentierungen spontan-sprachlicher Aufnahmen und dazugehörige automatische Segmentierungsverfahren vorgestellt. Es wurde gezeigt, dass solche automatischen Segmentierungsverfahren in ihrer Genauigkeit sehr anfällig sind, wenn die zu verarbeitenden Sprachdaten Abweichungen von einer Standardform, wie sie etwa Nachrichtensprecher produzieren, aufweisen. In zwei Experimenten wurden Methoden vorgestellt, um die Genauigkeit auch bei schwierigen Sprachdaten wie Diskussionsrunden mit mehreren Teilnehmern auf einem relativ hohen Niveau zu halten. Diese Methoden können prinzipiell auch auf „schwierige" Sprachdaten wie Dialektdaten angewendet werden, allerdings ist dort der Erfolg geringer als in den oben gezeigten Experimenten. Die Methoden beinhalten zum Einen eine adäquate Modellierung von spontansprachlichen Eigenschaften, wie etwa Hesitations- und Rezeptionsphänomenen, und zum Anderen nicht-sprachlicher Eigenschaften, wie etwa Störgeräuschen, Lachen und Klatschen. Eva-

luierungstests auf Sprachaufnahmen von Diskussionsrunden ergaben, wie in Abb. 9 gezeigt, Steigerungen der Segmentierungsgenauigkeit von bis zu 13%.[14]

Die Kurven in Abb. 9 – die Kurve mit den Kreis-Symbolen zeigt die automatische Segmentierung ohne und die Kurve mit den Karo-Symbolen mit Ganzwortmodellen – zeigen deutlich den Unterschied bei der Verwendung von Ganzwortmodellen und Einzelphonemmodellen. Besonders interessant ist dabei der enorme Anstieg der Genauigkeit im unteren Bereich des Auswertungsintervalls. Zusammenfassend gilt:

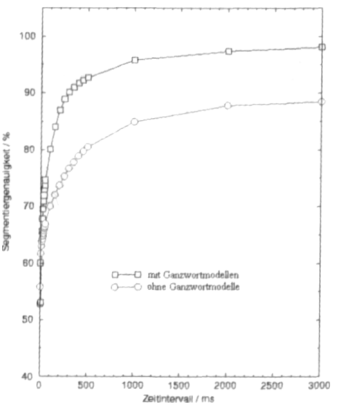

Abb. 9. Evaluierung auf „Talkshow"-Korpus

- Nicht-sprachliche Phänome bewirken globale Fehler bei der automatischen Segmentierung von Sprachdaten.
- Modelle, die auf solchen Phänomenen trainiert und eingesetzt werden, heißen Ganzwortmodelle.
- Ganzwortmodelle können anders als Einzelphonem-Modelle lange Signalsegmente adäquat modellieren.

7. Das Recherchieren in Diskurstranskripten mit der Volltextdatenbank COSMAS II

In diesem Abschnitt wollen wir eine am IDS entwickelte Datenbank vorstellen, mit deren Hilfe wir Diskurstranskripte in sehr detaillierter Weise recherchierbar machen. Bestimmt durch ihren Endzweck, nämlich das Untersuchen diskursanalytischer Vorgänge, besitzen diese Transkripte eine Diskursgliederung und eine Reihe spezifischer Annotationen, deren Verwaltung und Recherchierbarkeit gewährleistet werden muss. Zuerst wollen wir uns also den Grundmerkmalen zuwenden, die aus unserer Sicht Diskurstranskripte charakterisieren. Daraus leiten wir ein *Diskurstranskriptmodell* ab, das wir unseren Diskurstranskriptwerkzeugen zugrunde legen wollen. Danach stellen wir unsere Volltextdatenbank COSMAS II vor, die aufgrund dieses Modells ein diskursanalytisches Recherchieren in den Transkripten ermöglicht.

14 Korpus mit ca. 1 h Dauer, Mitschnitt einer Talkshow mit 5 Sprechern unterschiedlicher Sprechweise.

7.1. Vom Text zum Diskurstranskript

Das primäre Ziel von COSMAS I und II[15] ist das Volltext-*Retrieval* auf Korpora geschriebener Sprache. Als Ausnahme werden von diesen Datenbanken auch die drei folgenden Korpora gesprochener Sprache angeboten: das *Freiburger Korpus*, das *Pfeffer-Korpus* und das *Dialogstrukturen-Korpus*. Dabei waren bislang die Suchmöglichkeiten in beiden Typen von Korpora weitgehend identisch.

Transkripte mit einfacher Diskursstruktur (z.B. sind viele Transkripte Interviews, die hauptsächlich die Form eines Monologs und praktisch keine Interaktion zwischen den Teilnehmern aufweisen) und solche, bei welchen man auf das Markieren von Überlappungen ganz verzichtet hat, können mit Volltext-*Retrieval* (für geschriebene Sprache) ausreichend untersucht werden. Eine erste Schwierigkeit tritt auf, wenn Transkripte Kommentare (z.B. *ZEIGT MIT DER HAND*) oder in Fließtext verfasste Annotationen (z.B. *LACHT KURZ*) aufweisen. Mit COSMAS II werden alle Arten von Text, die keine Äußerungen sind, strukturell bzw. intern vom gesprochenen Text getrennt als Annotationen verwaltet.

Bei Transkripten, die im Hinblick auf diskursanalytische Untersuchungen verfasst wurden, muss man noch einen Schritt weiter gehen, weil die nach gewissen Transkriptionsrichtlinien markierte Diskursgliederung (Äußerungsgrenzen, Simultanpassagen usw.) zusätzlich von der Datenbanksoftware verwaltet werden muss, um explizit recherchierbar gemacht werden zu können. Wir wollen in der Folge zeigen, wie COSMAS II um ein *Volltext-Retrieval für Diskurstranskripte* erweitert wurde.

Die in diesem Sinne realisierte Erweiterung von COSMAS II schließt sich an eine Infrastruktur von spezialisierten Komponenten für die Handhabung von Diskurstranskripten an. Es sind dies: Editieren (DIDA), Darstellung (DIDA, COSMAS II), Ausdruck und Export (DIDA, COSMAS II), Konversion (DT-Manager[16]), Anreicherung (DT-Manager zusammen mit Annotationswerkzeugen) und Recherchieren (COSMAS II). Diese Werkzeuge erfordern ein gemeinsames *Diskurs(transkript)modell*.

7.2. Ein Modell für Diskurstranskripte

Das Modell für Diskurstranskripte besteht aus einer Sammlung von Regeln, die sowohl beim Transkribieren als auch bei der Konversion der Transkripte in das COSMAS-II-eigene Eingangsformat beachtet werden müssen. Durch die technische Umsetzung des Modells in COSMAS II wird ein diskursanalytisches

15 COSMAS = *Corpus Storage and Maintenance System*. WWW-Zugang zu COSMAS I: <http://www.ids-mannheim.de/~cosmas>.

16 DT-Manager = Diskurstranskript-Manager, ein am IDS entwickeltes Diskurstranskript-konversionswerkzeug.

Recherchieren, das über die Möglichkeiten herkömmlicher Volltextdatenbanken hinausführt, erst möglich (siehe die Beispiele am Ende dieses Abschnitts). Außerdem soll gewährleistet werden, dass Suchanfragen, die in den Transkripten durchgeführt werden, die gesuchten Textstellen unabhängig von der Transkriptkodierung (z.B. der Reihenfolge der Wörter in den Simultanpassagen) wiederfinden können. Wie wir gleich sehen werden, liegt der Schwerpunkt dieses Modells in der Beschreibung der Diskursgliederung und der Relation zwischen den Diskurseinheiten.

Das Modell:

- sprecherbezogener Redefluss: Der Text eines Transkripts besteht aus einem Redestrang pro Gesprächsteilnehmer. Jeder dieser Stränge muss über Unterbrechungen, sei es durch kurze Einschübe oder durch längere Pausen des Teilnehmers, hinweg verfolgt werden können.
- Simultanpassagen: Simultanpassagen müssen eine klare Anfangs- und Endbegrenzung haben. Innerhalb einer Simultanpassage besteht nur noch unter den Wörtern eines einzelnen Sprechers eine zeitliche Relation. Wörter mehrerer an einer Simultanpassage Beteiligter hingegen können zeitlich nicht in Relation gebracht werden. Für sie gilt, dass sie (in diesem Modell) gleichzeitig gesprochen wurden. Eine zeitliche Relation wäre in der Regel nur schwer zu hören und oft im Hinblick auf eine Untersuchung sowieso nicht notwendig. Einige Transkriptionsvorschriften schreiben keine Endbegrenzung vor. Das ist aber für Recherchen über die Gleichzeitigkeit von Äußerungen nicht ausreichend.
- Annotationen über Veränderungen des Redeflusses: Änderungen in der Dynamik oder im Tempo werden im Hinblick auf die Recherchen paarweise markiert.
- Wortfragment: Die kleinste Einheit ist das Wortfragment. In der Regel fällt ein Wort mit einem Fragment zusammen, außer wenn ein Ereignis innerhalb des Wortes festgehalten wurde und das Wort dadurch in mehrere Fragmente zerfällt. Ein solches Ereignis kann eine Mikropause oder die Grenze einer Simultanpassage sein.
- Ereignisse und nicht-lexikalisierte Äußerungen: Ereignisse wie Pausen und nicht lexikalisierte Äußerungen sind eigenständige Einheiten des Diskurses. Im Gegensatz zu anderen Annotationen (wie z.B. prosodische Annotationen) beziehen sie sich nicht direkt auf Äußerungen.
- Wortabstand: Zwischen den Wörtern eines Sprechers lässt sich ein Wortabstand definieren. Benachbarte Wörter haben einen Wortabstand von 1. Wir betrachten auch dann Wörter als Nachbarn, wenn Einschübe von anderen Sprechern vorkommen oder wenn sie durch Äußerungsgrenzen getrennt sind. Zwischen den Wörtern unterschiedlicher Sprecher wollen wir hingegen keine Abstandsrelation definieren. In Simultanpassagen wird deutlich, dass sich eine solche Relation nicht definieren lässt. Zwischen Wörtern und Ereignissen oder nicht-lexikalisierten Äußerungen wollen wir sinnvollerweise auch keine Wortabstandsrelation definieren.

- Zeitliche Abfolgerelation: Zwischen den Wörtern, Ereignissen und nichtlexikalisierten Äußerungen definieren wir eine zeitliche Abfolgerelation, so dass wir zwischen je zwei solchen Einheiten zweier Sprecher sagen können, ob sie nacheinander oder gleichzeitig vorkommen. In einer Simultanpassage treten alle Einheiten gleichzeitig ein.

7.3. Die Volltextdatenbank COSMAS II

Die Volltextdatenbank, die wir nun vorstellen wollen, setzt das soeben skizzierte Diskurstranskriptmodell um. In einem so genannten Indexierungsprozess werden die Diskurstranskripte in komprimierter Form in die Datenbank aufgenommen und auf einen schnellen Recherchezugriff vorbereitet. Parallel dazu wird ein separates Archiv mit den Audiodateien, die mit den Transkripten aligniert worden sind, aufgebaut. Der Datenbestand sieht zur Zeit wie folgt aus:

- über 330 Diskurstranskripte
- 1,3 Mio. Wortformen
- entspricht ca. 150 Aufnahmestunden
- davon ca. 20 Std. aligniert.

Zur Zeit stellt COSMAS II den Zugriff auf die Diskurstranskriptdatenbank nur innerhalb des IDS zur Verfügung.

7.4. Allgemeines Vorgehen beim Recherchieren

Transkriptauswahl oder -zusammenstellung: Vor dem eigentlichen Recherchieren muss man eine Transkriptauswahl treffen. Man hat die Wahl zwischen dem gesamten Datenbankinhalt, einem vordefinierten Korpus (in der Regel eine Sammlung von Transkripten, die für ein Projekt zusammengestellt worden sind) oder einem *virtuellen Korpus*, das man sich selbst zusammenstellt. Als Auswahlkriterien stehen dabei zur Verfügung:[17] das *Aufnahmedatum* (wenn bekannt), die *Textsorte* (eine Bezeichnung für die während der Aufnahme vorherrschende Art des Diskurses: *Interview*, *Schlichtung*, *Talkshow*, usw.), die *Korpusbezeichnung* und Textfelder[18] (z.Z. noch sehr rudimentär ausgefüllt). Diese Kriterien lassen sich untereinander kombinieren.

17 Die Liste der zur Auswahl stehenden Transkripteigenschaften, auch bibliografische Angaben genannt, wird in Zukunft sicherlich erweitert werden, um Angaben wie z. B. die Regionalität aufzunehmen.

18 Ein Transkript kann einen beliebig langen Titel, eine in freiem Text verfasste Beschreibung oder eine Liste von Schlüsselwörtern aufnehmen. In diesen Feldern kann mit einer Volltextsuche nach Suchbegriffen gesucht werden, die die entsprechenden Transkripte aktivieren.

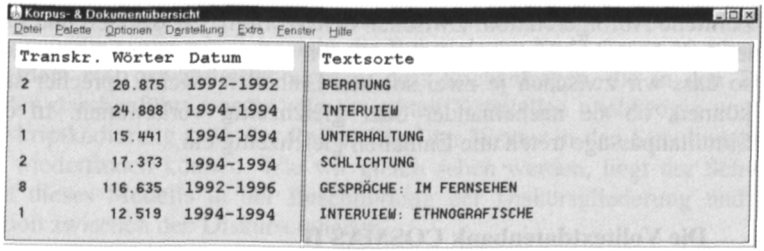

Abb. 10. Darstellung eines geöffneten Korpus nach der Textsorte

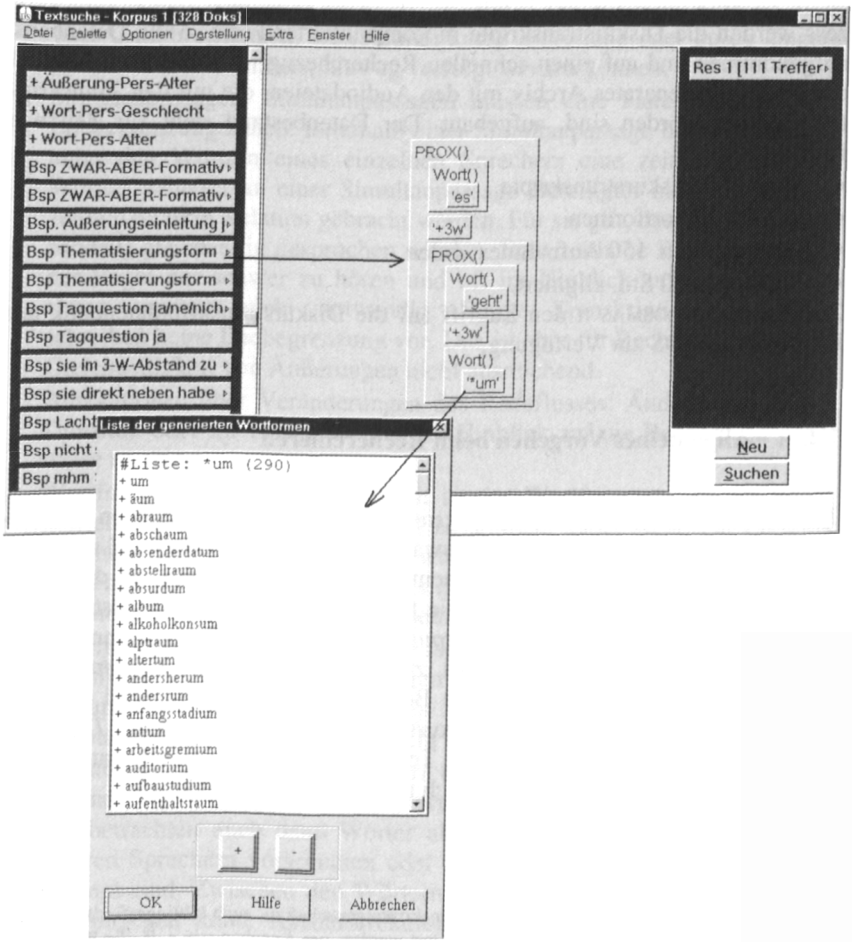

Abb. 11. Das Suchfenster von COSMAS II

Die Transkriptauswahl bzw. das zusammengestellte virtuelle Korpus kann gesichtet werden. Es werden verschiedene Darstellungsarten angeboten, bei welchen die Transkripte in nach den oben genannten Kriterien zusammengefasste und sortierte Übersichten präsentiert werden. Z.B. stellt Abb. 10 links den Inhalt des zuvor geöffneten virtuellen Korpus der alignierten Transkripte, zusammengefasst nach deren Textsorte, dar. Man kann z.B. auch über die Datierung (Aufnahmedatum) verschiedene Korpora bestimmen, etwa um Fragen der Sprachentwicklung zu untersuchen.

Recherchieren: Hat man die zusammengestellten Transkripte (fortan: virtuelles Korpus) aktiviert, kann mit dem Recherchieren begonnen werden. Die im letzten Teil dieses Abschnitts vorgestellten Suchoperatoren stehen als graphische Operatoren in einer Leiste zur Verfügung. Sie werden einzeln oder in Kombination verwendet. So lassen sich Wortformen, Prosodie,[19] Diskursgliederung und Sprecherdaten zu Suchanfragen kombinieren. Frühere Suchanfragen können in neue Anfragen eingefügt oder abgespeichert werden, um in späteren Sitzungen wiederverwendet werden zu können (siehe Abb. 11 auf der linken Seite). Bei der Wortformsuche können Platzhalteroperatoren eingesetzt werden. Diese erzeugen Auswahllisten, aus welchen man die unerwünschten Wortformen abwählen kann.

Ergebnispräsentation: Die primäre Darstellungsform der Treffer ist das KWIC (hier als einzeilige Kurzform des Wortlauts des Treffers ohne weitere Informationen aus dem Transkript). Diese Darstellung lässt sich nach verschiedenen Kriterien sortieren (z.B. alphabetisch oder chronologisch). Die Trefferstatistik kann nach unterschiedlichen Gesichtspunkten aufgelistet werden: z.B. zusammengefasst nach *Korpuszugehörigkeit* (vgl. Abb. 12 auf der folgenden Seite zur Veranschaulichung), *Diskurstyp* oder *Aufnahmedatum*. Vom KWIC aus gelangt man in die Partiturdarstellung (Abb. 12), die den Treffer in einem größeren Kontext präsentiert. Sowohl im KWIC als auch in der Partiturdarstellung kann man die Treffer in einem frei wählbaren Zeitfenster abspielen.

Weiterverarbeitung der Treffer außerhalb von COSMAS II: Das KWIC kann im ASCII-Format zum Zwecke der automatischen Weiterverarbeitung oder im RTF-Format (mit oder ohne Quellennachweis) exportiert werden.[20] Einzelne Treffer (aus alignierten Transkripten) stehen als Audio-Ausschnitte für die weitere Signalverarbeitung (z.B. in *Praat*) zur Verfügung. In einer späteren Version sollen statistische Werkzeuge, die am IDS für die geschriebene Sprache angewendet werden, versuchsweise auch auf die Transkripte angesetzt werden.

---

19 In Zukunft werden auch weitere Annotationstypen hinzukommen. Um nur ein Beispiel zu nennen: zur Zeit wird über die Anreicherung der Wortformen mit ihrer Aussprache (in Lautschrift) nachgedacht.

20 In einer folgenden Programmversion soll auch die Partiturdarstellung exportierbar werden.

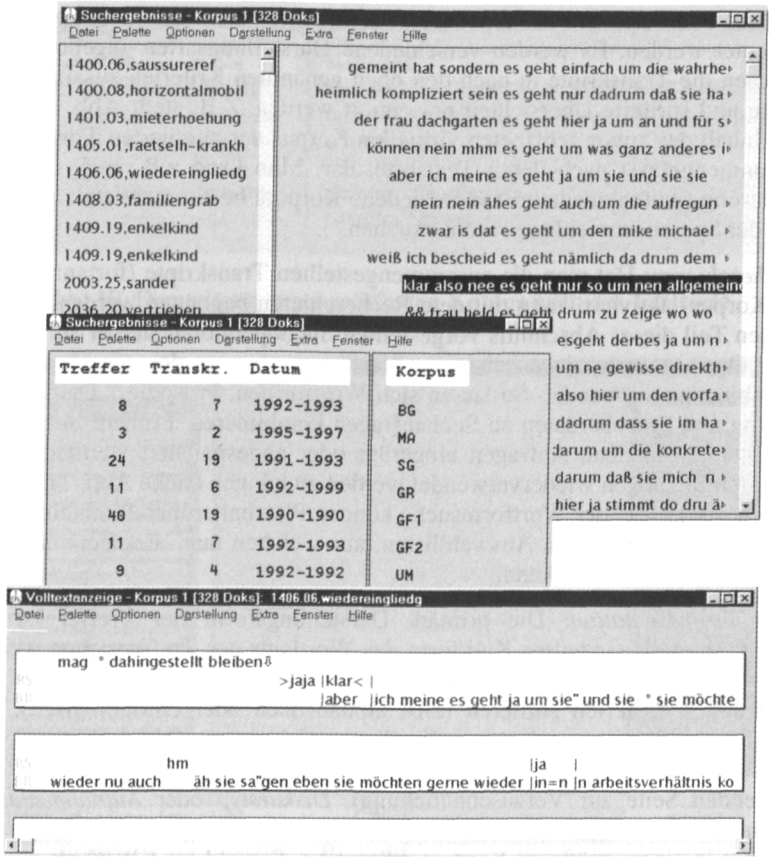

Abb. 12. Drei Ergebnispräsentationen: KWIC, Trefferstatistik nach Korpuszugehörigkeit und Partiturdarstellung

7.5. Die Suchoperatoren von COSMAS II

In diesem Abschnitt wollen wir einige der Suchoperatoren vorstellen, die COSMAS II speziell für das Recherchieren in Diskurstranskripten zur Verfügung stellt.

Der Wort-Operator WORT(x): *WORT(x)* sucht nach einem oder mehreren alternativen Suchbegriffen, die mit dem Platzhalteroperator * (Asterisk) versehen sein können. *WORT(indirekt*)* findet Vorkommnisse von *indirekt*, *indirekte*, *indirekter*, *indirektes*, etc. Falls in den Transkripten Großschreibung verwendet wird, kann durch entsprechende Optionen gesteuert werden, ob die Groß-/Kleinschreibung berücksichtigt werden soll.

Durch die Trennung von Text und Annotationen in COSMAS II ist dieser Operator im Stande, Wörter zurückzuliefern, in welchen Informationen über die Diskursgliederung notiert sind. Die Suchanfrage *WORT(indirekt)* findet demnach auch folgende Textstellen:

(4) ... *indi|rekt* ... (Wort auf einer Überlappungsgrenze)

(5) ... *indi*rekt* ... (Wort mit einer Mikropause)

(6) ... *indirekt*↓ ... (Wort mit Tonfallmarkierung)

Operatoren für die Suche nach Wörtern mit prosodischen Merkmalen: In diese Kategorie fallen mehrere Operatoren, wie z.B. *BETONUNG*, ein Suchoperator, der Wörter findet, bei welchen festgehalten wurde, dass eine Silbe auffallend betont wurde.

(7) ... *indi:rekt* ... (Wort, bei dem die Silbe *di* auffallend betont wurde)

Kombinieren von Wortform und prosodischen Merkmalen: Operator *GLEICH(X,Y)*:

(8) GLEICH(WORT(nicht), INTONATION(steigend))

(9) GLEICH(WORT(nicht), GLEICH(INTONATION(steigend), DEHNUNG-LANG))

Das zweite Beispiel, das durch Verschachteln von zwei *GLEICH*-Operatoren drei Bedingungen verknüpft, lässt sich als die Suche nach dem Wort *nicht* umschreiben, das sowohl eine steigende Intonation als auch eine besonders lang gedehnte Silbe aufweist. Mit solchen Suchanfragen lassen sich in unseren Diskurstranskripten Fragesätze aufspüren (da annotierte Äußerungssegmente vom Typ ‚Fragesatz' fehlen).

(10) ...glaubst du nicht↑ ...

(11) ...willst du ni::cht↑ ...

Operatoren für nicht-lexikalisierte Äußerungen: Nicht-lexikalisierte Äußerungen werden mit dem Operator VOCAL(Schlüsselwort) gesucht, z.B. VOCAL(grinst) oder VOCAL(lacht kurz).[21]

Operatoren für die Diskursgliederung: Die Suchoperatoren dieser Kategorie liefern Wörter zurück, die sich an markanten Stellen der Diskursgliederung befinden. Dazu gehören die Äußerungsgrenzen und die Simultanpassagen.

Operatoren für die Sprecherdaten: Zu den Sprecherdaten gehören z. B. die Altersangabe und das Geschlecht. Diese Operatoren schränken Suchanfragen auf Sprecher mit bestimmten Merkmalen ein.

---

21 *VOCAL* [vocalized]: Name übernommen aus dem TEI-Standard, der unserer Korpuskodierung zugrunde liegt.

Sprecherbezogener Wortabstandsoperator ABSTAND(X nw Y): Dieser Abstandsoperator gewährleistet, dass die gefundenen X- und Y-Textstellen vom selben Sprecher stammen und höchstens den angegebenen Abstand (in Worteinheiten) aufweisen. Wie schon weiter oben ausgeführt wurde, ist dieser Operator im Stande, X- und Y-Stellen miteinander zu kombinieren, zwischen denen Einschübe von anderen Sprechern vorkommen. Diese Besonderheit hebt ihn von ähnlichen Abstandsoperatoren gewöhnlicher Volltextdatenbanken ab.

(12) *ABSTAND(WORT(nicht) +1w INTONATION(steigend))* liefert *nicht*-Stellen zurück, die unmittelbar vor einem beliebigen Wort (*+1w*) mit steigender Intonation gesprochen wurden.

Diese Suchanfrage liefert in den beiden folgenden Transkriptausschnitten je einen Treffer:

(13)
```
AA: das |stimmt| so |nicht oder↑|
BB:     |also |    |absolut     |
```

(14)
```
AA: das stimmt so nicht      oder↑
BB:                    doch
```

Der Abfolgeoperator ABSTAND(X ns Y): Dieser Operator betrachtet das Transkript als Abfolge von Zeitsegmenten. Eine Simultanpassage gilt in dieser Betrachtungsweise als ein Zeitsegment. Wörter außerhalb von Simultanpassagen bilden jedes für sich ein Segment. Zwischen zwei Zeitsegmenten lässt sich immer eine zeitliche Abfolge festlegen. Innerhalb eines Segments hingegen (also für alle Wörter innerhalb einer Simultanpassage) gilt Gleichzeitigkeit. Im Gegensatz zum Wortabstand stellt dieser Abstandsoperator sowohl für Wörter eines selben Sprechers als auch verschiedener Sprecher einen Bezug her. Er ist insbesondere als Instrument zur Untersuchung von Phänomenen in und um Simultanpassagen wichtig und unterscheidet sich deshalb ebenfalls von Abstandsoperatoren herkömmlicher Volltextdatenbanken.

(15) *ABSTAND(WORT(nein) 0s WORT(weil))* liefert *nein-weil*-Wortpaare zurück, die innerhalb einer Simultanpassage auftreten. Durch die Nullsegmentangabe (*0s*) werden automatisch Simultanpassagen durchsucht.

Hier ein Beispiel für eine entsprechende Textstelle:

(16)
```
AA: das stimmt |weil ich | ich
BB:            |nein nein|     nein da will ich ihnen
```

(17) *ABSTAND(IN-SP-GESCHL(WORT(weil) fem) +1s IN-SP-GESCHL(WORT(nein nee nö) masc))*

Diese Suchanfrage sucht nach Vorkommnissen von *weil* in weiblichen Äußerungen, die unmittelbar (*+1s*) danach von einem *nein* (oder *nee* oder *nö*) eines männlichen Teilnehmers gefolgt werden. Mögliche Situationen treten bei Simultanpassagen oder Sprecherwechsel auf, z.B.:

(18) AA(♀): das stimmt **weil**|ich
BB(♂): |**nee nee**| das können sie so

(19) AA(♀): das stimmt **weil**↓
BB(♂): **nee** das können sie so nich

7.6. Anwendungsbeispiele für COSMAS-II-Recherche: Formen der Kommunikationsregulierung in Fernseh- und in Alltagsgesprächen[22]

Als Beispiel für eine gesprächsanalytische Recherche gehen wir der Frage nach, an welchen Stellen mit welchen Formulierungen auf Normen und Leitvorstellungen der Kommunikation Bezug genommen wird. Dabei geht man in COSMAS II so vor:

Korpusauswahl: Alle Recherchen werden einmal mit dem Korpus der Fernsehgespräche durchgeführt (das sind vor allem Talkshows; sie beginnen alle mit der Systemnummer 4050nnn – „nnn" = laufende Nummer) und im Vergleich dazu mit dem Gesamtkorpus von Gesprächsaufnahmen; für jede Anfrage sind also nacheinander zwei „virtuelle" Korpora zu laden.

Formulierung von Suchanfragen: Formen der Bezugnahme auf Normen und Leitvorstellungen sind natürlich in den Gesprächskorpora nicht explizit annotiert, etwa in Form von interpretativen Kommentaren. Darum kann man nicht direkt nach ihnen suchen, sondern muss sich „intelligente", „fantasievolle" bzw. „kreative" Suchanfragen überlegen: Durch welche Zeichenfolgen an der „Text-Oberfläche", gegebenenfalls in Verbindung mit welchen Annotationen (Prosodie / Pausen / Sprecherwechsel / Simultaneität) erhält man Belege für das Phänomen? Wenn die Suchanfrage dann zunächst zu unspezifisch ist, man also viele Belege bekommt, die nichts mit dem gesuchten Phänomen zu tun haben, muss man sich einschränkende Zusatzbedingungen überlegen.

Ein heuristischer Anhaltspunkt dazu sind vorliegende Einzelfallanalysen mit ausgesuchten Belegen und Transkriptbeispielen. In diesen Einzelfallanalysen war die Leitfrage: Wie formulieren die Gesprächsbeteiligten Normen und Leitvorstellungen der Kommunikation, wie fordern sie die Einhaltung dieser Normen und Leitvorstellungen mit welchen Formulierungen ein?

Ausgangspunkt der Recherche sind Formen der Rederechtssicherung (im Rahmen der sogenannten „Gesprächsorganisation"): Wie wehren Gesprächsbeteiligte Unterbrechungsversuche durch andere Beteiligte oder andere Störungen ab, wie versuchen sie, „am Ball" zu bleiben, also den *Turn* zu behalten, weiter zu reden? Dazu gibt es sicherlich Standardformulierungen wie *Lassen Sie mich bitte ausreden!* Man könnte also in COSMAS II zunächst als einfache Wortanfrage nach allen Vorkommen für *ausreden* suchen; Problem

22 Dieser Abschnitt ist in erweiterter Version auf der beiliegenden CD-ROM unter BODMER_FACH_SCHMIDT_SCHUETTE\ANNEXES\ANNEX_4.PDF zu finden.

ist dabei möglicherweise die Homonymie mit dem Plural des Nomens *Ausrede* (im Sinne von *Ausflucht*); die beiden Lesarten von *ausreden* werden ja in DIDA bzw. COSMAS II weder durch Klein-/Großschreibung noch durch morphosyntaktische Annotationen auseinander gehalten. Man könnte also die Anfrage etwa so spezifizieren: Gibt es Belege, wo *lassen* im Abstand von maximal 4 Wörtern von *ausreden* vorkommt? Diese Anfrage könnte man dann mehrfach variieren:

a. *Variation bei der Anrede*: *lassen* oder *lass* (auch die alte Rechtschreibung *laß* muss zugelassen werden!). Dabei sollte auch das Vorkommen von „lateraler Adressierung" überprüft werden (z.B. *ich möchte, dass er / man mich ausreden lässt*, also alternativ auch nach *lässt / läßt* im Kontext von *ausreden* suchen).
b. *Veränderung des Abstands von* lassen *und* ausreden: Gibt es mehr Belege, wenn man den Abstand erhöht? Tauchen dann auch Pseudo-Belege auf, wo *lassen* nur zufällig in der Nähe von *ausreden* steht, aber nicht zur selben syntaktischen Struktur gehört? Was passiert, wenn man den Abstand verkleinert, z.B. auf „1w"? Dann könnten eigentlich nur Formulierungen wie *(...) ausreden lassen (...)* als Rechercheergebnisse vorkommen, sofern es sie im Korpus überhaupt gibt.
c. Um die *Verteilung von Höflichkeitsformeln* mit zu berücksichtigen, *bitte* hinzunehmen (Anfrage also etwa [„lassen" 4w „ausreden"] 4w „bitte").
d. Vermutlich tauchen solche Floskeln *kurz nach Simultanpassagen* auf, in denen zwei oder mehr Sprecher um das Rederecht konkurrieren. Als Zusatzbedingung könnte man also die Nähe von Simultanpassagen angeben (z.B. mit dem Abstandsoperator „+3w"; „+" bedeutet dabei, dass erst die Simultanpassage kommt, dann die Floskel zur Rederechtssicherung).

Die Recherche zur Rederechtssicherung kann man dann ausweiten: Gibt es andere Formulierungen, mit denen ein Sprecher das Rederecht einfordert und andere am Sprechen hindern will?

Wenn das Thema „Rederechtssicherung" für die COSMAS-II-Recherche erschöpft ist, lässt sich nach weiteren Formen der Kommunikationsregulierung suchen. Vermutlich wird man vor allem bei metakommunikativen Kommentaren zum ablaufenden Gespräch fündig werden. Auch hierfür können die Beispiele aus vorliegenden Analysetexten als Ausgangspunkt, sozusagen als „Schablone", genommen werden, um Suchanfragen zu formulieren und zu spezifizieren. Vermutlich werden dabei besonders häufig Modalverben (*müssen*, *dürfen*, *sollen*) vorkommen, wenn Sprecher Ansprüche an die eigene Gesprächsbeteiligung oder die ihrer Gesprächspartner formulieren oder ihre Partner kritisieren. Dabei kann man die Zweckmäßigkeit einer *Wildcard* (*) in der Suchanfrage nach dem Stamm (z.B. „soll*") überprüfen, um nach allen Flexionsformen solcher Modalverben zu suchen.

„Gelungene" Anfragen, bei denen die Rechercheergebnisse das erbringen, was man sich bei der Formulierung der Suchanfrage vorgestellt hat, kann man im linken Feld von COSMAS II sichern. Die Rechercheergebnisse (im KWIC-Format) lassen sich jeweils in RTF-Dateien mit eindeutigen Dateinamen exportieren; der Gang der Recherche lässt sich protokollieren.

Typische Formulierungen für Metakommunikation lauten in der COSMAS-II-Syntax beispielsweise *‚was' 1w ‚sie' 3w ‚da' 3w ‚sagen'*, *‚is* unsinn'*, *‚das' 10w ‚gemeint'*, *‚ich' 3w ‚möchte' 10w ‚wissen'*, *‚ich' 2w ‚frage' 2w ‚sie'*, *‚ich' 3w ‚finde' 3w ‚das'*, *‚meine* ihre* seine* der die' +1w ‚auffassung'*.[23] Als Formulierungen zur Rederechtssicherung werden z.B. *ausreden*, *unterbrechen*, *‚lassen' +8w ‚ausreden'*, *‚ausreden' +4w ‚lassen'*, *(‚lassen' 6w ‚sie') 6w ‚mich'*, *‚hören' 6w ‚sie' 6w ‚zu'*: *‚hören*' 3w ‚mal'*: , *‚nich*' 3w ‚stören'* verwendet.[24]

Wir schließen diese Überlegungen zur Strukturierung von Recherchen ab mit zwei Beispielen für Ergebnisse einer gesprächsanalytisch motivierten COSMAS-II-Suche, die das Ausgabeformat verdeutlichen sollen.

Tag questions sind Vergewisserungsfragen, mit denen der Sprecher eine Bestätigung oder Rezeptionssignale des oder der Adressaten einfordert, mit denen also die etablierte Rederechtsverteilung „einer redet, die anderen hören zu" ratifiziert werden soll. Die Suche ergibt im kleinen Korpus der alignierten Transkripte 14 Treffer.[25] Diese Vorkommen von Vergewisserungsfragen können nun zwecks Vergleich und Sprecheridentifizierung an ein Programm zur Prosodieanalyse wie *Praat* übergeben und dort z.B. mit Grundfrequenzanalysen und Spektrogrammen näher bestimmt werden.

Thematisierungsformeln (mit zugelassener Inversion) vom Typ *es (...) geht (...) um* (mit zugelassenen Wortabständen von maximal drei Wörtern, um syntaktisch angereicherte Konstruktionen, z.B. *es geht jetzt / mir / vor allem um (...)* mit zu erfassen) ergeben in einem ersten schnellen Zugriff Hinweise auf das Gesprächsthema und auf Gesprächsphasen, in denen das Thema Gegenstand einer Aushandlung zwischen den Gesprächsbeteiligten, möglicherweise auch strittig ist. Typischerweise tauchen solche Formeln in kontroversen, argumentativen Gesprächen im weiteren Verlauf auf, wenn aus Sicht eines der Beteiligten das Gespräch in eine falsche, nicht mit den vordefinierten oder eingangs ausgehandelten Themen kompatible Richtung zu laufen droht – mit *es geht um (...)* versuchen Beteiligte dann, die verabredete thematische Ausrichtung des Gesprächs einzuklagen; sie benutzen dann oft kontrastierende Formelpaare wie *es geht nicht um (...)*, *es geht vielmehr um (...)*.[26]

---

23 Die Trefferlisten im KWIC-Format für Suchanfragen zur Metakommunikation in Fernsehgesprächen sind auf der beiliegenden CD-ROM unter BODMER_FACH_SCHMIDT_SCHUETTE\ANNEXES\ANNEX_2.PDF beigefügt.

24 Die exportierten Treffer (KWIC) für Suchanfragen zur Rederechtssicherung in Fernsehgesprächen finden sich auf der beiliegenden CD-ROM unter ANNEX_3.PDF.

25 Die exportierten Treffer (KWIC) für Suchanfragen zu *Tag Questions* in Fernsehgesprächen sind auf der beiliegenden CD-ROM unter ANNEX_4.PDF (S. 4f.) zu finden.

26 Der Anfang einer exportierten Liste von KWIC-Belegen zu Thematisierungsformeln aus den IDS-Gesprächskorpora, dargestellt an Beratungsgesprächen, findet sich auf der beiliegenden CD-ROM unter ANNEX_4.PDF (S. 6).

8. Ausblick

Aufnahmen aus authentischen natürlichen Gesprächssituationen sind für die gesprächsanalytische und pragmatische Forschung eine unverzichtbare empirische Grundlage. Transkripte stellen dabei notwendige Hilfsmittel dar. Ihre Zuverlässigkeit und Genauigkeit ist vom Aufwand bei der Erfassung und Korrektur sowie von wechselnden Forschungsinteressen abhängig. Transkripte sollten deshalb immer nur im Zusammenspiel mit den Aufnahmen genutzt werden, an denen die Beschreibungen und Analysen letztlich überprüft werden müssen. In der Forschungspraxis war freilich bislang der Rückgriff vom Transkript auf bestimmte Stellen in der Aufnahme umständlich und unterblieb darum oft – mit dem Resultat transkriptinduzierter Ungenauigkeiten oder Fehldeutungen. Über einen gezielten Zugriff auf Ausschnitte aus dem Sprachsignal mit der Möglichkeit, diese Ausschnitte wiederholt anzuhören und computergestützt weiter zu analysieren, wird diese Ungleichgewichtigkeit aufgehoben: Transkript und Aufnahme kommen bei der Analyse gleichermaßen zu ihrem Recht. Eine Recherche mit COSMAS II ermöglicht zudem einen Zugriff auf eine Vielzahl von gleichartigen Belegen aus dem Gesamt-Korpus oder aus beliebig zusammengestellten Teilkorpora.

Ziel der Datenbankrecherche in Gesprächskorpora ist in der Regel die Mustererkennung: Welche sprachlichen Muster (also etwa Formulierungen, prosodische Merkmale bestimmter Äußerungssegmente sowie deren Kookkurrenzen) sind universell, welche korrelieren mit bestimmten Interaktionstypen, mit anderen Merkmalen, die für bestimmte Teilkorpora von Gesprächen konstitutiv sind, oder mit soziodemografischen Sprechermerkmalen?

In diesem Sinne haben Sprachtechnologie und Gesprächsanalyse gemeinsame Interessen; für Gesprächsforscher können sehr große Korpora natürlicher und dialogischer Gespräche, die sprachtechnologisch aufbereitet werden können, für neue Anstrengungen motivierend sein, Korpora zusammenzustellen und sie in einem Datenbankformat zu verwalten. Indem sie an der Entwicklung von Korpustechnologie und der Modellierung dialogischer Sprechsprache zum Zwecke statistisch basierter automatischer Analyseverfahren teilnehmen, verschaffen sich Linguisten einen Zugang zu aktuellen Entwicklungsmöglichkeiten für das *Retrieval* und die Analyse von Texten (vgl. Kallmeyer 1997).

Literatur

Becker-Mrotzek, Michael / Meier, Christoph 1999: Arbeitsweisen und Standardverfahren der Angewandten Diskursforschung; in: Brünner, Gisela / Fiehler, Reinhard / Kindt, Walther (Hrsg.): *Angewandte Diskursforschung*. Band 1: *Grundlagen und Beispielanalysen*. Opladen / Wiesbaden: Westdeutscher Verlag, 18-45.

Deppermann, Arnulf 1999: *Gespräche analysieren*. Opladen: Leske und Budrich.

Fach, Marcus 2000: Trigger für die Automatische Spracherkennung. *International Journal for Language Data Processing (Sprache und Datenverarbeitung)* 24, 35-51.

— 2001: Automatische Segmentierung, Verwaltung und Abfrage von Korpora gesprochener Sprache. Stuttgart: Universität Stuttgart (Diss.); erscheint in: *phonetikAIMS, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Lehrstuhl für Experimentelle Phonetik*. Stuttgart: Universität Stuttgart.

Greenberg, Steven *et al.* 2000: Automatic Phonetic Transcription of Spontaneous Speech (American English); in: *Proceedings of the International Conference on Spoken Language Processing*. Beijing, 330-333.

Hosom, John-Paul 2000: Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information. Beaverton, OT: CSLU, Oregon Graduate Institute of Science and Technology (PhD-Dissertation); im Internet unter <http://www.cse.ogi.edu/~hosom/hosom_thesis.ps>.

Kallmeyer, Werner 1997: Vom Nutzen des technologischen Wandels in der Sprachwissenschaft: Gesprächsanalyse und automatische Sprachverarbeitung. *Zeitschrift für Literaturwissenschaft und Linguistik* 107, 124-149.

Kipp, Andreas / Wesenick, Maria Barbara 1995: *Das Münchner AUtomatische Segmentationssystem (MAUS)* (= Verbmobil Memo 95). München: Ludwig-Maximilians-Universität / Institut für Phonetik und Sprachliche Kommunikation; im Internet unter <http://www.phonetik.uni-muenchen.de/Forschung/Publications/VMMemo-95-95.ps>.

Rapp, Stefan 1995: Automatic phonetic transcription and linguistic annotation from known text with Hidden Markov Models / An aligner for German; in: *Integration of Language and Speech in Academia and Industry. ELSNET goes east and IMACS*. Moskau; im Internet unter <http://www.ims.uni-stuttgart.de/~rapp/aligner.ps.gz>.

— 1998: *Automatisierte Erstellung von Korpora für die Prosodieforschnung* (= phonetikAIMS, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Lehrstuhl für Experimentelle Phonetik; 4:1). Stuttgart: Universität Stuttgart.

Schmidt, Rudolf / Neumann, Robert 1999: Automatic Text-to-Speech-Alignment: Aspects of Robustification; in: Matousek, Václav *et al.* (Hrsg.): *Text, Speech and Dialogue*. Berlin / Heidelberg: Springer, 72-76.

Selting, Margret *et al.* 1998: Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91-122.