Korpora gesprochener Sprache im IDS und ihre Bearbeitung – von der Aufnahme über Dokumentation und Transkription zur Datenbankrecherche

Wilfried Schütte, Mannheim

In meinem Beitrag möchte ich zeigen, in welcher Weise die aktuellen Korpus-Projekte und Datenbankentwicklungen in der Abteilung Pragmatik des Instituts für Deutsche Sprache Mannheim (IDS) Fragestellungen und Forschungspraktiken der Gesprächsanalyse bedienen. Abschließend stelle ich die Entwicklungsarbeit an der Nachfolge-Version 2.0 der Datenbank Gesprochenes Deutsch für die gegenwärtig angebotene Version vor. Anforderungen an Korpora gesprochener Sprache und ihre Transkription werden auch im Beitrag von Johannes Schwitalla in diesem Band behandelt.

1. Korpora für die Gesprächsanalyse

Typische Fragestellungen der Gesprächsanalyse sind (vgl. Deppermann 2008, 79ff.):

- Welche sprachlichen Formen gibt es für Interaktionspraktiken und welche Funktionen haben sie?
- Welche Interaktionsaufgaben, -probleme und -zwecke gibt es und mit welchen Formen werden sie bearbeitet?
- Welche kommunikativen Gattungen und Ereignisse gibt es, welches Repertoire haben sie?
- Welche Kommunikationsanlässe und Beteiligtenstrukturen gibt es für das Kommunikationsportrait eines sozialen Felds, welche Kommunikationsideologien werden dabei vertreten, welche Formen der Beziehungskonstitution sind manifest und welche Stilpräferenzen erkennbar? Solche Fragestellungen lassen sich nicht sinnvoll an einzelnen Gesprächen untersuchen, man benötigt dazu vielmehr ein größeres Korpus, z.T. auch mit zeitlicher Staffelung, um Sprachentwicklungsprozesse dokumentieren zu können.

Die Gesprächsanalyse verfolgt diese Fragestellungen mit aufeinander aufbauenden, z.T. rekursiv durchgeführten Arbeitsschritten: Nach der Entwicklung einer Fragestellung wird der Feldzugang hergestellt durch Kontakt, Klärung der Forscherrolle im Feld sowie die Einwilligung der aufgenommenen Personen in eine Aufnahme und ihre wissenschaftliche Auswertung. Heute sind digitale Aufnahmen Stand der Technik; die Wahl einer Audio- oder Videoaufnahme ist von der Fragestellung, aber auch von der Einwilligung und von Überlegungen abhängig, inwieweit die Aufnahme das Gespräch in seiner Natürlichkeit stört. Videoaufnahmen sind für multimodale Fragestellungen (z.B. wie das verbale und gestische Verhalten korrelieren) unverzichtbar, greifen aber stets stärker in die Aufnahmesituation ein als Audioaufnahmen. Die Daten werden durch Digitalisierung (vor

allem, wenn auf ältere noch analog vorliegende Korpusbestände zurückgegriffen werden soll), Dokumentation, Inventarisierung, Segmentauswahl und Transkription aufbereitet. Die Datenanalyse befasst sich zunächst mit Einzelfällen, dann für einen Fallvergleich mit Kollektionen. Auf eine Theoretisierung folgen Publikation und Bericht, tunlichst auch als Rückmeldung an die Untersuchten.

Die Gesprächsanalyse stellt mithin als Voraussetzung für ihre empirische Arbeit Korpora zusammen, sie bietet nach Abschluss der Projekte ihre Korpora der wissenschaftlichen Gemeinschaft an, und sie kann somit auch für die Entwicklung von Fragestellungen, für die Datenanalyse und vor allem für Kollektionen zum Fallvergleich auf vorhandene Korpora zugreifen. Die aktuellen korpusbezogenen Projekte und Datenbankentwicklungen in der Abteilung Pragmatik des IDS sollen diese Arbeit mit Korpora ermöglichen, vorhandene Angebote zusammenführen und für die Recherche verbessern:

- AGD (Archiv für Gesprochenes Deutsch, URL 1)
- Neufassung der DGD (Datenbank Gesprochenes Deutsch, URL 2)
- FOLK (Forschungs- und Lehrkorpus des gesprochenen Deutsch, URL 3) als Prototyp eines Referenzkorpus zur gesprochenen Sprache
- Projekt ‚Deutsch heute' (URL 4).

Daneben gibt es im IDS hinführende Arbeiten, die die Voraussetzungen für eine Erstellung von konsistenten Gesprächskorpora schaffen:

- Mitarbeit an der Neufassung der GAT-Transkriptionskonventionen (GAT-2, Selting et al. 2009) im Hinblick auf die computergestützte Weiterverarbeitung von Transkripten und auf multimodale Transkripte
- Transkripteditor FOLKER für die Eingabe von Transkripten in FOLK nach einem definierten Datenmodell (s.u.)
- GAIS (Gesprächsanalytisches Informationssystem, URL 5) mit infrastrukturellen Informationen zur Gesprächsforschung. Neben einer bibliografischen Datenbank zur Gesprächsforschung bietet GAIS vor allem ein Online-Handbuch zu den Arbeitsphasen bei der Korpuserstellung, u.a. zu Aufnahmegeräten, zu Audio- und Videoeditoren, zu Transkriptionskonventionen und –programmen und zu Hilfsmitteln zur Korpusverwaltung.

Korpora gesprochener Sprache dienen als empirische Grundlage für die Pragmatik, Stilistik, Soziolinguistik, insbesondere für die Gesprächsanalyse. Für ihre Fragestellungen ist es unabdingbar, ein Korpus von natürlichen Gesprächen zu erheben, d.h. von Gesprächen, die nicht eigens zum Zweck der Aufnahme arrangiert wurden, während dialektologische und variationslinguistische Fragestellungen auch mit elizitierten Materialien arbeiten können. Korpora gesprochener Sprache zu erstellen ist gegenüber Textkorpora wesentlich aufwändiger — u.a. weil aus forschungsethischen Gründen Aufnahmen nur gemacht werden dürfen, wenn die Gesprächsbeteiligten zuvor schriftlich ihr Einverständnis zur Aufnahme und wissenschaftlichen Auswertung gegeben haben, und weil die Gesprächs-

aufnahmen nach einem konsistenten Schema mit Metadaten dokumentiert und Aufnahmen für die analytische Arbeit transkribiert werden müssen. Zudem sind Maskierungen zu personenrelevanten Angaben in der Audioaufnahme und im Transkript vor jeder öffentlichen Präsentation und Einstellung in eine Datenbank unabdingbar.

Traditionell wurden solche Korpora angelegt an unterschiedlichen Orten und in einzelnen Projekten, also an Universitäten, Forschungsinstituten oder für Qualifikationsarbeiten wie Magisterarbeiten und Dissertationen. Korpora gesprochener Sprache sollten nach Abschluss des einzelnen Projekts auch anderen Projekten und Forschern zur Verfügung stehen. Dazu ist es freilich sinnvoll, dass sie in Bezug auf ihre Dokumentation und die Transkriptionskonventionen standardisiert sind.

2. Das Archiv Gesprochenes Deutsch (AGD)

Am IDS wurden im Archiv Gesprochenes Deutsch (AGD) und seinen institutionellen Vorgängern schon immer Korpora aus abgeschlossenen Forschungsprojekten übernommen, um sie für zukünftige Forschung und Lehre zu erhalten. Das Archiv verfügt über 17.000 Aufnahmen von über 5000 Stunden Dauer, ein Großteil davon steht öffentlich zur Verfügung und wird auf Anforderung auf CD verschickt. Darüber hinaus verfügt es über 6700 Transkripte zu den Aufnahmen, die es ebenfalls auf Anforderung per Mail verschickt. 3000 dieser Transkripte befinden sich in der DGD, können also online abgerufen werden. Ihre Text-Ton-Synchronisierung, das sogenannte ‚Alignment', bewirkt, dass Recherchetreffer durch einen Mausklick ins das Transkript in einem 30-Sekunden-Ausschnitt aus der Aufnahme zu hören sind. Die AGD-Website (URL 6) gibt einen Überblick und Detailinformationen über die gesamten Bestände.

Öffentlich abgeschlossene Projektkorpora bereit zu stellen, ist leider mit einer Reihe von Problemen verbunden:

- Dokumentation: In Projekten werden oft viele Informationen nicht dokumentiert, weil sie allen Mitarbeitern selbstverständlich sind — bei der Übergabe ans Archiv fehlen diese Informationen dann. Zudem hat natürlich jedes Projekt gemäß seiner Fragestellung eigene Vorgaben, welche und wie viele Informationen überhaupt erfasst werden. Das macht eine systematische Recherche im Archiv über die Metadaten problematisch.
- Transkription: Auch die Transkripte werden gemäß den Projektzielen unterschiedlich vollständig, unterschiedlich detailliert und nach unterschiedlichen Konventionen angefertigt. Das erschwert eine systematische Recherche im Archiv über alle Korpora.
- Datenschutz: Häufig haben die Sprecher nur der Nutzung im Projekt zugestimmt. Oft wurden diese Einverständniserklärungen eingeholt lange bevor es die Digitalisierung und das Internet gab. Es ist in vielen

Fällen unklar und rechtlich strittig, ob die Korpora weiter verwendet werden dürfen.

- Aktualität: Oft werden Korpora erst 10-20 Jahre nach Projektabschluss ins Archiv eingestellt. Sie waren zunächst für eine Auswertung im Sinne der projektspezifischen Ziele reserviert, Abschlusspublikationen können sich verzögern, und viele Korpora stehen erst nach Emeritierungen zur Verfügung. So hinken sie dem aktuellen Sprachgebrauch hinterher.
- Stratifikation: Die Zusammensetzung der Bestände ist zufällig und hängt nur davon ab, welche Interaktionstypen in welchen sozialen Milieus die Projekte untersucht haben, die dem AGD ihre Korpora übergeben haben. Aussagen über den Sprachgebrauch können also auch nur für diese speziellen Kontexte Gültigkeit beanspruchen.

Um diesen historisch bedingten Einschränkungen zu begegnen, entstehen derzeit in der Abteilung Pragmatik des IDS zwei moderne Forschungskorpora: das ‚Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) und das Korpus ‚Deutsch heute' (DH, URL 4). Sie sollen die notwendigen Anforderungen erfüllen:

- Dokumentation: Die Korpora werden systematisch nach einem im IDS neu entwickelten universellen Metadaten-Schema dokumentiert, das eine systematische und umfangreiche Recherche ermöglicht.
- Transkription: Zumindest FOLK wird vollständig transkribiert (DH in den relevanten Teilen), und zwar nach einer neu entwickelten Version von GAT, die speziell für die maschinelle Recherche optimiert wurde.
- Datenschutz: Alle Sprecher stimmen von vorne herein der öffentlichen Nutzung in einer Datenbank zu.
- Aktualität: Die Korpora werden nach der Erhebung so schnell wie möglich zur Verfügung gestellt.
- Stratifikation: Beide Korpora werden nach einem Erhebungsplan erstellt, der auf eine möglichst vollständige Erfassung ausgerichtet ist, bei DH regional, bei FOLK sozial.

FOLK wird authentisches, medial mündliches gesprochenes Deutsch in Interaktionen enthalten. Das Ziel ist, ein nachhaltig archiviertes und ausgewogenes Korpus zu erstellen, dem ein einheitliches Datenformat zu Grunde liegt — sowohl in Bezug auf die Transkriptionskonventionen als auch das Audiodatenformat. Die Daten sollen vollständig über das Internet zugänglich und recherchierbar sein. Als Nutzer sind Linguisten, aber auch Vertreter aus Deutsch als Fremdsprache und Deutsch als Zweitsprache und anderer Sozialwissenschaften anvisiert. Bis zum Jahr 2010 wird die Basisfunktionalität der Korpusdatenbank entwickelt sein, die auch den FOLKER-Editor zur Transkripterstellung umfasst. Die Daten werden standardisiert als GAT-2-Rohtranskript erfasst und in einem XML-Format gespeichert. Für die Metadaten wird ein geschlossener Satz von Deskriptoren, ein sog. ‚Controlled Vocabulary' entwickelt, der für spezifische Teilkorpora erweitert werden kann.

Die Ausgewogenheit eines derartigen Korpus ist problematisch, da allgemein gängige Kriterien einer Gattungssystematik (wie etwa die von Isenberg 1984 aufgestellten Exhaustivität, Monotypisierung, Homogenität) für den kommunikativen Haushalt im Deutschen nicht erfüllt werden können. Große Bereiche der Kommunikationspraxis sind wissenschaftlich nicht bekannt, das Gattungskonzept ist nicht ausreichend definiert, und oft bestehen keine klaren Inklusionsverhältnisse der einzelnen Gattungen. So ist eine Systematisierung des Korpus nach Sprechergruppen vorzuziehen, die auf Grund von soziodemografischen Merkmalen und sozialstilistischen Milieus gebildet werden. Diese können am besten ausgearbeitet werden, indem man einem Sprecher über einen langen Zeitraum folgt und dokumentiert, in welchen Milieus er sich bewegt.

Spezifische Lösungen muss es auch geben für die rechtlichen Implikationen eines öffentlich zugänglichen Korpus in Bezug auf die Datenschutz- und Persönlichkeitsrechte sowie das Urheberrecht. Ein Ziel des Projektes ist es, zusammen mit der DFG hier eine ‚best practice'-Richtlinie zu entwickeln. Als konkrete Lösung wird in der Praxis ein graduierter Zugang zu schützenswerten Daten für unterschiedliche Nutzergruppen realisiert.

Die Strategie für die ersten Jahre des Korpusaufbaus geht dahin, bereits bestehende und leicht zugängliche Daten in guter Tonqualität und mit ausreichender Dokumentation in FOLK zu integrieren. Einbezogen werden insbesondere solche Gattungen, die von breitem Interesse sind und rechtlich unbedenkliche Daten beinhalten. Bis 2010 ist ein Umfang von 71 Stunden angestrebt.

Wichtige Arbeitsanforderungen bei der Datenerhebung sind eine Metadatenerhebung nach vorgegebenem Schema und Einverständniserklärungen der aufgenommenen Personen, gegebenenfalls auch der Erziehungsberechtigten und der Verantwortlichen in einer Institution. Die Metainformationen umfassen:

- Basisdaten: Kennung, Zeit, Ort und Teilnehmer.
- Aufnahmebedingungen, das ‚Wer' und ‚Wie': Verantwortliche für die Aufnahme, Aufnahmegerät, das Mikrofon und deren Position, besondere Aufnahmebedingungen und -störungen, Vollständigkeit der Aufnahme, Datenträger, Mediendateien.
- Interaktionsmerkmale, die ‚Beschreibung': Anzahl der Teilnehmer, mediale Realisierung, Gesprächstyp, in der Interaktion verwendete Sprachen, Sprachmischung, Gesprächszusammenfassung, Gesprächsthemen, Gesprächsverlauf relevante Zusatz-Informationen oder sonstige Umstände und Zusatzmaterial.
- Forscherbeteiligung: an Interaktion beteiligte Forscher, Forscherinvolviertheit, Grad der Elizitierung, Vorgaben.
- Sprecherinformationen.

Einverständniserklärungen müssen Angaben zum Ziel des Projekts, Zusicherungen zum sensiblen Umgang mit Aufnahmen und Transkripten sowie zum Recht auf informationelle Selbstbestimmung enthalten.

3. ‚Deutsch heute'

Ziel des Projektes ist es, durch eine formalisierte Aufnahmesituation und ein engmaschiges Ortsnetz die Vielfalt der gesprochenen deutschen Standardsprache im gesamten deutschen Sprachgebiet zu erfassen. In über 160 Orten wurden 4-6 Informanten systematisch mit verschiedenen Sprechaufgaben wie Interview, Lesetexte, Wortlisten und Map Task aufgenommen und auf diese Weise über 1000 Stunden Aufnahmen zusammengetragen. Zur Zeit wird das Korpus im Projekt dokumentiert, transkribiert und ausgewertet, mittelfristig soll es über die DGD allgemein zur Verfügung stehen.

4. Transkription

Eine Transkription im gesprächsanalytischen Sinne umfasst neben dem Sprechtext u.a. die literarische Umschrift, Pausen, Rezeptions- und Häsitationsphänomene, Ein- und Ausatmen, nichtsprachliche Vorgänge, Überlappungen (Simultanpassagen). Die Transkription muss den späteren Recherchemöglichkeiten in Gesprächsdatenbanken Rechnung tragen. Die zusätzliche literarische Umschrift ist unverzichtbar für die Gesprächsforschung, standardsprachliche ‚Übersetzungen' würden ein unzureichendes Bild der tatsächlichen Formulierungsverfahren vermitteln (Beispiel: *hasse mal_n euro* vs. *hast du mal einen euro*). So dient die literarische Umschrift der gesprächsanalytischen Sequenzanalyse, die orthografische Transkription als segmentweise synchronisierte Variante des Transkripts der Datenbankrecherche.

Granularität und Layout der Transkription sind abhängig von Forschungsfragen und Erkenntnisinteressen: Ein Feintranskript mit genauer Notation von Phonetik und Prosodie ist zu aufwändig für viele Fragestellungen, z.B. inhaltsanalytischer Art für eine Nutzung außerhalb der Linguistik in der Sozialforschung, und schwierig zu lesen. Andererseits müssen viele kommunikative Phänomene parametrisiert dargestellt und können nicht einfach in der Transkription interpretativ und apodiktisch behauptet werden (z.B. *ironisch*).

Wichtig für den schnellen Zugriff vom Transkript auf die zugehörige Stelle in der Audiodatei ist das sog. ‚Text-Ton-Alignment': Es kann entweder automatisch als nachträgliche computergestützte Synchronisation von Transkript und Ton oder manuell im Vollzug des Transkribierens mit technisch fortschrittlichen Editoren erfolgen.

5. FOLKER

Seit 10 Jahren gibt es den Standard der GAT-Transkriptionskonventionen (Selting et al. 1998). Diese Konventionen wurden in den letzten beiden Jahren unter der Bezeichnung GAT-2 so überarbeitet, dass sie sich optimal und vollständig in formale Regeln für die digitale Verarbeitung von Transkripten umsetzen lassen (Selting et al. 2009). Diese Formalisierbarkeit ist die entscheidende Voraussetzung für leistungsfähige Werkzeuge in

zwei Phasen im Forschungsprozess — Editoren für die Erstellung von nachhaltig archivierbaren und ohne Informationsverlust zwischen Forschern austauschbaren Transkripten und Suchmaschinen für die systematische Recherche in Transkripten, z.B. in Datenbanken.

Der neue Transkriptionseditor FOLKER stellt ein solches Werkzeug dar. Er wurde von Thomas Schmidt, dem Entwickler von EXMARaLDA, für das Projekt ‚Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) entwickelt. FOLKER unterstützt das Transkribieren nach GAT-2 optimal, indem er die Einhaltung der Konventionen des Minimaltranskripts überprüft, gewährleistet die Erstellung von Transkripten in einem nachhaltig archivierbaren Format, das ohne Verluste zwischen Forschern austauschbar ist und auch in neuen Programmversionen und auf anderen Plattformen unverändert genutzt werden kann, und ist deshalb für den Aufbau von großen, konsistent recherchierbaren Beständen von Gesprächsdaten in Datenbanken geeignet.

Mit dem FOLKER-Editor werden Transkripte in einem spezifischen XML-Format erstellt. In den Editor eingebaut sind ein Audioplayer und eine Darstellung des Sprachsignals als Oszillogramm. So kann man komfortabel Segmente zum Transkribieren auswählen und dabei automatisch Zeitmarken für diese Segmente setzen. FOLKER arbeitet auf der Grundlage eines modifizierten Datenmodells für GAT-2-Minimaltranskripte (cGAT) und überprüft segmentweise bei der Texteingabe, ob die cGAT-Konventionen eingehalten werden und keine zeitlichen Inkonsistenzen auftreten. Beispielsweise sind Transkriptions-Segmente unzulässig, die sich überlappen und demselben Sprecher zugeordnet werden (‚Selbst-Überlappungen').

Im Editor lässt sich das Transkript wahlweise darstellen als Folge von Segmenten, als Partitur oder als Folge von Sprecherbeiträgen, bei denen aufeinander folgende und einem Sprecher zugeordnete Segmente zusammengefasst werden. Diese Darstellungsmöglichkeiten passen zu einer sinnvollen Abfolge von Arbeitsschritten beim Transkribieren.

Abb. 1: Beiträge-Ansicht in FOLKER

Diese Ansichten sind unterschiedliche Visualisierungen desselben Transkript-Modells. Wer transkribieren möchte, muss sich also nicht vorab und unumkehrbar für eine bestimmte Darstellung entscheiden, sondern kann in FOLKER flexibel mit unterschiedlichen Visualisierungen für dieselbe Transkript-Datei arbeiten.

FOLKER ist für ein bestimmtes Nutzungsszenario ausgelegt und unterstützt dieses effizient. Im Gegensatz dazu sind andere in der Gesprächsforschung verbreitete Transkriptionsprogramme wie der EXMARaLDA-Partitur-Editor (URL 7), ELAN (URL 8) und Praat (URL 9), Mehrzweckeditoren, die möglichst viele Nutzungsszenarien in einer Oberfläche unterstützen wollen. Mehr Informationen zu FOLKER finden sich auf der AGD-Website (URL 10), wo auch das Programm selbst nach einer kostenlosen Registrierung heruntergeladen werden kann in Versionen für Windows und Mac OS X.

6. Datenbank Gesprochenes Deutsch (DGD)

Die jetzige Version 1.0 der ‚Datenbank Gesprochenes Deutsch' (DGD) steht von der AGD-Website aus zur Verfügung. Es gibt sie in zwei Versionen: als öffentliche Version, die nur mit wenigen Diskursen bestückt ist, um die Funktionalität der Recherche und Trefferanzeige zu demonstrieren, und als Wissenschaftler-Version, die nach kostenloser Anmeldung für Wissenschaftler zur Recherche zur Verfügung steht.

Für die Version 2.0 der DGD wird derzeit ein generisches Korpusverwaltungssystem am IDS entwickelt; Hauptziele sind die Integration von historischen und aktuell erhobenen Sprechkorpora unter Nachhaltigkeits-Aspekten sowie eine objekt-orientierte Benutzerschnittstelle für die Exploration von und Recherche in Korpora (Gasch 2010). Eine auf XML-Schemata basierende Standardisierung auf den beiden Ebenen der dokumentarischen Metadaten (Gasch/Brinckmann/Dickgießer 2008) und der Transkripte definiert präzise Systemschnittstellen für den Export von vorhandenen Korpora gesprochener Sprache oder den zukünftigen Import neuer Korpora. Dabei werden auch die spezifischen Charakteristika individueller Sprechkorpora bewahrt.

Der Standardisierungs-Zugang für die Informationsverwaltung über Korpusgrenzen hinweg basiert auf XML-Schemata für Metadaten. Alle Sprechkorpus-Systeme verwalten Metainformationen zu den Quellsignalen (Mediendateien). Hingegen können die Informationsstrukturen der Datenkomponenten, die in unterschiedlichen Sprechkorpora verwendet werden, sehr voneinander abweichen im Hinblick auf die unterschiedlichen linguistischen Forschungsfragen, die von den Korpus-Erstellern formuliert wurden. Diese Unterschiede zwischen Sprechkorpora können beispielsweise bedingt sein durch die vertretenen Genres, den Grad inhaltlicher Beschränkungen, die physikalische Datenstruktur oder das Forschungsgebiet, das im Fokus steht (z.B. natürliche vs. elizitierte Sprache). Dieses Problem einer Korpusvielfalt wird durch eine webbasierte Plattform zur Navigation im Korpus gelöst. Sie beruht auf einem abstrakten Standardisierungskonzept – die Plattform soll für große Sammlungen von Sprechkorpora geeignet sein, nicht so sehr spezifische Lösungen für Datensätze von einzelnen Sprechkorpus-Projekten anbieten. Die korpusübergreifende Perspektive führt zur Definition eines generischen systemweiten Datenmodells, das eine sanfte Integration von Daten ohne Informationsverlust ermöglicht. Die Komponenten dieses Modells sind hierarchisch miteinander verknüpft (s. Abb. 2): Strukturierte XML-Dokumentations-Instanzen auf der Korpus-, Sprechereignis- und Sprecher-Ebene, unstrukturierte, halbstrukturierte oder zeitsynchronisierte Transkripte, Quellaufnahmen und gegebenenfalls zusätzliche unstrukturierte Sekundärdokumente werden mit systemweiten eindeutigen Identifikatoren verknüpft. Die Ausgabequalität einer korpusübergreifenden Informationsverarbeitung kann unterschiedlich ausfallen, weil sie immer direkt von der Datenkomponente mit der niedrigsten Struktur- und Datenqualität im Korpus vorgegeben wird.

Abb. 2: Datenmodell — Komponenten von Sprachkorpora (Gasch 2009)

6. Datenbank Gesprochenes Deutsch (DGD)

Die jetzige Version 1.0 der ‚Datenbank Gesprochenes Deutsch' (DGD) steht von der AGD-Website aus zur Verfügung. Es gibt sie in zwei Versionen: als öffentliche Version, die nur mit wenigen Diskursen bestückt ist, um die Funktionalität der Recherche und Trefferanzeige zu demonstrieren, und als Wissenschaftler-Version, die nach kostenloser Anmeldung für Wissenschaftler zur Recherche zur Verfügung steht.

Für die Version 2.0 der DGD wird derzeit ein generisches Korpusverwaltungssystem am IDS entwickelt; Hauptziele sind die Integration von historischen und aktuell erhobenen Sprechkorpora unter Nachhaltigkeits-Aspekten sowie eine objekt-orientierte Benutzerschnittstelle für die Exploration von und Recherche in Korpora (Gasch 2010). Eine auf XML-Schemata basierende Standardisierung auf den beiden Ebenen der dokumentarischen Metadaten (Gasch/Brinckmann/Dickgießer 2008) und der Transkripte definiert präzise Systemschnittstellen für den Export von vorhandenen Korpora gesprochener Sprache oder den zukünftigen Import neuer Korpora. Dabei werden auch die spezifischen Charakteristika individueller Sprechkorpora bewahrt.

Der Standardisierungs-Zugang für die Informationsverwaltung über Korpusgrenzen hinweg basiert auf XML-Schemata für Metadaten. Alle Sprechkorpus-Systeme verwalten Metainformationen zu den Quellsignalen (Mediendateien). Hingegen können die Informationsstrukturen der Datenkomponenten, die in unterschiedlichen Sprechkorpora verwendet werden, sehr voneinander abweichen im Hinblick auf die unterschiedlichen linguistischen Forschungsfragen, die von den Korpus-Erstellern formuliert wurden. Diese Unterschiede zwischen Sprechkorpora können beispielsweise bedingt sein durch die vertretenen Genres, den Grad inhaltlicher Beschränkungen, die physikalische Datenstruktur oder das Forschungsgebiet, das im Fokus steht (z.B. natürliche vs. elizitierte Sprache). Dieses Problem einer Korpusvielfalt wird durch eine webbasierte Plattform zur Navigation im Korpus gelöst. Sie beruht auf einem abstrakten Standardisierungskonzept – die Plattform soll für große Sammlungen von Sprechkorpora geeignet sein, nicht so sehr spezifische Lösungen für Datensätze von einzelnen Sprechkorpus-Projekten anbieten. Die korpusübergreifende Perspektive führt zur Definition eines generischen systemweiten Datenmodells, das eine sanfte Integration von Daten ohne Informationsverlust ermöglicht. Die Komponenten dieses Modells sind hierarchisch miteinander verknüpft (s. Abb. 2): Strukturierte XML-Dokumentations-Instanzen auf der Korpus-, Sprechereignis- und Sprecher-Ebene, unstrukturierte, halbstrukturierte oder zeitsynchronisierte Transkripte, Quellaufnahmen und gegebenenfalls zusätzliche unstrukturierte Sekundärdokumente werden mit systemweiten eindeutigen Identifikatoren verknüpft. Die Ausgabequalität einer korpusübergreifenden Informationsverarbeitung kann unterschiedlich ausfallen, weil sie immer direkt von der Datenkomponente mit der niedrigsten Struktur- und Datenqualität im Korpus vorgegeben wird.

Abb. 2: Datenmodell — Komponenten von Sprachkorpora (Gasch 2009)

Abb. 3 illustriert Suchfunktionalitäten zur Volltext-Recherche in DGD-Transkripten (Gasch 2010):

Abb. 3: Volltext-Recherche in der DGD 2.0

Hier einige Suchfunktionalitäten:

- Wildcards: [_] für ein beliebiges Zeichen, [%] für eine Folge beliebiger Zeichen.
- Groß-/Kleinschreibung wird beachtet.
- Erweiterte Suchoptionen: Suchbegriffe können mit [AND] bzw. [OR] verknüpft werden: *Italien AND Österreich, Frankreich OR Luxemburg OR Belgien.* Mit dem [NEAR]-Operator lässt sich der genaue Wortabstand von Suchbegriffen spezifizieren: z.B. *...Uraufführung des Freiburg-Films anlässlich der 850-Jahrfeier von Freiburg... NEAR((Uraufführung, Jahrfeier), 6, TRUE)* oder *...Lesetext1: Fabel (Nordwind und Sonne), normales Tempo... NEAR((Lesetext1, normales Tempo), 4, TRUE)*

Literaturverzeichnis

Deppermann, Arnulf (2008): Gespräche analysieren. Eine Einführung. 4. Aufl. Wiesbaden (=Qualitative Sozialforschung 3).

Gasch, Joachim (2010): DGD 2.0: A Web-based Navigation Platform for the Visualization, Presentation and Retrieval of German Speech Corpora. In prep.

Gasch, Joachim/ Brinckmann, Caren/ Dickgießer, Sylvia (2008): memasysco: XML schema based metadata management system for speech corpora. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko
http://www.lrec-conf.org/proceedings/lrec2008/pdf/729_paper.pdf.

Isenberg, Horst (1984): Texttypen als Interaktionstypen. Eine Texttypologie. In Zeitschrift für Germanistik 5, S. 261-270.

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

Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung. Online-Zeitschrift zur verbalen Interaktion.
http://www.gespraechsforschung-ozs.de/.

URL 1: http://agd.ids-mannheim.de/html/index.shtml [30.11.2009].
URL 2: http://agd.ids-mannheim.de/html/dgd.shtml [30.11.2009].
URL 3: http://agd.ids-mannheim.de/html/folk.shtml [30.11.2009].
URL 4: http://www.ids-mannheim.de/prag/AusVar/Deutsch_heute/ [30.11.2009].
URL 5: http://gais.ids-mannheim.de [30.11.2009].
URL 6: http://agd.ids-mannheim.de/html/korpora/korpus-index.shtml [30.11.2009].
URL 7: http://www.exmaralda.org/partitureditor.html [30.11.2009].
URL 8: http://www.lat-mpi.eu/tools/elan/ [30.11.2009].

URL 9: http://www.praat.org [30.11.2009].
URL 10: http://agd.ids-mannheim.de/html/folker.shtml [30.11.2009].

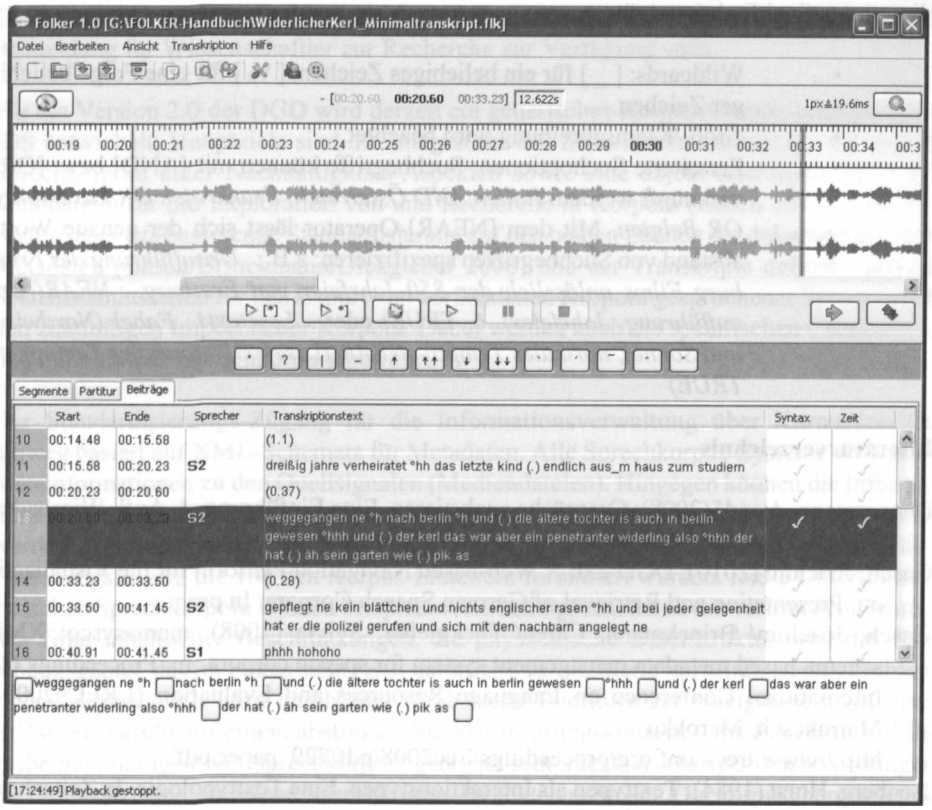

Abb. 1: Beiträge-Ansicht in FOLKER

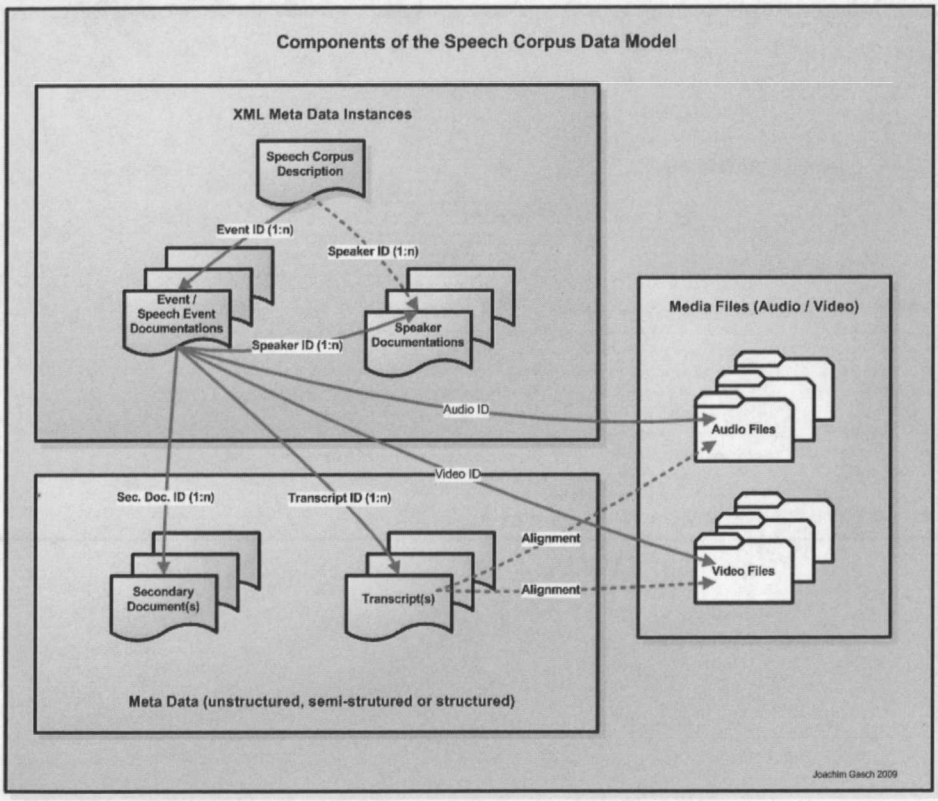

Abb. 2: Datenmodell — Komponenten von Sprachkorpora (Gasch 2009)

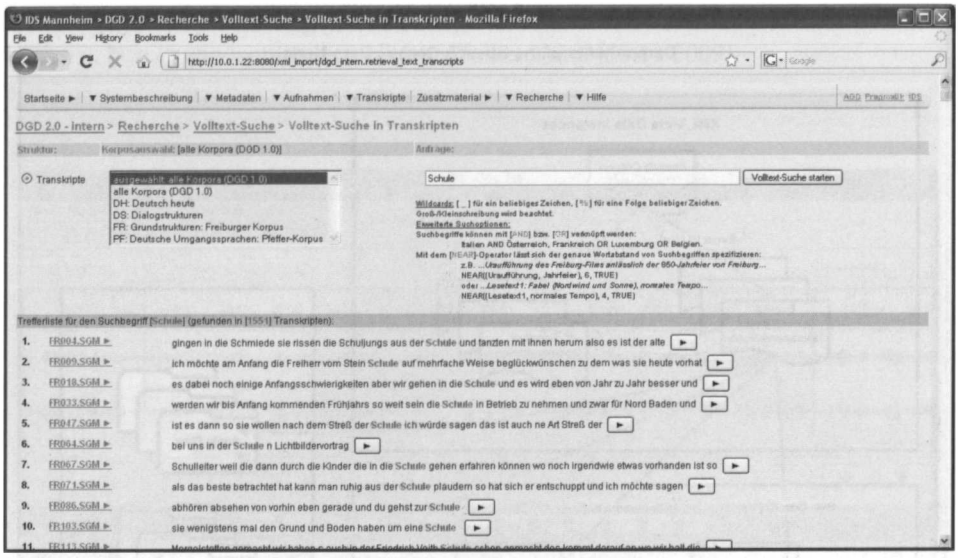

Abb. 3: Volltext-Recherche in der DGD 2.0