Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 6 (2005), Seite 103-126 (www.gespraechsforschung-ozs.de)

Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze

Thomas Schmidt

Abstract
Dieser Aufsatz befasst sich mit Fragen, die sich im Zusammenhang mit der Archivierung und öffentlichen Bereitstellungen von gesprächsanalytischen Daten (Audio- bzw. Videoaufnahmen und deren Transkriptionen) stellen. Er gibt zunächst einen Überblick über die Forschungsperspektiven, die eine verbesserte Praxis der Datenarchivierung für die Gesprächsforschung bieten würde, und nennt dann einige der wesentlichen Probleme, die in der derzeitigen Praxis der Schaffung solcher Archive im Wege stehen können. Anschließend werden vorhandene Lösungsansätze vorgestellt, die helfen können, diese Probleme zu überwinden.

Keywords: Gesprächsforschung, Konversationsanalyse, Diskursanalyse, Datenarchivierung, Korpora, Sprachressourcen.

English Abstract
This paper discusses some questions concerning the archiving and publication of conversation analytical data (audio or video recordings and their transcriptions). It sets out with an overview of the research perspectives that an improved practice of archiving and publishing data would hold for the research community and then summarizes the most important problems, which, at this point in time, may make such an improved practice difficult to realize. The last part of the paper presents some existing proposals for solutions to these problems.

Keywords: Conversation analysis, discourse analysis, data archiving, corpora, language resources.

1. Einleitung

> Vielleicht findet sich ja [...] jemand, der mal die zahllosen so entstandenen Korpora sammelt und für weitere Arbeiten zugänglich macht, damit nicht jeder wieder bei Null anfängt? Oder gibt es so etwas sogar schon in Ansätzen, nur ist es zu wenig bekannt? Es wäre toll, wenn wir da unsere Forschung etwas effektivieren könnten. [Ruth Albert, MG 2003]

Dieses Zitat ist ein Ausschnitt aus einer Diskussion, die im Februar 2003 auf der Mailingliste Gesprächsforschung geführt wurde (MG 2003). Ausgangspunkt der Diskussion war die Beobachtung einer Teilnehmerin, dass viele wissenschaftliche Vorhaben, die sich mit gesprochener Sprache befassen (hier: eine Dissertation zum kindlichen Spracherwerb), darunter leiden, dass ein großer Teil der investierten Zeit und Mühe in die Erhebung und Transkription von Gesprächsdaten fließt, die so entstehenden Korpora aber oft trotzdem nicht umfangreich genug sind, um aussagekräftige Analysen zu ermöglichen. Es wird daher nach den Möglichkeiten gefragt, solche Korpora zu sammeln und der Forschungsgemeinschaft zur Verfügung zu stellen.

Der vorliegende Beitrag versucht, einige Antworten auf diese Fragen zu formulieren. Dass er dabei mit einem Verweis auf eine eher informelle Diskussion und nicht auf eine "reguläre" wissenschaftliche Veröffentlichung beginnt, hat einen guten Grund: obwohl die Frage nach der Erhebung und Aufbereitung von Primärdaten ganz ohne Zweifel eine zentrale Rolle in der gesprächsanalytischen Forschungsarbeit spielt, findet nach meiner Erfahrung die (durchaus lebhafte) Auseinandersetzung mit Problemen der Veröffentlichung, Archivierung oder Wiederverwendung solcher Daten in der Regel außerhalb der offiziellen wissenschaftlichen Kanäle statt. Ich werde mich daher in diesem Beitrag zu einem großen Teil auf Aussagen stützen, die nicht in zitierfähigen Publikationen festgehalten sind, sondern unter den Begriff der "Personal Communication" fallen. Außer der bereits erwähnten Diskussion auf der Mailingliste Gesprächsforschung und der fortdauernden Arbeit an der Datenbank "Mehrsprachigkeit" am SFB 538 in Hamburg sind in diesem Zusammenhang vor allem die folgenden Veranstaltungen zu nennen:

- ein Gastvortrag von Dr. Michael Lautenschlager vom MPI für Meteorologie (Hamburg) zum Thema "Zitierfähigkeit wissenschaftlicher Primärdaten" am SFB 538 (September 2002)
- ein Workshop zum Thema "Multilingual databases" anlässlich des Gastaufenthaltes von Brian MacWhinney am SFB 538 (Juli 2003)
- ein Workshop zum Thema "Heterogenität in linguistischen Datenbanken" am SFB 632 "Informationsstruktur" in Potsdam (Juli 2004, siehe dazu Dipper/Goetze 2005)
- ein Workshop zum Thema "Nachhaltigkeit in der Arbeit mit linguistischen Daten" am SFB 441 "Linguistische Datenstrukturen" in Tübingen (Februar 2005)
- mehrere Workshops und Konferenzen, die sich im Kontext sprachtechnologischer (also gerade nicht gesprächsanalytischer oder 'klassisch' linguistischer) Forschung u.a. mit dem Thema "Primärdaten gesprochener Sprache" befassten. Exemplarisch sei die LREC (Language Resources and Evaluation Conference) genannt, die im Mai 2004 in Lissabon stattfand (siehe Lino et al. 2004)

In diesen Vorträgen und den sie begleitenden Diskussionen habe ich den Eindruck gewonnen, dass die Auseinandersetzung über Datenarchive gesprochener Sprache nicht nur immer um die gleichen Erwartungen kreist, sondern dass sich darüber hinaus auch einige wenige zentrale Probleme herauskristallisieren, die einer Erfüllung dieser Erwartungen im Wege stehen können. Die Abschnitte 2 und 3 rekapitulieren diese Perspektiven bzw. Probleme. In Abschnitt 4 wird dann – gewissermaßen als Versuch einer Antwort auf die im einleitenden Zitat gestellte Frage, ob "es so etwas in Ansätzen schon [gibt] " – eine Auswahl von Konzepten und Initiativen vorgestellt, die Ansätze zur Lösung der sich stellenden Probleme aufzeigen.

Die Behandlung des Themas bleibt aber aus zwei Gründen notwendigerweise unvollständig: zum einen befinden sich die Gebiete, die mit Archiven von Daten gesprochener Sprache zu tun haben, ausnahmslos noch in einer Phase des Experimentierens – von endgültigen und einheitlichen Lösungen für die verschiedenen Teilprobleme sind sie alle noch relativ weit entfernt, folglich können hier auch

keine solchen Lösungen vorgestellt werden. Zum anderen verteilen sich selbst die Lösungs*ansätze* über so viele grundverschiedene Bereiche (Sprachtechnologie, Dokumentation bedrohter Sprachen, Gesprächsforschung, Spracherwerbsforschung, an einer Stelle sogar Geowissenschaften), dass ihre schlüssige Gesamtdarstellung nicht nur den Rahmen dieses Beitrags sprengen würde, sondern vielleicht sogar prinzipiell noch nicht zu leisten ist. Ziel dieses Beitrags ist demnach zunächst, die in den genannten Quellen geführte Diskussion zu dokumentieren und damit eine Basis zu ihrer Fortführung zu schaffen.

2. Ausgangslage

Wenn auch eine Grundannahme diese Beitrags ist, dass Datenarchive für die Gesprächsforschung noch nicht annähernd in dem Ausmaße existieren, wie es wünschenswert wäre, so soll dennoch nicht der Eindruck erweckt werden, dass es noch keine erfolgreiche Unternehmungen auf diesem Gebiet gegeben hätte.

Zunächst ist in diesem Zusammenhang das Repertorium "Deutsche Transkripte" (Glas/Ehlich 2000) zu nennen, das Transkriptbände und Transkripte, die als Bestandteile von Aufsätzen und Monographien veröffentlicht wurden, katalogisiert und damit vielen – aber bei weitem nicht allen – Erwartungen nachkommt, die einem Datenarchiv für die Gesprächsforschung entgegen gebracht werden: das Hauptdefizit eines solchen Repertoriums liegt darin, dass die gesprächsanalytischen Daten, zu denen es Zugang bietet, in der Regel lediglich gedruckte Transkripte sind. Dass diese – wie in der Gesprächsforschung als unbestritten gelten dürfte – bereits eine Vielzahl von theorie- und zielsetzungsabhängigen Abstraktionen gegenüber den eigentlichen Primärdaten (den Aufnahmen[1]) beinhalten, schränkt ihre Wiederverwendbarkeit wesentlich ein: nur ein kleiner Teil der Fragestellungen, die (durch eine geeignete neue Transkription) an der Aufnahme bearbeitet werden könnten, lassen sich auch noch anhand des Transkripts sinnvoll angehen. Darüber hinaus wird eine Wiederverwendung der Daten dadurch erschwert, dass sie nur in gedruckter, nicht aber in digital verarbeitbarer Form verfügbar sind.[2]

Den nicht nur diesbezüglich zeitgemäßen Maßstab setzt die "Datenbank Gesprochenes Deutsch (DGD)" am Institut für Deutsche Sprache in Mannheim (Fiehler 2005). In dieser Datenbank sind die wesentlichen Anforderungen, die gewöhnlich an ein Datenarchiv für die Gesprächsforschung gestellt werden, realisiert: verschiedene Transkriptionskorpora lassen sich computergestützt nach transkribierten Einzelphänomenen abfragen, gleichzeitig ist ein Zugang zu vollständigen Transkriptionen und den zugrunde liegenden Aufnahmen möglich. Insbesondere stehen die so aufbereiteten Korpora ohne größere Hürden der gesamten

---

[1] Wie z.B. Sager (2001:1028f.) darlegt, müssen Aufnahmen streng genommen bereits als Sekundärdaten bezeichnet werden, da "die Realität durch den Aufnahmeprozess in spezifischer Weise verkürzt wird". Transkriptionen wären demnach Tertiärdaten, und die Gespräche selbst die eigentlichen Primärdaten. Da letztere sich aber einer Archivierung generell entziehen, ist es nicht sinnvoll, diese Unterscheidung hier aufrecht zu erhalten.

[2] Das gilt auch für digital publizierte Transkriptbände wie Boettcher et al. (2005), deren Veröffentlichungsformat (in diesem Fall PDF) es nicht erlaubt, die Transkriptionsdaten unmittelbar mit einem Rechner zu bearbeiten (also etwa zu durchsuchen, in einen Editor zu laden etc.).

Forschergemeinschaft zu klar definierten Bedingungen zur Verfügung. Auch wenn sich bezüglich der Details ihrer technischen Realisierung, ihres Aufbaus und ihrer Benutzung sicherlich einige Kritikpunkte formulieren ließen, so hat diese Datenbank für sich genommen dennoch exemplarischen Status – würden alle gesprächsanalytischen Daten in vergleichbarer Form zugänglich gemacht, wäre der dem einleitenden Zitat zugrunde liegenden Kritik der Boden weitestgehend entzogen. Zu fragen ist daher weniger nach Unzulänglichkeiten dieses Unternehmens, sondern nach den Gründen für seine Einzigartigkeit: der wohl wichtigste Grund dürfte die Kontinuität sein, die eine auf Dauer eingerichtete Institution wie das IDS den allermeisten anderen Zusammenhängen, in denen Daten gesprochener Sprache erhoben werden, voraushat. Zwar sind auch die IDS-Korpora in zeitlich befristeten Projekten entstanden, ihre derzeitige Form der Aufbereitung haben sie aber ausnahmslos im Rahmen ihrer Integration in eine dauerhaft fortbestehende, übergeordnete Einrichtung erhalten. Eine wichtige Frage für diesen Beitrag ist daher, mit welchen Mitteln und Konzepten sich eine erfolgreiche Datenarchivierung auch für Projekte, denen ein solcher Überbau fehlt, erreichen lässt.

Die Liste der möglichen Orientierungspunkte ist mit diesen beiden Beispielen sicherlich nicht abgeschlossen. Als weitere Unternehmen, die die Ausgangslage für diesen Beitrag mitbestimmen, könnten z.B. das Bayerische Archiv für Sprachsignale (BAS, Draxler/Schiel 2002), das Kiel Corpus (Simpson et al. 1997) oder prosoDB (Gilles 2001) herangezogen werden. Mehr als durch das Vorhandensein solcher positiven Beispiele wird das Grundproblem, mit dem sich dieser Beitrag auseinandersetzen will, aber durch ein Nichtvorhandensein definiert: Trotz solcher gegenteiliger Belege bleibt nämlich der Normalfall der, dass gesprächsanalytische Daten, nachdem sie erstellt und in einem relativ eng begrenzten Zusammenhang analysiert wurden, der wissenschaftlichen Öffentlichkeit nicht, oder zumindest nicht in einer brauchbaren Form, zur Verfügung stehen, ja dass größtenteils sogar noch nicht einmal verlässliche Informationen über ihre Existenz und ihre Zusammensetzung erhältlich sind.[3] Im Folgenden geht es darum, ob und wie es möglich ist, das Gegenteil zum Normalfall zu machen.

[3] Vgl. dazu auch Glas/Ehlich (2000): "Die Projektgebundenheit vieler Transkriptionen hat zur Folge, dass die erheblichen Arbeitsaufwendungen, die für das Erstellen und Verwalten von Transkripten erforderlich sind, sich häufig im Einzelprojekt erschöpfen. Die einmal gewonnenen Daten könnten aber für viele andere Fragestellungen weitergenutzt werden - wenn nur die Kenntnis über sie besser verbreitet wäre".
Es sei noch angemerkt, dass dies auf die Gesprächsforschung im deutschsprachigen Raum in besonderem Maße zuzutreffen scheint, während andernorts die Veröffentlichung von gesprächsanalytischen Primärdaten u.U. nicht ganz so ungewöhnlich ist. Vergleiche dazu die Anmerkung einer Teilnehmerin der Diskussion auf der Mailingliste "Gesprächsforschung": "Als ich vor einiger Zeit bei [...] in Dänemark war, bat er mich sofort um die Transkription zum gehaltenen Vortrag, denn auf der Seite www.conversation-analysis.net können Transkriptionen herunter geladen werden. In Dänemark scheint dies kein solch großes Problem zu sein und [...] bemerkte, dass es ein spezifisches Phänomen in Deutschland sei, die Daten nicht weiterzugeben" [Kirsten Nazarkiewicz, MG 2003].

3. Perspektiven

> Es ist nun einmal so, dass, wenn jede(r) wieder anfängt, Daten zu erheben, in Anbetracht der zur Verfügung stehenden Zeit nie wirklich aussagefähige Analysen entstehen können. [...] Wenn sich jemand finden würde, der solche Video-/Transkriptsammlungen [...] zusammenstellt und für Forschungszwecke bereitstellt [...], könnten wir erheblich professioneller arbeiten [Ruth Albert, MG 2003].

Ein Datenarchiv für die Gesprächsforschung wäre also eine nach klaren Regeln öffentlich zugängliche Zusammenstellung von digitalen Video- und Audio-Aufnahmen und deren Transkriptionen. Unter den vielfältigen Meinungen, die zum Thema der Datenarchivierung bestehen, ist die, dass solche Archive generell unnütz sind, die am wenigsten verbreitete. Vielmehr gibt es eine ganze Reihe von offensichtlichen Gründen, die es sinnvoll erscheinen lassen, gesprächsanalytische Daten verstärkt für eine Wiederverwendung durch die wissenschaftliche Gemeinschaft zu erschließen. Die wichtigsten davon, die zudem in vielfältiger Weise untereinander verknüpft sind, werden im Folgenden rekapituliert:

I. *Ökonomie*: Unbestreitbar ist das Erheben und Transkribieren von Gesprächsaufnahmen ein Unterfangen, das erhebliche zeitliche und personelle – und damit aus der Perspektive des Geldgebers finanzielle – Ressourcen beansprucht. Verlässliche Zahlen sucht man vergebens; die Annahme, dass in einem typischen gesprächsanalytischen Forschungsprojekt mehr als die Hälfte der Zeit für Datenerhebung und -aufbereitung aufgewendet wird, dürfte aber nicht unrealistisch sein. Bei der Darstellung von Forschungsergebnissen wird dies allerdings in der Regel nicht reflektiert – hier liegt der Schwerpunkt eindeutig auf den Ergebnissen, die *nach* der aufwändigen Erstellung der Korpora gewonnen wurden. Gesprächsanalytische Daten zu archivieren und der wissenschaftlichen Gemeinschaft zur Verfügung stellen, könnte damit nicht nur aus Sicht der Forschenden zu einer erheblichen Zeit- und Arbeitsersparnis führen, es würde auch aus Sicht der Geldgeber das Verhältnis von Korpuserstellung und -auswertung in Einklang mit den publizierten Ergebnissen bringen. Dabei darf der Anspruch natürlich nicht sein, dass ein Forschungsprojekt sich nur noch auf bereits vorhandene Daten stützt: vorhandene Aufnahmen zu verwenden, aber neu zu transkribieren oder ein bestehendes Korpus durch eigene Aufnahmen und Transkriptionen zu ergänzen, bergen im Vergleich zur gängigen Praxis bereits das Potential zu einer ganz erheblichen Einsparung von Zeit- und Arbeitsaufwand.

II. *Quantität*: Zwar spielt in der Gesprächsforschung – im Vergleich etwa zu einer typischen sogenannten 'korpuslinguistischen' Untersuchung oder vielen Spracherwerbsstudien – die Quantifizierung eine vergleichsweise geringe Rolle, weil der sehr detaillierten Analyse von interessanten Einzelbeispielen in der Regel der Vorzug vor einer zahlenmäßigen Auswertung vieler Belege gegeben wird. Dennoch ist die Größe eines Korpus nicht unbedeutend. Zum einen erhöht eine größere Datenmenge die Wahrscheinlichkeit, gesprächsanalytisch fruchtbare Beispiele aufzufinden bzw. verbessert die Möglichkeit zu beurteilen, wie interessant (weil z.B. 'typisch' oder 'ungewöhnlich') ein Einzelbeispiel ist.[4] Zum anderen wird

[4] Genau dieses Problem wurde in der einleitend zitierten Diskussion anfänglich thematisiert.

auch in der Gesprächsforschung stellenweise gefordert, die Signifikanz von Einzelfallanalysen durch quantitative oder gar statistisch beurteilende Verfahren zu belegen,[5] und ein solcher Beleg setzt zwangsläufig eine gewisse 'kritische Masse' an Korpusdaten voraus. Da wegen des bereits beschriebenen zeitlichen und personellen Aufwands die Erstellung großer Korpora ein einzelnes Projekt in der Regel überfordern wird, ist auch hier die nächstliegende Lösung, vorhandene Korpora verfügbar zu machen und für solche quantitativen Verfahren zu größeren Einheiten zu bündeln.

III. *Vergleichbarkeit*: Dass viele gesprächsanalytische Ergebnisse einzigartig bezüglich ihrer Kombination aus Fragestellung und Datenmaterial sind, macht sie untereinander schwerer vergleichbar. Eine Möglichkeit, dies zu umgehen, wäre es, die gleiche Fragestellung anhand verschiedener Korpora zu verfolgen. In welchem Ausmaß diese wahrgenommen wird, vermag ich nicht zu beurteilen. Hinsichtlich der umgekehrten Möglichkeit – verschiedenen Fragestellungen anhand ein und desselben Korpus nachzugehen – sind dem Gesprächsforscher durch die praktischen Gegebenheiten wiederum enge Grenzen gesteckt, denen sinnvoll nur dadurch begegnet werden kann, dass Korpora der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden.

IV. *Methodik*: Auch wenn (oder gerade weil) die Erstellung von Gesprächskorpora sich in den vergangenen Jahren merklich professionalisiert hat (Redder 2002), ist der Bedarf an Richtlinien, Erfahrungsberichten und anderen Entscheidungshilfen, die einen Gesprächsforscher bei der Erstellung eines Korpus unterstützen, nach wie vor groß (vgl. z.B. zahlreiche diesbezügliche Anfragen auf der Mailingliste Gesprächsforschung). Wären mehr gesprächsanalytische Korpora öffentlich verfügbar, so könnte deren Beispiel sicherlich einen wesentlichen Beitrag zur Identifizierung einer 'best practice' der Korpuserstellung leisten und somit zu einem nützlichen Fortschritt auf einem oft vernachlässigten Gebiet der gesprächsanalytischen Methode führen. Dies gilt ganz genauso für die von Gesprächsforschern verwendeten Transkriptionssysteme: würde die Praxis des Transkribierens durch eine große Anzahl öffentlich verfügbarer Transkriptionen (von verschiedenen Personen und verschiedenartigen Aufnahmen) dokumentiert, könnte die Adäquatheit der in den jeweiligen Konventionen formulierten Maximen und Regeln mit wesentlich größerer Verlässlichkeit und Aussagekraft überprüft werden, als dies derzeit der Fall ist.[6] Es ist alles andere als unwahrscheinlich, dass eine solche

---

[5] Z.B. bei Deppermann (1999): "Eine umfassende Darstellung gesprächsanalytischer Methodik müsste allerdings einiges Weiteres enthalten, was hier ausgespart wird. Wichtig wären z.B.: [...] die Verbindung gesprächsanalytischer Methoden mit Quantifikation" und bei Rehbein/Kameyama (2005): "Die Erarbeitung eines umfangreichen Corpus gesprochener Sprache (authentischer Kommunikationsdaten) ist vonnöten, um die leeren Flecken bei der empirischen Begründung einer gesellschaftsbezogenen Diskurs- und Textformenlehre und -klassifikation zu füllen".

[6] In diesem Zusammenhang sei auf die Arbeiten von Sabine Kowal und Daniel O'Connell verwiesen, die die mangelnde empirische Überprüfung der Güte von Transkriptionssystemen bemängeln (z.B. in O'Connell/Kowal 1995:651 – "Essentially, what Du Bois has done is to select a set of signs for notation and declare them good, accessible, robust, economical and adaptable. He provides no empirical evidence that his notation system is preferable to any other with respect to these qualities, nor does he establish that these five qualities are the necessary

Überprüfung Mängel offen legen würde, die als Anlass zu einer Verbesserung der Systeme genommen würde.

IV. *Gute wissenschaftliche Praxis*: Die Empfehlung Nr. 7 der DFG zur guten wissenschaftlichen Praxis (siehe DFG 1998) lautet:

> Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Datenträgern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.

In den zugehörigen Erläuterungen finden sich dann u.a. die folgenden Formulierungen:

> Auf die Aufzeichnungen später zurückgreifen zu können, ist schon aus Gründen der Arbeitsökonomie in einer Gruppe ein zwingendes Gebot. Noch wichtiger wird dies, wenn veröffentlichte Resultate von anderen angezweifelt werden. Daher hat jedes Forschungsinstitut, in dem lege artis gearbeitet wird, klare Regeln über die Aufzeichnungen, die zu führen sind, und über die Aufbewahrung der Originaldaten und Datenträger. [...] Die Berichte über wissenschaftliches Fehlverhalten sind voll von Beschreibungen verschwundener Originaldaten und der Umstände, unter denen sie angeblich abhanden gekommen waren. Schon deshalb ist die Feststellung wichtig, dass das Abhandenkommen von Originaldaten aus einem Labor gegen Grundregeln wissenschaftlicher Sorgfalt verstößt und prima facie einen Verdacht unredlichen oder grob fahrlässigen Verhaltens rechtfertigt.

Zwar ist hier nur von einer adäquaten Archivierung, nicht unbedingt auch von einer öffentlichen Bereitstellung von Primärdaten die Rede, letztere setzt aber erstere voraus, und es darf bezweifelt werden, dass gesprächsanalytische Forschungsprojekte auch nur erstere immer im geforderten Umfang zu leisten imstande sind.[7] In dieser Weise birgt der Mangel an Lösungen zur Archivierung auch die Gefahr, gegen Regeln guter wissenschaftlicher Praxis zu verstoßen; oder positiv formuliert: wenn es zum Normalfall würde, dass gesprächsanalytische Primärdaten archiviert und der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden, würde das einen Nachvollzug oder eine kritische Überprüfung vorhandener Forschungsergebnisse wesentlich erleichtern und die Ergebnisse in dieser Weise aufwerten.

V. *Wissenschaftliche Lehre und Angewandte Gesprächsforschung*: Was bereits oben unter Punkt IV angesprochen wurde, gilt in gleichem Maße für die wissenschaftliche Lehre und für Anwendungen der Gesprächsforschung. Öffentlich verfügbare Gesprächskorpora stellen eine Form der Anschauung dar, die nicht nur für eine Fortentwicklung der Methode fruchtbar gemacht werden kann, sondern auch einem Studierenden der Linguistik oder einem Teilnehmer an einer gesprächsanalytischen Schulung sicherlich ganz andere Perspektiven auf den Gegenstand

---

and sufficient ones for adequate notation") und sie verschiedentlich anhand publizierter Transkriptausschnitte nachzuholen versuchen.

[7] Nach meiner persönlichen Erfahrung ist es alles andere als ungewöhnlich, dass Aufnahmen aus abgeschlossenen Forschungsprojekten entweder nicht mehr auffindbar oder nicht mehr zu gebrauchen sind. Noch viel weniger ungewöhnlich ist es, dass Transkriptionen oder die Detailkenntnis der angewandten Konventionen verloren gehen oder es nicht mehr möglich ist, aus einer Menge verschiedener Versionen ein und derselben Transkription die "gültige" zu identifizieren.

eröffnet, als sie die bloße Rezeption der Ergebnisse und aufbereiteten Einzelbeispiele, die aus solchen Korpora hervorgehen, bietet.[8]

4. Probleme

Obwohl also theoretisch eine Reihe von guten Argumenten dafür sprechen, dass für die Gesprächsforschung erhobene Daten archiviert und der wissenschaftlichen Gemeinschaft zugänglich gemacht werden, geschieht dies in der Praxis nur sehr selten. Mögliche Gründe dafür werden im Folgenden rekapituliert. Dabei wird weitestgehend darauf verzichtet, auf die weiter unten angesprochenen Lösungsansätze vorzugreifen.

I. *Technologische Probleme*: Die Uneinheitlichkeit von Rechnersystemen und der schnelle technologische Wandel machen das Archivieren und Austauschen digitaler Daten ganz allgemein zu einem schwierigen Unterfangen, und es ist keine neue Erkenntnis, dass gesprächsanalytische Daten von diesen Schwierigkeiten ganz besonders betroffen sind. Weil dies an anderer Stelle (z.B. Schmidt 2002) ausführlich besprochen worden ist, soll hier ein kurzer Blick auf die Situation der beiden wichtigsten deutschen gesprächsanalytischen Transkriptionssysteme zur Illustration genügen.

Viele nach HIAT transkribierte Daten liegen entweder als syncWriter- oder als HIAT-DOS-Dateien vor. Diese beiden Formate sind inkompatibel und nicht ausreichend dokumentiert, um geeignete Konvertierungsmechanismen entwickeln zu können. Da außerdem die beiden Programme für verschiedene Betriebssysteme konzipiert sind und nicht mehr weiterentwickelt werden, wären diese Daten, wenn sie in ihrer derzeitigen Form archiviert würden, jeweils nur für einen Teil der wissenschaftlichen Gemeinschaft und wahrscheinlich nur noch für befristete Zeit nutzbar. Nach GAT transkribierte Daten dürften hingegen in der Regel als Text- oder Worddateien vorliegen. Weil Texteditoren zum Grundbedarf jedes Computernutzers gehören, sind sie damit auf den ersten Blick weniger abhängig von einer speziellen Software. Einige Details des Transkriptionssystems (insbesondere die Zeichen zur Markierung prosodischer Information) gehören aber genau in den Bereich, der sich ohne weitere Spezifizierung (z.B. der Zeichenkodierung) nicht zuverlässig zwischen verschiedenen Anwendungen austauschen lässt.[9] Da in GAT

---

[8] Damit soll nicht der gängigen Praxis von Lehre und Anwendung (die ich auch gar nicht umfassend beurteilen kann) ein genereller Mangel an Anschauung unterstellt werden – es scheint mir ganz im Gegenteil durchaus der Normalfall zu sein, dass Studierende, die mit Gesprächsforschung zu tun haben, früh selbst Aufnahmen erheben und transkribieren. Es ist aber nicht von der Hand zu weisen, dass der Schritt von solchen exemplarischen Einzelübungen zur Konfrontation mit einem größeren, "professionellen" Gesprächskorpus oft überhaupt nicht in die Phase des Erststudiums zu fallen braucht oder zumindest erst beim Anfertigen einer Abschlussarbeit (und dann mit einem entsprechenden Mangel an Erfahrung) getan wird. Dies rührt sicherlich nicht etwa daher, dass Korpusarbeit als didaktisch ungeeignet angesehen wird, sondern daher, dass es für die Lehre geeignete (also entsprechend gut dokumentierte und "benutzerfreundlich" aufbereitete) Datensammlungen schlicht nicht gibt.

[9] Bei nicht-deutschsprachigen Transkripten, die von Erweiterungen des lateinischen Alphabets (z.B. Portugiesisch, baltische Sprachen), von nicht-lateinischen Alphabeten (z.B. Kyrillisch oder Griechisch) oder von nicht-alphabetischen Schriftsystemen (z.B. Japanisch, Chinesisch) Gebrauch machen, ist ein solcher Austausch noch problematischer.

dezidiert auf eine solche Spezifizierung verzichtet wurde (das System richtet sich "an Transkribenten, nicht an Computer", Selting et al. 1998:92), existiert also auch hier keine verlässliche Grundlage für eine uneingeschränkt nutzbare Datenarchivierung.[10]

Dass gesprächsanalytische Transkriptionsdaten durch die Uneinheitlichkeit ihrer Datenformate in dieser Weise software- bzw. betriebssystemabhängig sind, ist derzeit wahrscheinlich das drängendste technologische Problem. Hingegen stellt die Speicherung und Verarbeitung digitalisierter Audio- und Videoaufnahmen, die noch vor einem Jahrzehnt nur mit Hilfe besonders ausgestatteter Rechner angegangen werden konnte, heutzutage kaum noch ein substantielle Schwierigkeit dar. Das Problem des Zugangs schließlich, also die Frage, mit welchen technischen Mitteln (Online-Zugang vs. Offline-Distribution, Gestaltung von Benutzer-Schnittstellen, etc.) Forscher Zugriff auf archivierte Daten erhalten können, bleibt solange von nachgeordneter Bedeutung, wie die Archive selbst noch nicht existieren.[11]

II. *Rechtliche und ethische Probleme*:

> Der nächste Grund, den ich vermute, betrifft die Vertraulichkeit der Datenbehandlung, die ich bisher immer bei Aufzeichnung/Aufnahme persönlich, verbindlich zusagen musste. Da fällt es einem schwer, selbst ein nicht zu identifizierendes Transkript aus der Hand zu geben, weil man die weitere Verwendung nicht kontrollieren kann. [Kirsten Nazarkiewicz, MG 2003]

Gesprächsanalytische Aufnahmen sind ihrer Natur nach fast immer sensible Daten, weil ihre Authentizität im Normalfall einhergeht mit einer Nähe zur Persönlichkeitssphäre der aufgenommenen Personen. Es ist daher üblich, vor (oder ggf. unmittelbar nach) der Aufnahme deren Einverständnis zu einer wissenschaftlichen Verwendung der Daten einzuholen. Dabei besteht eine erhebliche Unsicherheit über die Rechtsgültigkeit solcher Einverständniserklärungen, die zum Problem wird, wenn Daten weitergegeben oder veröffentlicht werden sollen.[12] Deren Klärung ist nicht nur insofern problematisch, als sie einen fortgeschrittenen juristischen Sachverstand erfordert, sondern wird auch oft als Gefahr gesehen, "schlafende Hunde zu wecken" (personal communication), denn unter Umständen

---

[10] Wenn ein erfolgreicher Datenaustausch also somit bereits unter dem Dach eines einzigen Transkriptionssystems nicht garantiert ist, ist er über diese Grenze hinweg umso problematischer.

[11] Auch hier ließen sich aber anhand des Beispiels der "Datenbank gesprochenes Deutsch" (s.o.) bereits einige Erkenntnisse gewinnen.

[12] Beispiele für weitestgehend ungeklärte rechtliche Fragen in dieser Hinsicht sind:
- Dürfen Daten auch für andere wissenschaftliche Zwecke als die ursprünglich vorgesehenen verwendet werden (z.B. Wiederverwendung gesprächsanalytischer Daten in der qualitativen Sozialforschung) oder bedarf es dazu einer erneuten Einwilligung der aufgenommenen Personen?
- Kann ein Forscher für eine nicht-vorsätzliche Verletzung von Datenschutzbestimmungen (z.B. wenn er beim Anonymisieren einer Transkription einen Personennamen übersehen hat) haftbar gemacht werden?
- Kann die aufnehmende Person für eine missbräuchliche Verwendung der Daten durch andere haftbar gemacht werden?
- Kann bei Aufnahmen von Kindern die Einverständniserklärung der Eltern bei Volljährigkeit des Kindes ihre Gültigkeit verlieren?

würde sich eine solide rechtliche Grundlage für gesprächsanalytische Aufnahmen als wesentlich restriktiver erweisen als die gängige Praxis.

Dort wo der Schutz der Persönlichkeitssphäre kein Problem ist – also vor allem bei Aufnahmen aus dem öffentlichen Raum (Fernsehen, Radio) –, können es Fragen des Urheberrechts oder Copyrights sein, die einer Veröffentlichung von Daten im Wege stehen.

III. *Methodologische Probleme*: Der Aufwand, den die Aufbereitung, Archivierung und Veröffentlichung von gesprächsanalytischen Daten bedeuten würde, ist nur dann zu rechtfertigen, wenn die so entstehenden Sammlungen für eine große Gruppe von Anwendern nutzbar sind. Neben technologischen (s.o.) sind es vor allem methodologische Probleme, die dem im Wege stehen können. Dies betrifft zunächst die Verschiedenheit der in der Gesprächsforschung verwendeten Transkriptionssysteme: trotz diverser Vereinheitlichungsversuche existieren nach wie vor mehrere solcher Systeme, die jeweils nur in ihrem eigenen Anwenderkreis volle Akzeptanz finden und außerhalb dieses Kreises oft als ungeeignet (weil z.B. "schwer lesbar" oder "unsystematisch", personal communication) für eine gemeinschaftliche gesprächsanalytische Forschung (also u.a. einen Datenaustausch) angesehen werden. Darüber hinaus darf auch bezweifelt werden, dass die einzelnen Systeme innerhalb ihres Anwenderkreises allen Ansprüchen gerecht werden, die man im Kontext eines Datenarchivs vernünftigerweise an sie stellen würde. Fraglich ist insbesondere, ob die von einem Transkriptionssystem gegebenen Regeln soweit eindeutig und verlässlich sind, dass auch bei einem Datenaustausch über kleinere Gruppen hinweg sich so etwas wie ein gemeinsames Verständnis eines Transkripts ergeben kann. Der folgende Ausschnitt aus der eingangs erwähnten Diskussion nährt diese Bedenken:

> Ich vermute, die meisten haben – wie ich – eine Unsicherheit über die 'Korrektheit' der Transkription (in den Augen anderer). Wir haben [...] bei einem Transkriptionsvergleich [...] sage und schreibe so viele Transkriptvarianten wie Personen zustande gebracht [...]. Alle, die wir mit Transkription zu tun haben, werden natürlich auf einer abstrakten Ebene zugeben, dass es das 'korrekte' Transkript nicht gibt, aber im konkreten Fall scheint eine Gesichtsbedrohung damit verbunden zu sein. [Kirsten Nazarkiewicz, MG 2003]

Methodologischer Natur ist schließlich auch die Frage, ob gesprächsanalytische Daten überhaupt sinnvoll von einer anderen als der aufnehmenden Personen genutzt werden können. Hier geht es weniger um die transkribierten Phänomene als um das, was Deppermann (2000) "Ethnographie"[13] nennt, also die Vielzahl von soziokulturellen Variablen, die den Kontext dieser Phänomene bilden. Da es kaum möglich ist, diese Variablen systematisch in ihrer Gesamtheit so zu erfassen, wie sie der mit der Datenerhebung befasste Forscher wahrnimmt, kann argumentiert

[13] "Danach beinhaltet Ethnographie die Arbeit mit unterschiedlichen Datenquellen (wie Interviews, informellen Gesprächen oder schriftlichen und visuellen Dokumenten) und vor allem die teilnehmende Beobachtung im Feld über mehrere Monate. Das Ziel besteht darin, dass der Forscher einen Einblick in möglichst viele Facetten und Situationen eines sozialen Feldes gewinnt. Ethnographisches Arbeiten hat also grundsätzlich einen holistischen Anspruch - es geht darum, die Kultur eines Feldes in ihrer "Gesamtheit", was heißt: in ihren wesentlichen Strukturen, Prozessen und deren Bezügen zueinander in den Blick zu bekommen [...]" (Deppermann 2000).

werden, dass jedem Datenarchiv zwangsläufig Informationen fehlen, die zu seiner sinnvollen Nutzung notwendig wären:

> Eine mögliche Antwort auf Deine Frage, warum das Bedürfnis nach einer Video- und Transkriptsammlung offensichtlich kaum besteht, ist vermutlich der Umstand, dass einem beim 'fremden' Material einfach wichtige Informationen fehlen, die der/die Filmende/Aufnehmende durch diese Arbeit automatisch aufnimmt, nicht immer ausreichend zusätzlich schriftlich dokumentiert bzw. dokumentieren kann – vieles läuft da unbewusst im Umfeld ab [Heidi Abfalterer, MG 2003].

IV. *Wissenschaftlicher Wettbewerb*:

> Der letzte Grund, der mir einfällt, scheint mir zugleich jedoch der stärkste zu sein: Konkurrenz. Viel Arbeit und zum Teil jahrelange Vorbereitung steckt in der 'Ergatterung' von Daten, von der Transkription ganz zu schweigen. Das eigene Korpus ist ein Pfund, mit dem man z.T. [...] lange wuchern muss und will. Das stellt sich nicht so leicht zur Verfügung. [Kirsten Nazarkiewicz, MG 2003]

Diesem Zitat ist eigentlich nichts hinzuzufügen. Die Verfügbarkeit gesprächsanalytischer Daten ist die erste und wichtigste Voraussetzung für eine erfolgreiche wissenschaftliche Arbeit. Weil die Erstellung dieser Daten einen so großen Raum einnimmt, aber in der Regel nicht als eigenständig publizierbare Leistung angesehen wird (s.o.), wird mit der Veröffentlichung von Gesprächskorpora immer auch ein gewisser Wettbewerbsvorteil gegenüber anderen Forschern eingebüßt.

V. *Ökonomie*: Schließlich können Gründe der Zeit- und Arbeitsersparnis, die ja oben bereits als ein Argument *für* eine Veröffentlichung von Gesprächskorpora genannt wurden, genau so gut *gegen* eine solche Veröffentlichung sprechen. Dies ist eine Frage der Perspektive, die zum einen eng verwandt mit dem Problem des wissenschaftlichen Wettbewerbs ist: der Zeit- und Arbeitsaufwand, den Nutzer von öffentlich verfügbaren Gesprächskorpora sparen, ist zu einem nicht unerheblichen Teil genau der Aufwand, den der Urheber der Daten in deren Aufbereitung und Dokumentation sowie in die technische Ausstattung zur Archivierung gesteckt hat, und kann insofern einen Wettbewerbsnachteil mit sich bringen. Zum anderen ist die Tatsache, dass die Bereitstellung eines Korpus für andere überhaupt einen Mehraufwand bedeutet, dass die Daten also in ihrer in Einzelprojekten verwendeten Form in der Regel nicht unmittelbar für andere nutzbar sind, im Wesentlichen eine Konsequenz der oben dargestellten methodologischen Probleme der Datenarchivierung.

5. Lösungsansätze

Im einleitenden Zitat wird gefragt, ob es Lösungen zur nachhaltigen Archivierung gesprächsanalytischer Daten schon gebe, und sie vielleicht nur nicht genug bekannt seien. Es gibt sie, und sie sind – wohl weil sie größtenteils nicht aus der Gesprächsforschung selbst und nicht aus dem deutschsprachigen Forschungsraum stammen – nach meinem Eindruck viel zu wenig bekannt. In diesem Abschnitt sollen deshalb die wichtigsten solcher Lösungsansätze kurz vorgestellt werden.

5.1. Standardisierte Datenformate

Um den oben beschriebenen technologischen Problemen einer Datenarchivierung zu begegnen, existieren bereits viel versprechende und erprobte Konzepte:

Weil sich die technische Umgebungen schnell und in oft unvorhersagbarer Weise ändern, ist es zunächst wichtig sicherzustellen, dass digitale Daten auch dann noch zuverlässig verarbeitet werden können, wenn die Software, mit der sie ursprünglich erstellt wurden, nicht mehr lauffähig ist. An die Stelle software-zentrierter Lösungen sollten daher ganz allgemein datenzentrierte Lösungen treten, die alle relevante Information in den Daten selbst ablegen. Langfristige Nutzbarkeit kann dann durch eine geeignete Dokumentation der Datenformate, durch eine Trennung von Inhalts- und Formaspekt und durch den Rückgriff auf offene Standards sichergestellt werden (vgl. Schmidt 2002). Für textuelle Daten (bzw. textähnliche Daten wie Transkriptionen) sollte dies in der Regel eine Entscheidung für XML und Unicode bedeuten, da andere in Frage kommende Standards entweder nicht offen (z.B. RTF, PDF), weniger verlässlich (z.B. HTML, RTF), weniger flexibel bzw. umfassend (z.B. ANSI-Codepages, HTML) oder weniger weit verbreitet (z.B. SGML) sind. Dass XML mittlerweile in fast alle Bereiche der digitalen Datenverarbeitung Eingang gefunden hat und seine Pflege und Weiterentwicklung als Standard damit im Interesse einer sehr großen Anwenderschaft liegt, stellt eine fast optimale Voraussetzung für eine dauerhafte Datenarchivierung dar. Hinsichtlich geeigneter Formate für digitale Audio- oder Videoaufnahmen ist die Situation ähnlich gut überschaubar: Wer heute Tonaufnahmen als Wave- oder MP3-Dateien und Videoaufnahmen nach den durch die *Moving Pictures Expert Group* (MPEG) erarbeiteten Standards archiviert, dürfte damit sichergestellt haben, dass diese Daten auf absehbare Zeit nutzbar bleiben, denn auch hier besteht weit über die Gesprächsforschung hinaus ein Interesse an der Nutzbarhaltung von Daten in diesen Formaten.

Auf der Ebene der physikalischen Datenbeschreibung lässt sich also bereits eine klare Leitlinien für eine adäquate Archivierung formulieren: bei der Speicherung von Transkriptionen und Aufnahmen sollte von den genannten offenen Standards Gebrauch gemacht werden. Die Gesprächsforschung wäre aufgerufen, diesen Leitlinien zu folgen bzw. die dafür nötigen Voraussetzungen zu schaffen, indem sie insbesondere bei der Formulierung von Transkriptionskonventionen die computerseitige Repräsentation der Daten berücksichtigt.

5.2. Generische Datenmodelle

Einen Schritt weiter – also etwa unter der Voraussetzung, dass eine Spezifikation existiert, wie HIAT-, GAT- oder DIDA-Transkriptionen als Unicode-kodierte XML-Dateien zu repräsentieren sind – stellt sich die Frage, wie ein Austausch zwischen verschiedenen solchen Formaten zu bewerkstelligen ist. Wie an anderer Stelle (z.B. Schmidt 2005) ausgeführt, bildet eine XML-basierte Datenbeschreibung alleine noch keine hinreichende Basis für einen solchen Austausch, weil inhärente Eigenschaften von Daten gesprochener Sprache (insb. zeitlich parallele Relationen) es notwendig machen, über das kanonisch mit XML assoziierte OHCO-Datenmodell hinauszugehen. Um es beispielsweise möglich zu machen, sowohl HIAT- als auch GAT-Daten in einer gemeinsamen Software-Umgebung

(z.B. einem Transkriptionseditor oder einer Schnittstelle zur Korpusabfrage) nutzen zu können, bedarf es daher übergeordneter Lösungen, die in der Texttechnologie unter den Bezeichnungen "generische Datenmodelle", "Frameworks" oder "Modellierungsparadigmen" firmieren. Die wichtigste Leistung solcher Datenmodelle besteht darin, dass sie in der notwendigen Heterogenität von digitalen Gesprächsdaten einen gemeinsamen homogenen Kern identifizieren, der zwischen verschiedenen Datenformaten und Anwendungen vermitteln kann.[14] Verschiedene solcher Datenmodelle sind vorgeschlagen worden, z.B.:

- der Annotationsgraphenformalismus (Bird/Liberman 2001)
- andere (weniger mächtige und komplexe) zeitbasierte Datenmodelle: EXMARaLDA/Transkriptionsgraphen (Schmidt 2005), TASX (Milde/Gut 2001), ELAN (Brugman/Russel 2004)
- das NITE Object Model (Evert et al. 2003)
- Das Diskurs(transkript)modell des IDS (Bodmer et al. 2002)[15]

Die Entwicklung solcher Datenmodelle ist zweifelsohne nicht Aufgabe der Gesprächsforschung, sondern der Texttechnologie. Hinsichtlich der technologischen Probleme, die sich bei der Archivierung stellen, wäre es aber wünschenswert, dass gesprächsanalytische Daten sich in wenigstens eines dieser Modelle integrieren lassen. Dies könnte durch das Heranziehen texttechnologischer Expertise beim oben angesprochenen Entwurf der Datenformate sicher gestellt werden.

5.3. Nine access levels

Hinsichtlich der Probleme des Daten- und Personenschutzes gilt es zunächst, ein weit verbreitetes Missverständnis aus der Welt zu räumen: Wenn das WWW als Plattform für eine Archivierung und Distribution von gesprächsanalytischen Daten genutzt wird, so muss dies nicht gleichbedeutend mit einem unbeschränkten und unkontrollierten Zugang für jedermann sein. Zwar ist es wünschenswert, Informationen über das Vorhandensein und die Zusammensetzung von Gesprächskorpora zunächst mit möglichst wenig praktischen Hürden bereitzustellen, und es ist schwer vorstellbar, dass eine diesbezügliche Lösung nicht an irgendeiner Stelle eine Website beinhaltet. Sensible Daten können dabei aber zum Beispiel mit einem Passwort geschützt werden oder müssen gar überhaupt nicht online verfügbar sein, solange sich ihre Existenz über das WWW recherchieren lässt. Sinnvollerweise wird man also gesprächsanalytische Daten für die Veröffentlichung auf einem Spektrum zwischen unbeschränktem und maximal kontrolliertem Zugang einordnen. In MacWhinney (2001) werden dafür neun verschiedene Ebenen vorgeschlagen, z.B.:

---

[14] In den Begriffen der Datenbanktheorie gehören sie damit auf die konzeptuelle bzw. logische Ebene, während etwa XML-basierte Dateiformate der physikalischen und Software-Anwendungen der Anwendungsebene zuzuordnen sind.

[15] Hier sind die in publizierter Form zugänglichen Informationen sehr spärlich, was umso bedauerlicher ist, als es sich bei diesem Datenmodell um eine vornehmlich für die Gesprächsforschung konzipierte Lösung handelt.

> Level 1: Data are fully public (public speeches, public interviews, etc.) and generally viewable and copyable over the Internet, although they may still be copyrighted. [...] Level 5: Access is restricted to researchers who have signed non-disclosure forms. In addition, copying is disallowed. [...] Level 7: This level would only allow viewing and listening in controlled conditions under direct on-line supervision. This level is needed for data of a highly personal or revealing nature. This level has been used in the past for the viewing of material from psychiatric interviews. [...]

Übertragen auf gesprächsanalytische Daten könnten danach beispielsweise Aufnahmen aus dem öffentlichen Rundfunk (Schmitt 2003 oder Barth-Weingarten 2003) dem Level 1, Aufnahmen aus einem Kommunikationstraining (Hablützel 2002) dem Level 5 und sehr sensible Daten wie die Aufnahme eines Arzt-Patienten-Gesprächs (Meyer 2003) dem Level 7 zugeschlagen werden. Für Transkriptionen könnte analog verfahren werden, wobei die Möglichkeit einer vollständigen Anonymisierung hier in der Regel nur eine niedrige Sicherheitsstufe notwendig machen sollte.

Die Einführung solcher Zugangsbestimmungen dürfte dabei gegenüber den derzeit gebräuchlichen Datenschutz-Zusicherungen an die aufgenommenen Personen gar keine einschneidende Änderung bedeuten. Entscheidend wäre vielmehr, dass die Bedingungen, zu denen Transkriptionen und Aufnahmen für dritte verfügbar (bzw. nicht verfügbar) sind, von Vorneherein klar definiert und die oben angesprochenen Unsicherheiten hinsichtlich rechtlicher Verantwortung etc. damit weitestgehend beseitigt wären.

5.4. Zitierfähigkeit wissenschaftlicher Primärdaten

Nicht nur in der Gesprächsforschung sind Primärdaten eine für die Forschungsarbeit so wertvolle Ressource, dass ihre Urheber sich aus Gründen des wissenschaftlichen Wettbewerbs mit einer Veröffentlichung schwer tun. Das folgende Zitat aus einem Bericht einer CODATA[16]-Arbeitsgruppe deutet an, dass sich ähnliche Probleme über alle Disziplingrenzen hinweg stellen:

> In principle, scientists are prepared to provide data, but for the time being it is unusual to appreciate the necessary extra work for processing, context documentation and quality assurance. [...] Project data is widely spread among research institutes and is collected and governed by scientists. Due to the lack of acknowledgement of this extra work, project data is often poorly documented, therefore badly accessible and not maintainable over long time periods. Large amounts of data are unused as they are only known and accessible to a small group of scientists.

Als Grundproblem wird dabei ein "lack of acknowledgment" identifiziert, also eine mangelnde Würdigung für die wissenschaftliche und organisatorische Arbeit, die in die Aufbereitung von Primärdaten gesteckt werden muss, damit sich diese von Dritten nutzen lassen. Dies äußert sich vor allem darin, dass nicht festgelegt ist, ob und wie solche Daten bei einer Wiederverwendung zitiert werden sollen. Wissenschaftliche Primärdaten zitierfähig zu machen, ist daher ein entscheidender Schritt, um ihre Veröffentlichung für den Urheber attraktiv zu machen. Die CODATA-Initiative schlägt vor, die Zitierfähigkeit über einen so genannten Pri-

[16] Committee on Data for Science and Technology – www.codata.org.

märdaten-DOI (Document Object Identifier), vergleichbar etwa einer ISBN-Nummer für gedruckte Publikationen, zu ermöglichen, der von einer zentralen Agentur vergeben und verwaltet wird. Bei aller Professionalität hat diese Lösung jedoch den Nachteil, dass sie organisatorische Infrastrukturen – etwa eine gemeinsame Vertretung für die gesamte Disziplin in der DOI-Agentur – voraussetzt, die in der Gesprächsforschung nicht unbedingt gegeben sind. Besser geeignet, weil mit weniger bürokratischem Überbau verbunden, mag daher eine Lösung sein, wie sie im Rahmen der CHILDES-Datenbank bereits seit langem erfolgreich praktiziert wird: hier wird einfach gefordert, dass jeder Nutzer öffentlich bereitgestellter Daten eine bestimmte Veröffentlichung von deren Urheber zitiert.[17] Idealerweise ist dies eine Veröffentlichung, die das Korpus selbst beschreibt, zumindest aber eine auf ihm basierende Untersuchung zum Thema hat. Diese Form der Würdigung geht also vollständig in der üblichen Zitierpraxis auf und sollte somit auch in gewohnter Weise von der wissenschaftlichen Gemeinschaft kontrolliert werden können.

5.5. Infrastrukturen

Wie oben kurz am Beispiel der 'Datenbank gesprochenes Deutsch' erläutert, kann es eine wesentliche Voraussetzung für die Archivierung von Primärdaten sein, dass sich eine geeignete übergeordnete und dauerhafte Institution um die Pflege und Bereitstellung der Daten kümmert. Eine Organisation, die sich explizit mit Belangen der Gesprächsforschung in dieser Hinsicht befasst, gibt es bislang nicht, deshalb mag es nützlich sein, einen Blick auf vorhandene Infrastrukturen in anderen (bzw. übergeordneten) Gebieten zu werfen.

5.5.1. LDC und ELRA

Das Linguistic Data Consortium (LDC) an der University of Pennsylvania beschreibt sich als "an open consortium of universities, companies and government research laboratories [which] creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes". Es unterstützt Forscher und Entwickler bei der Bereitstellung und Nutzung von Sprachressourcen. Dabei erfolgt zunächst keine Festlegung auf einen bestimmten Adressatenkreis oder einen bestimmten Datentypus – ein Blick auf den Katalog der angebotenen Ressourcen macht aber deutlich, dass das LDC derzeit in erster Linie für die sprachtechnologische Forschung und Anwendung von Interesse ist: ein Großteil der Daten (insgesamt etwa 300 Korpora, Lexika, Datenbanken etc.) sind Aufnahmen und Transkriptionen, die beispielsweise dem Trainieren und Evaluieren von Spracherkennungstechnologie dienen,[18] und die für die Gesprächsforschung nur von geringem Interesse sein dürften. Dennoch finden sich in den LDC-Beständen auch vereinzelt Korpora, die gesprächsanalytischer

[17] Genauer heißt es unter *Ground rules*: "All articles using TalkBank or CHILDES corpora should cite the references for those corpora as listed in the documentation. This is crucial!"

[18] Ein typisches solches Korpus besteht zum Beispiel aus einer großen Anzahl von Aufnahmen, in denen verschiedene Sprecher Ziffernfolgen, Ortsnamen, Menü-Befehle o.Ä. aussprechen. Es handelt sich also nicht um authentische Interaktionsdaten im Sinne der Gesprächsforschung.

Forschung entstammen oder zumindest für eine solche verwendet werden könnten, z.B.:

- Das Santa Barbara Corpus of Spoken American English, das auf Initiative der University of Santa Barbara unter der Leitung von John DuBois erhoben wurde und damit eine offensichtliche Nähe zu konversationsanalytischer Forschung aufweist.
- Das SLX Corpus of Classic Sociolinguistic Interviews (Labov et al. 1972), das aus den Aufnahmen und Transkriptionen der soziolinguistischen Interviews von William Labov besteht.
- Das HCRC Map Task Corpus (Isard 2001), ein Korpus zur aufgabenorientierten Kommunikation, das an der University of Edinburgh erhoben wurde.[19]

Die European Language Resource Association (ELRA) mit Sitz in Paris ist gewissermaßen das europäische Gegenstück zum US-amerikanischen LDC. Die dort angebotenen Daten umfassen daher einerseits eine wesentlich größere Sprachenvielfalt, andererseits weisen sie aber eine noch deutlichere Ausrichtung auf sprachtechnologische Anwendungen auf. Insbesondere entstammen hier viele Anbieter nicht der universitären Forschung, sondern den Forschungs- und Entwicklungsabteilungen von Wirtschafts-Unternehmen. Trotzdem versteht sich auch die ELRA als Ansprechpartner für Sprachdatenarchivierung und -distribution im universitären Umfeld, und für einige wenige der angebotenen Ressourcen – z.B. das Spoken Dutch Corpus und das Basque Spoken Corpus – ist zumindest vorstellbar, dass sie für die Gesprächsforschung von Interesse sein könnten.

Wenn somit vielleicht zweifelhaft bleibt, ob LDC und ELRA geeignete Organisationen für eine Datenarchivierung in der Gesprächsforschung wären, so kann ihr Beispiel doch immerhin eine recht konkrete Vorstellung davon vermitteln, wie eine solche Infrastruktur aussehen könnte.

5.5.2. OLAC

Einen anderen Ansatz zur Schaffung einer Infrastruktur für die Archivierung und den Austausch von Sprachdaten verfolgt die *Open Language Archive Community* (OLAC). Sie definiert sich als "an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources". Anders als LDC und ELRA, die als zentrale Institutionen für die Katalogisierung und Distribution von sprachlichen Daten fungieren, versteht sich die OLAC eher als ein dezentrales offenes Netzwerk von Personen und Organisationen, die an solchen Daten interessiert sind.[20] Koordiniert wird dieses Netzwerk daher nicht von einer übergeordneten Anlaufstelle, sondern über eine Website sowie verschiedene Arbeitsgruppen, Workshops und Konferenzen. Die bislang über die OLAC zugänglichen Daten sind wesentlicher diverser als bei LDC und ELRA: sie umfassen nicht nur Kor-

[19] Siehe auch http://www.hcrc.ed.ac.uk/maptask/.

[20] Tatsächlich gehören auch LDC und ELRA zur OLAC, was deutlich macht, dass es sich hierbei nicht um konkurrierende Unternehmungen handelt.

pora geschriebener und gesprochener Sprache, sondern auch Software-Kataloge, thematische (Online-)Bibliographien und Ähnliches. Im Vergleich zu den Beständen von LDC und ELRA spielen dabei Ressourcen aus und für die universitäre Forschung eine deutlich größere Rolle. Zum gegenwärtigen Zeitpunkt wird man zwar auch dort vergeblich nach Daten suchen, die von speziellem Interesse für die Gesprächsforschung sind. Die Initiative lässt jedoch keinen Zweifel daran, dass sie für solche Daten offen ist und sie gleichberechtigt neben die vorhandenen stellen würde:

> The Open Language Archives Community seeks to embrace all members of the language resources community, from well-established institutional archives to individuals who want to share research results. Would-be participants thus operate in a wide variety of circumstances.

5.6. Best Practice-Richtlinien

Die bisher genannten Lösungsansätze befassen sich jeweils nur mit einem *Teilaspekt* der im vorigen Abschnitt thematisierten Probleme. Abschließend seien daher noch zwei Arbeiten angesprochen, die sich in exemplarischer Weise mit der *Gesamtheit* der Schwierigkeiten befassen, die sich im Hinblick auf Archivierung und Austausch von Sprachdaten stellen.

5.6.1. Seven Dimensions of Portability for Language Documentation and Description [21]

Ziel der Ausführungen von Bird/Simons (2002) ist es, Probleme zu identifizieren, die sich hinsichtlich der Portabilität von Sprachdokumentationen und -beschreibungen[22] stellen, sowie Vorschläge zu einer 'best practice' zu formulieren, mit denen diesen Problemen begegnet werden kann. Sie arbeiten zunächst sieben Problembereiche heraus, die jeweils in weitere Teilaspekte zerfallen:

(1) *Content*: Hier geht es um *inhaltliche* Eigenschaften der Daten selbst. Als potentiell problematisch für einen Datenaustausch wird erstens die Abdeckung (*coverage*) einer Ressource eingestuft, also ihre Vollständigkeit im Hinblick auf eine

[21] Es scheint mir weder möglich noch sinnvoll, hier eine vollständige Paraphrase des gesamten Papiers zu versuchen, zum einen, weil es an Klarheit und Verständlichkeit kaum zu überbieten ist, zum anderen, weil es sich als so einflussreich für die gegenwärtige Diskussion erwiesen hat, dass die Lektüre des Originals für jeden am Thema interessierten Forscher ohnehin unumgänglich sein sollte. Stattdessen sei hier zunächst ein grober Überblick über die dargestellten Ideen gegeben, der deutlich macht, dass Bird/Simons (2002) einerseits die hier bereits behandelten Aspekte vollständig berücksichtigen, andererseits aber auch viele weitere essentielle Fragen thematisieren. Weiter unten werden dann zwei Bereiche exemplarisch im Detail dargestellt.

[22] Wie die Wortwahl andeutet, ist der Ausgangspunkt der Überlegungen dabei die Praxis der "field linguistics", insbesondere vor dem Hintergrund der Dokumentation bedrohter Sprachen. Daher werden an einigen Punkten Probleme angesprochen, die für die Gesprächsforschung eher von untergeordnetem Interesse sind. Zu einem weitaus größeren Teil sind die Ausführungen jedoch unmittelbar für die Situation der Gesprächsforschung relevant und auch problemlos auf diese übertragbar.

gegebene Zielsetzung. Zweitens und drittens werden *accountability* und *terminology* als Problembereiche identifiziert, die beide die Möglichkeit betreffen, Interpretationen und Kategorien, die für die Erstellung der Daten eine Rolle gespielt haben, bei deren späterer Nutzung nachvollziehen zu können.

(2) *Format*: Dieser Punkt betrifft *äußerliche* Eigenschaften der Daten, insbesondere die weiter oben thematisierte technologische Frage einer geeigneten computerseitigen Repräsentation (*openness*, *encoding*, *markup*), aber auch die Frage, wie aus solchen Daten eine für den Menschen optimierte Darstellung abgeleitet werden kann (*rendering*).

(3) *Discovery*: Hierbei geht es darum, wie ein Forscher Kenntnis über eine zum Austausch angebotene Sprachressource erlangen kann. Zum einen muss sichergestellt sein, dass er sich über ihr Vorhandensein (*existence*) informieren kann, zum anderen, dass er möglichst einfach ermitteln kann, ob sie für seine Zwecke geeignet ist (*relevance*).

(4) *Access*: Unter diesem Stichwort werden Probleme des Zugangs zu den Daten behandelt. Dies betrifft erstens die Reichweite des Zugangs (*scope of access*), also z.B. die Frage, ob ein Korpus als Ganzes erworben werden kann, oder ob sich der Zugang etwa auf die Ergebnisse beschränkt, die man über die Nutzung restriktiver Abfrageschnittstellen erhält.[23] Zweitens und drittens fällt unter diesen Punkt die Notwendigkeit, den Zugangsprozess (*process for access*) so zu gestalten und zu beschreiben, dass die Nutzung einer Ressource nicht mit unnötigen Schwierigkeiten verbunden ist (*ease of access*).

(5) *Citation*: Unter *Citation* wird zum einen das thematisiert, was weiter oben unter den Stichworten "Wissenschaftlicher Wettberwerb" und "Zitierfähigkeit" angesprochen wurde. Bemerkenswerterweise betrifft dabei nur einer von vier Unterpunkten (*bibliography*) das Interesse des Autors einer Sprachressource, bei deren Nutzung angemessen zitiert zu werden. Die übrigen drei Punkte befassen sich mit der umgekehrten Perspektive, d.h. mit Anforderungen, die ein Zitierender an ein Datenarchiv stellen muss. Dies beinhaltet erstens die dauerhafte Gültigkeit (*persistence*) einer Zitierform, die beispielsweise eine Web-Adresse nicht zu garantieren vermag. Zweitens werden die Unveränderlichkeit (*immutability*) und drittens die Granularität (*granularity*) einer Ressource als Faktoren genannt, die aus Sicht des Zitierenden die Nennung einer Quelle problematisch machen können.

(6) *Preservation*: Unter dieser Überschrift werden Kernprobleme des Archivierungswesens selbst thematisiert. Langlebigkeit (*longevity*) der archivierten Materialien und ihre Sicherheit (*safety*) gegenüber unvorhergesehenen Ereignissen (wie Brände, Naturkatastrophen) spielen bei digitalen Archiven eine nicht minder wichtige Rolle als in traditionellen Bibliotheken. Von zusätzlicher Bedeutung sind bei digitalen Beständen Entscheidungen über geeignete Speicherformen (*media*).

---

[23] Dass solche Abfrageschnittstellen den Zugang zum Korpus beschränken, muss dabei nicht unbedingt gewollt sein, sondern kann aus einer wohlgemeinten Bemühung um "Benutzerfreundlichkeit" resultieren.

(7) *Rights*: Für den Bereich der rechtlichen und ethischen Fragen identifizieren Bird/Simons die adäquate Formulierung von Nutzungsbedingungen (*terms of use*) für eine Sprachressource als das zentrale Problem. Wie auch oben schon angesprochen, muss dafür vor allem ein ausgewogener Kompromiss (*balance*) zwischen einem Nutzen (*benefit*) für die interessierte Gemeinschaft und einem Schutz von Persönlichkeitsrechten (*sensitivity*) gefunden werden.

Nach diesem umfassenden Überblick formulieren Bird/Liberman für jeden der genannten Punkte eine so genannte "Best Practice Recommendation", also eine Empfehlung, welche Praxis im Bezug auf die einzelnen Problemfelder zum gegenwärtigen Zeitpunkt als die beste anzusehen ist[24]. Zwei dieser Empfehlungen seien hier exemplarisch besprochen:

> RELEVANCE: We value the ability of any potential user of a language resource to judge its relevance without first having to obtain a copy. Thus the best practice is one that makes it easy for anyone to judge the relevance of a resource based on its description.

Dieser Empfehlung, die in den Bereich *Discovery* fällt, liegt die Erkenntnis zugrunde, dass ein ungeleiteter Zugang zu einem unbekannten Sprachkorpus in der Regel sehr mühsam ist. Nur anhand der Daten selbst lassen sich Zusammensetzung, Qualität und Umfang eines Korpus nur sehr aufwändig ermessen, selbst wenn diese Daten umfangreiche Meta-Informationen einschließen. Zu einer adäquaten Veröffentlichung einer Ressource würden demnach immer auch eine übersichtsartige Beschreibung und ein kommentiertes Beispiel aus dem Korpus gehören, die einem potentiellen Nutzer möglichst mühelos zu beurteilen erlaubt, ob die Ressource für seine Zwecke interessant ist.

> IMMUTABILITY: We value the ability of users to cite a language resource without that resource changing and invalidating the citation. Thus the best practice is one that makes it possible for users to cite particular versions that never change.

In dieser Empfehlung (aus dem Bereich *Citation*) wird ein Problem der Zitierfähigkeit eines Primärdatums thematisiert. Das Zitieren einer Ressource ist nur dann sinnvoll, wenn die Quelle Dritten in derselben Form zur Verfügung steht wie dem Zitierenden. Gerade bei digitalen Transkriptionskorpora ist jedoch die Gefahr gegeben, dass durch die leicht möglichen und oft auch notwendigen nachträglichen Korrekturen oder Erweiterungen eine neue Fassung eine ältere in einer für Außenstehende nicht nachvollziehbaren Weise ersetzt. Um die Zitierfähigkeit sicherzustellen, sollten solche Korpora deshalb möglichst in kontrollierten Versionen (wie das z.B. von Software bekannt ist) veröffentlicht werden, wobei jede einzelne Fassung für sich dann archiviert und somit zuverlässig zitierbar würde.

---

[24] Dabei weisen Sie darauf hin, dass eine solche Empfehlung untrennbar davon abhängt, welche Ziele man als erstrebenswert ansieht, und es deshalb im Hinblick auf eine offene Diskussion sinnvoll ist, sie in eine Formel der Form "If you value X, then A is better than B." zu bringen.

5.6.2. Guide Corpus Oraux

Als letzter Lösungsansatz sei hier der "Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux"[25] erwähnt, der im Mai 2005 von einer Gruppe französischer Forscher im Auftrag des Centre National de la Recherche Scientifique in einer ersten Version veröffentlicht wurde (Baude et al. 2005). Dieser ist insofern dem im vorigen Abschnitt beschriebenen Aufsatz vergleichbar, als er eine Übersicht über bestehende Probleme im Umgang mit Korpora gesprochener Sprache gibt und auf dieser Basis Richtlinien für eine 'best practice' formuliert. Indem er sich zunächst auf die französische Forschungslandschaft und auf französischsprachige Korpora beschränkt, erhebt er einerseits jedoch keinen vergleichbaren Anspruch auf Allgemeingültigkeit, kann dafür aber andererseits viele Probleme wesentlich detaillierter und fokussierter darlegen. Das Dokument behandelt auf insgesamt über 100 Seiten in sehr ausführlicher Form zahlreiche rechtliche, technologische und methodologische Fragen, die sich bei der Arbeit mit Korpora gesprochener Sprache stellen. Eine vollständige Inhaltsangabe würde daher den Rahmen dieses Beitrags sprengen. Es sei stattdessen hier einfach der Vorbildcharakter betont, der dieses Unternehmen für die deutsche Gesprächsforschung haben könnte; dies umso mehr, als die Bibliographie des Leitfadens bereits zahlreiche gesprächsanalytische Arbeiten aus dem deutschsprachigen Raum berücksichtigt.

6. (Vorläufiges) Fazit

Dieser Beitrag hat versucht, die wichtigsten Perspektiven, die Datenarchive für die Gesprächsforschung bieten, sowie die Probleme, die sie mit sich bringen, darzulegen. Wie bereits oben angedeutet, kann sein Fazit beim derzeitigen Stand der Dinge nur ein vorläufiges sein.

Zunächst ist festzuhalten, dass kaum eines der in Abschnitt 3 angesprochenen Probleme gänzlich ohne Lösungsansatz bleibt. Den technologischen Problemen kann durch eine geeignete Nutzung offener Standards und generischer Datenmodelle entgegengetreten werden; die rechtlichen und ethischen Probleme werden im Modell der "Nine Access Levels" angegangen; die Probleme des wissenschaftlichen Wettbewerbs finden in den Überlegungen zur "Zitierfähigkeit wissenschaftlicher Primärdaten" Beachtung; Infrastrukturen aus anderen sprachwissenschaftlichen Bereichen geben ein geeignetes Vorbild für die Gesprächsforschung ab; Vergleichbares leisten auch Best-Practice-Richtlinien aus dem nicht-deutschsprachigen Raum. Wären diese Lösungsansätze in der deutschen Gesprächsforschung hinreichend bekannt (nach meinem Eindruck sind sie es nicht), wäre bereits eine sehr solide Basis gegeben, aus der sich für die gesprächsanalytische Forschung optimierte Methoden der Datenarchivierung entwickeln ließen.

Dabei scheinbar unberücksichtigt bleiben jedoch zunächst Probleme, die oben unter dem Stichwort der "Methodik" dargestellt wurden. Dies ist einerseits insofern kaum verwunderlich, als methodische Fragen besonders eng an die Einzelwissenschaften gebunden sind und folglich nicht erwartet werden kann, dass dies-

[25] "Leitfaden guter Praxis für die Erstellung, Nutzung, Erhaltung und Distribution von Korpora gesprochener Sprache."

bezügliche Lösungsansätze z.B. aus der Sprachtechnologie oder der Spracherwerbsforschung unmittelbar auch für die Gesprächsforschung fruchtbar gemacht werden können. Andererseits mögen diese methodischen Probleme weniger ein Hindernis auf dem Weg zu Datenarchiven für die Gesprächsforschung sein, als die Datenarchive ein Weg zur Lösung solcher Probleme sind. Wenn heute eine Unsicherheit über die Validität von Transkriptionen oder von ethnographischen Meta-Daten besteht, so dürfte es kaum einen besseren Weg zu deren Überwindung geben, als die gängige Praxis in Form öffentlicher Archive zu dokumentieren und in dieser Weise einer Diskussion und Weiterentwicklung zugänglich zu machen.

Nicht zuletzt aus diesem Grund kann allerdings die Einrichtung und Betreibung eines Datenarchivs nicht die Aufgabe eines Einzelnen sein, der "die zahllosen so entstandenen Korpora sammelt und für weitere Arbeiten zugänglich macht" (vgl. einleitendes Zitat). Mehr noch als der zeitliche Aufwand einer solchen Arbeit macht die Tatsache, dass die technologischen. methodischen, infrastrukturellen etc. Probleme der Datenarchivierung einer andauernden Diskussion bedürfen, es notwendig, dass eine solche Anstrengung von der wissenschaftlichen Gemeinschaft ausgeht und getragen wird.

Nach alldem bietet das Feld der Archivierung gesprächsanalytischer Korpora derzeit wahrscheinlich eine ideale Gelegenheit zu Pionierarbeit. Ein gesprächsanalytisches Korpus von nicht mehr als zehn Aufnahmen und Transkriptionen, das unter größtmöglicher Berücksichtigung der hier angeführten Lösungsansätze der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt wird, dürfte zum gegenwärtigen Zeitpunkt bereits eine konkurrenzlose Leistung sein. Die Aufmerksamkeit, die es (und die zugehörigen gesprächsanalytischen Arbeiten) erhalten würde, sollte eventuelle Wettbewerbs-Nachteile, die sich der betreffende Forscher durch die 'Preisgabe' seiner Daten einhandelt, mehr als aufwiegen. Gleichzeitig würde ein solches Beispiel die Grundlage für weitere Entwicklungen, deren Darstellung in diesem Beitrag notgedrungen abstrakt geraten ist, wesentlich besser greifbar machen.

7. Literatur

Barth-Weingarten, Dagmar (2003): Zur (Aus-)Nutzung konzessiver Konstruktionen in radio interviews. Eine qualitativ-quantitative Untersuchung zur Kontextabhängigkeit von Äußerungen. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (4), 251-281.

Baude, Olivier / Blanche-Benveniste, Claire / Calas, Marie-France / Cordereix, Pascal / De Lambertierie, Isabelle / Goury, Laurence / Jacobson, Michel / Marchello-Nizia, Christiane / Mondada, Lorenza (2005): Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux. Paris: Centre National de la Recherche Scientifique.

Bird, Steven / Liberman, Mark (2001): A formal framework for linguistic annotation. In: Speech Communication 33(1,2), 23-60.

Bird, Steven / Simons, Gary (2002): Seven Dimensions of Portability for Language Documentation and Description. In: Language 79, 557-582.

Bodmer, Franck / Fach, Marcus / Schmidt, Rudolf / Schütte, Wilfried (2002): Von der Tonbandaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für

die Arbeit mit Gesprächskorpora. In: Pusch, Claus / Raible, Wolfgang (Hg.): Romanistische Korpuslinguistik. Korpora und Gesprochene Sprache. Tübingen: Narr, 209-243.

Boettcher, Wolfgang / Limburg, Anika / Meer, Dorothee / Zegers, Vera (2005): "Ich komm (0) weil ich wohl etwas das thema meiner hausarbeit etwas verfehlt habe". Sprechstundengespräche an der Hochschule. Ein Transkriptband. Radolfzell: Verlag für Gesprächsforschung.

Brugman, Hennie / Russel, Albert (2004): Annotating Multi-Media / Multi-Modal resources with ELAN. In: Lino et al. (2004), 2065–2068.

Deppermann, Arnulf (1999): Gespräche analysieren: eine Einführung. Opladen: Leske & Budrich.

Deppermann, Arnulf (2000): Ethnographische Gesprächsanalyse: Zu Nutzen und Notwendigkeit von Ethnographie für die Konversationsanalyse. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (1), 96-124.

DFG (1998): Deutsche Forschungsgemeinschaft: Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft" – Vorschläge zur Sicherung guter wissenschaftlicher Praxis, Januar 1998. Bonn: DFG.

Draxler, Christoph / Schiel, Florian (2002): Three New Corpora at the Bavarian Archive for Speech Signals - and a First Step Towards Distributed Web-Based Recording. In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC) 2002, Las Palmas, Gran Canaria, Spain. Paris: ELRA.

Evert, Stefan / Carletta, Jean / O'Donnell, Timothy J. / Kilgour, Jonathan / Vögele, Andreas / Voormann, Holger (2003): The NITE Object Model. Version 2.1. (24 March 2003). NITE Internal document. [http://www.ltg.ed.ac.uk/NITE/documents.html]

Fiehler, Reinhard (2005): Datenbank Gesprochenes Deutsch (DGD). Angekündigt in: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (6).

Gilles, Peter (2001): prosoDB: Eine multimediale Datenbankumgebung für konversationelle und prosodische Analysen. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (2), 75-89.

Glas, Reinhold / Ehlich, Konrad (2000): Deutsche Transkripte: Ein Repertorium. Arbeiten zur Mehrsprachigkeit, Serie A (63). Hamburg.

Halblützel, Susanna (2002): Kommunikationstraining in der Bank. Diskursanalytische Untersuchung eines Trainings im Bereich der Finanzanlageberatung. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (3), 165-191.

Isard, Amy (2001): An XML Architecture for the HCRC Map Task Corpus. In: Kühnlein, P. / Rieser, H. / Zeevat, H. (Hg.): Proceedings of BI-DIALOG 2001. Bielefeld.

Labov, William, Malcah Yaeger & Richard Steiner (1972): A Quantitative Study of Sound Change in Progress. Philadelphia: U. S. Regional Survey.

Lino, M. / Xavier, M. / Ferreira, F. / Costa, R. / Silva, R. (Hg.) (2004): Proceedings of the Fourth International Conference on Language Resources and Evaluation. Paris: ELRA.

MacWhinney, Brian (2001): From CHILDES to TalkBank. In: Almgren, M. / Barreña, A. / Ezeizaberrena, M. / Idiazabal, I. / MacWhinney, B. (Hg.): Research on Child Language Acquisition. Somerville, MA: Cascadilla, 17-34

Meyer, Bernd (2003): Dolmetschertraining aus diskursanalytischer Sicht: Überlegungen zu einer Fortbildung für zweisprachige Pflegekräfte. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (4), 160-185.

MG (2003): Diskussion auf der Mailingliste Gesprächsforschung unter dem Betreff "Video- und/oder Transkriptsammlung zu Kindersprache", 07. Februar – 22. Februar 2003. http://www.gespraechsforschung.de/liste.htm

Milde, Jan-Torsten / Gut, Ulrike (2001): The TASX-Environment: an XML-based corpus database for time aligned language data. In: Bird, Steven / Buneman, Peter / Liberman, Mark (Hg.): Proceedings of the IRCS Workshop On Linguistic Databases, 11-13 December 2001. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania, 174-180.

O'Connell, Daniel / Kowal, Sabine (1995): Transcription Systems for Spoken Discourse. In: Verschueren, Jef / Östman, Jan-Ola / Blommaert, Jan (Hg.): Handbook of pragmatics. Amsterdam: Benjamins, 646-656.

Redder, Angelika (2002): Professionelles Transkribieren. In: Jäger, Ludwig / Stanitzek, Georg (Hg.): Transkription – Medien/Lektüre. München: Fink, 115-131.

Rehbein, Jochen / Kameyama, Shinichi (2005): Pragmatik/Pragmatics. In: Ammon, Ulrich / Dittmar, Norbert / Methheier, Klaus / Trudgill, Peter (Hg.): Sociolinguistics. An International Handbook of the Science of Language and Society. Berlin, New York: Walter de Gruyter, 556-588.

Sager, Sven F. (2001): Formen und Probleme der technischen Dokumentation von Gesprächen. In: Brinker, Klaus / Antos, Gerd / Heinemann, Wolfgang / Sager, Sven (Hg.): Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. Berlin, New York: Walter de Gruyter, 1022-1033.

Schmidt, Thomas (2002): Gesprächstranskription auf dem Computer: das System EXMARaLDA. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (3), 1-23.

Schmidt, Thomas (2005): Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a.M.: Peter Lang (zgl. Dissertation, Universität Dortmund).

Schmitt, Reinhold (2003): Inszenieren: Struktur und Funktion eines gesprächsrhetorischen Verfahrens. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (4), 186-250.

Selting, Margret / Auer, Peter / Barden, Birgit / Bergmann, Jörg / Couper-Kuhlen, Elizabeth / Günthner, Susanne / Meier, Christoph / Quasthoff, Uta / Schlobinski, Peter / Uhmann, Susanne (1998): Gesprächsanalytisches Transkriptionssystem (GAT). In: Linguistische Berichte 173, 91-122.

Simpson, A.P. / Kohler, K.J. / Rettstadt, T. (1997): The Kiel Corpus of Read/ Spontaneous Speech - Acoustic data base, processing tools and analysis results. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 32.

Thomas Schmidt
Projekt Z2 'Computergestützte Erfassungs- und
Analysemethoden multilingualer Daten'
SFB 538 'Mehrsprachigkeit“
Max Brauer-Allee 60
22765 Hamburg
thomas.schmidt@uni-hamburg.de

Veröffentlicht am 9.6.2005