Marc Kupietz (Mannheim)/Thomas Schmidt (Mannheim)

Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung

Abstract: Die Programmbereiche „Korpuslinguistik“ und „Mündliche Korpora“ haben am IDS die Aufgabe, Grundlagen für die empirische Erforschung der deutschen Sprache zu legen. Unter anderem sammeln und erstellen sie schriftliche und mündliche Korpora, bereiten sie für eine wissenschaftliche Nutzung auf und stellen sie über Web-Oberflächen (COSMAS, DGD2 demnächst KorAP) zur Verfügung. Unser Beitrag gibt zunächst einen Überblick über Entstehungsgeschichte und aktuellen Stand dieser Arbeiten. Mit einem Blick in die Zukunft widmen wir uns auch der Frage, ob und in welcher Weise das Schlagwort ‚Big Data‘ für diese Arten linguistischer Ressourcen relevant ist.

In Bezug auf die schriftlichen Korpora wird dabei insbesondere über die diesjährige DEREKO-Erweiterung um über 17 Milliarden Wörter und die damit verbundenen Arbeiten berichtet. In diesem Zusammenhang werden u.a. DEREKOs Design, die zugrundeliegende Akquisitionsstrategie und Überlegungen zu Dispersion und Stratifizierbarkeit diskutiert.

Die spezifischen Herausforderungen, die sich beim Aufbau eines großen Gesprächskorpus stellen, werden am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) diskutiert. Dabei steht außer Frage, dass angesichts des Aufwandes, den Feldzugang sowie Erschließung der im Feld gewonnenen Audio- und Videodaten bedeuten, vergleichbare Datenmengen und Wachstumsraten wie bei Textkorpora nicht zu erreichen sind. Für den Aufbau umfangreicher mündlicher Korpora ist daher die Entwicklung eines eigenen Methodeninstrumentariums notwendig.

1 Einleitung

Die Programmbereiche „Korpuslinguistik“ und „Mündliche Korpora“ haben am IDS die Aufgabe, Grundlagen für die empirische Erforschung der deutschen Sprache zu legen. Unter anderem sammeln und erstellen sie schriftliche und mündliche Korpora, bereiten sie für eine wissenschaftliche Nutzung auf und stellen sie über Web-Oberflächen (COSMAS, demnächst KorAP, DGD2) zur Verfügung. Im Folgenden geben wir einen Überblick über Entstehungsgeschichte und aktuellen

Stand dieser Arbeiten. Mit einem Blick in die Zukunft widmen wir uns auch der Frage, ob und in welcher Weise das Schlagwort ‚Big Data‘ für diese Arten linguistischer Ressourcen relevant ist.

2 Geschichte schriftlicher Korpora am IDS

Die lange Geschichte sehr großer Schriftkorpora am IDS begann bereits weniger als drei Jahre nach seiner Gründung, als 1967 die Direktoren Paul Grebe und Ulrich Engel die Abteilung „Dokumentation der Deutschen Sprache“ gründeten, deren Aufgabe es insbesondere war, in großem Umfang deutschsprachige Texte zu sammeln und diese elektronisch – auf Lochkarten – zu speichern. Mit dem Mannheimer Korpus I (MK I), das 2,2 Millionen Wörter in 293 Texten, hauptsächlich aus den Bereichen Belletristik und Zeitungstexte umfasste, wurde bereits 1969 das erste Korpus veröffentlicht. 1972 wurde ein kleinerer zweiter Teil, MK II genannt, mit weiteren Texttypen hinzugefügt. Im folgenden Projekt „Grundstrukturen der deutschen Sprache“ wurde das Mannheimer Korpus dann intensiv analysiert und diente zwischen 1971 und 1981 als empirische Grundlage für 17 publizierte Bände zu grammatischen Themen (Teubert/Belica 2014, S. 301 ff.). Über die Jahre kamen immer mehr Korpora hinzu, u.a. von Zweigstellen des IDS, an denen spezielle Projekte wie etwa das Bonner Zeitungskorpus (Hellmann (Hg.) 1984) beheimatet waren. Die Bezeichnung Deutsches Referenzkorpus DEREKO wurde 1999 im Zusammenhang mit dem gleichnamigen, vom MWK Baden-Württemberg finanzierten Verbundprojekt mit den Universitäten Stuttgart und Tübingen als Kooperationspartner des IDS eingeführt. Ziel des 2002 abgeschlossenen Projekts war es, die deutsche Gegenwartssprache möglichst breit und der Sprachwirklichkeit angemessen zu repräsentieren, mit modernen korpuslinguistischen Verfahren aufzubereiten und der Wissenschaft zu Verfügung zu stellen. Nachdem dann 2004 am IDS die Arbeiten zum „Korporausbau“ als eigenständiges Dauerprojekt etabliert wurden, fiel die Entscheidung, die Bezeichnung DEREKO künftig für die Sammlung aller am IDS zur Verfügung gestellten Schriftkorpora der Gegenwartssprache zu verwenden. Mit dem neuen Namen und dem neu erschaffenen Projekt wurden dann viele der bereits vorher vorhandenen und heute noch gültigen Prinzipien und Ziele des Korporausbaus am IDS stärker formalisiert und akzentuiert.

3 Das Deutsche Referenzkorpus DEREKO

3.1 Ziele und Eigenschaften

Das wichtigste Ziel und gleichzeitig der höchste Anspruch von DEREKO ist, als so genanntes *very large general-purpose corpus* als Grundlage für ein möglichst breites Spektrum germanistisch-linguistischer Forschungsbereiche innerhalb und außerhalb des IDS abzudecken und entsprechend möglichst viele Aspekte der Schriftsprache kontinuierlich über die Zeit abzubilden. Um diese Ziele mit vertretbaren Kosten erreichen zu können, ist DEREKO weniger als direkt verwendbares Korpus konzipiert, sondern als eine Ur-Stichprobe, aus der sich seine Nutzer der jeweiligen Fragestellung angemessene Unterstichproben, so genannte *virtuelle Korpora* (al-Wadi 1994; Kupietz/Keibel 2009a) selbst zusammenstellen können. Trotz der Einbeziehung von Kostenaspekten in die Ausbaustrategie handelt es sich bei DEREKO jedoch nicht um ein opportunistisches Korpus, für das wahllos kostenfreie Texte gesammelt werden. Um seinem Anspruch als möglichst universell verwendbare empirische Grundlage gerecht zu werden, ist neben der Maximierung seiner Größe die möglichst breite Abdeckung potenziell relevanter Texttypen und anderer Strata die wichtigste Maxime des DEREKO-Ausbaus. Entsprechend enthält DEREKO neben Zeitungstexten – in absoluten Zahlen betrachtet und verglichen mit so genannten ausgewogenen Korpora – auch große Mengen an belletristischen, wissenschaftlichen und populärwissenschaftlichen Texten, Gebrauchstexten, verschriftlichter gesprochener Sprache und internet-basierter Kommunikation.

3.2 Ur-Stichproben-Design und Ausbaustrategie

Entsprechend seines Ur-Stichproben-Designs strebt DEREKO anders als viele andere bekannte Korpora keinerlei Ausgewogenheit an, da von der jeweiligen Fragestellung und der zu untersuchenden Sprachdomäne abhängig ist, welcher Begriff von Ausgewogenheit angemessen ist (vgl. Perkuhn/Keibel/Kupietz 2012, S. 47 f.). Fragen wie „Was ist die Sprachdomäne?“, „Was ist ausgewogen, 20% – 30% – 50%, oder ...?“, „Welche Strata sind überhaupt relevant?“ „Ist eine Proportionierung überhaupt relevant?“ sollte jeder Forscher selbst beantworten dürfen und er sollte in die Lage versetzt werden, selbst Sub-Stichproben stratifiziert zu ziehen, die möglichst repräsentativ bezüglich seiner jeweiligen Fragestellung sind.

Darüber hinaus sind weitere Vorteile des Ur-Stichproben-Designs, dass die Daten optimal für die maximale Anzahl potenziell relevanter Fragestellungen

nachnutzbar sind und dass bezüglich der Besetzung einzelner Strata ausschließlich absolute Größen relevant sind (vgl. auch Kupietz et al. 2010). So kann das Kosten-/Nutzenverhältnis bei der Akquisition dahin gehend optimiert werden, dass etwa kostenlos akquirierbare Daten nicht abgelehnt werden müssen, nur weil dies einer zuvor festgelegten Proportionierung zuwider läuft (Kupietz in Vorb.).

Angesichts der immensen wissenschaftlich-methodischen und wirtschaftlichen Vorteile liegt die Frage nahe, warum nicht bei Korpora wie dem British National Corpus (BNC) (Aston/Burnard 1998), dem Mannheimer Korpus I oder zumindest bei modernen Korpora wie dem Contemporary Corpus of American English (COCA) (Davies 2010) ähnlich vorgegangen wurde bzw. wird. Dafür gibt es jedoch eine Vielzahl potenzieller Gründe: (1.) Zumindest bis in die 1990er Jahre herrschte möglicherweise eine optimistischere Einschätzung bezüglich der Repräsentativität von Korpora für Sprache schlechthin vor; (2.) die wirtschaftlichen Vorteile des Ur-Stichproben-Designs wirken sich erst bei dauerhaft angelegten Projekten aus; (3.) die Bildung von virtuellen Teilkorpora ist nur bei sehr großen Ur-Stichproben sinnvoll (absolute Größen sind relevant, siehe oben); (4.) der technische Overhead für die Implementierung persistenter, wiederverwendbarer virtueller Korpora in Korpusplattformen ist immens; (5.) die stratifizierte Ziehung von Sub-Stichproben ist entscheidend von zuverlässigen Metadaten abhängig, was viele Web-Korpora ausschließt; (6.) die Kosten für Akquisition und Aufbereitung waren bis in die späten 1990er-Jahre über verschiedene Quellen und über die Zeit weitgehend konstant, während heute das Verhältnis etwa zwischen Zeitungen und Romanen 1:25.000 erreichen kann (vgl. Kupietz 2014, S. 323).

Insbesondere die Kombination aus den Gründen (3.) und (6.) ist entscheidend. Während etwa für das BNC eine im Voraus geplante, stratifizierte Stichprobenbildung möglich und sinnvoll war, ist das IDS bezüglich DEREKO zum einen in der Lage, zum anderen aber auch rechtlich und wirtschaftlich darauf angewiesen, kontinuierlich innerhalb der über die Zeit variierenden, überhaupt vorhandenen Angebote seine Akquisitionsauswahl zu treffen.

Vor diesem Hintergrund kann man die für den DEREKO-Ausbau relevanten Faktoren und Parameter folgendermaßen skizzieren:

1. Maximierung von Größe und Dispersion,
2. Nachfrage von IDS-internen/-externen Nutzern,
3. Angebot von Rechteinhabern und IDS-internen und -externen Projekten,
4. Kosten,
 - a) Akquisition (Verhandlungsaufwand, Lizenzgebühren),
 - b) Erschließung (Analyse, Aufbereitung, Konvertierungsprogramme),
 - c) Folgekosten (zu erwartende Formatänderungen, Lizenzgebühren),
5. Prognosen bezüglich der Entwicklung der Faktoren,

6. langfristige Akquisitionsstrategie z.B. Verbesserung der eigenen Rahmenbedingungen (Lüngen et al. 2011),
7. kontinuierliche Anpassung der Parameter an aktuelle Verhältnisse.

3.3 Rechtlicher Rahmen

Eine Besonderheit der Linguistik und angrenzender Disziplinen ist, dass ihre Forschungsprimärdaten typischerweise von Dritten betroffen sind (Kupietz/Bankhardt (Hg.) 2009). Im Falle von Textkorpora sind dies insbesondere die Urheberrechte von Autoren und die Verwertungsrechte, die Verlage von den Autoren erworben haben. Entsprechend ist das IDS auch nicht der Eigentümer der in DEREKO enthaltenen Texte, sondern hat nur bestimmte, in ca. 200 Lizenzverträgen festgelegte Nutzungsrechte. Um etwaige Lizenzgebühren möglichst niedrig zu halten, sind die Nutzungsrechte außerdem eng begrenzt. Z.B. ist der Download vollständiger Texte typischerweise ausgeschlossen, da nur so eine unkontrollierbare Verbreitung und eine gewöhnliche, so genannte konsumtive Nutzung der Texte ausgeschlossen werden kann. Ohne diesen Ausschluss würden für DEREKO jährlich schätzungsweise mehrere 100 Millionen Euro Lizenzkosten anfallen. Da das IDS alle vorhandenen Rechte weitergibt und für sich selbst keine besonderen Rechte an Aufbereitung und Bereitstellung reklamiert, ist das, was das Recherchesystem des IDS, COSMAS II (Bodmer 2005), an Recherchemöglichkeiten zulässt, im Wesentlichen genau das, was die Lizenzvereinbarungen mit den Rechteinhabern zulassen (Kupietz et al. 2010, 2014).

4 Aktuelle DEREKO-Entwicklungen

4.1 Neu-Akquisitionen in 2013/2014

Die Jahre 2013 und 2014 waren hinsichtlich neuer Akquisitionen für DEREKO in besonderem Maße erfolgreich. Zum einen konnte Belletristik in einem Umfang von über 6 Millionen Wörtern akquiriert und damit eine Verdoppelung in diesem vergleichsweise teuren und schwach besetzten Stratum seit 2010 erzielt werden. Zum anderen konnten durch zwei große Lizenzvertragsabschlüsse in 2013 knapp 100 neue regionale und überregionale Zeitungen und Zeitschriften, wie etwa die Süddeutsche Zeitung (ab 1992), ZEIT, FOCUS, Neue Zürcher Zeitung, Falter, profil, Weltwoche, Luxemburger Tag (alle ab 2000), im Umfang von insgesamt etwa 16 Milliarden Wörtern in DEREKO integriert werden.

Weitere substanzielle Erweiterungen gab es in den Bereichen internetbasierte Kommunikation bzw. konzeptionelle Mündlichkeit und konzeptionell schriftlicher gesprochener Sprache. Ersterer hat vor allem durch eine Kooperation mit dem CLARIN-Teilprojekt am IDS sowie Michael Beißwenger (TU Dortmund) entstandenen, qualitativ stark verbesserten und aktualisierten Auflage der Wikipedia-Diskussionen (wdd13) profitiert. In letzterem wurden in Kooperation mit dem PolMine-Projekt (Uni Duisburg) das Plenarprotokoll-Korpus akquiriert und um den Landtag von Niederösterreich erweitert.

Das DEREKO-2014-I-Release (IDS 2014) konnte so gegenüber dem vorherigen um 17 auf etwa 24 Milliarden Wörter erweitert werden.

4.2 More Data is Better Data

Angesichts der nun erreichten Größe von DEREKO, die ansonsten nur noch von wenigen reinen Web-Corpora übertroffen wird (Jakubíček et al. 2013), mag man sich die Frage stellen, ob Mercers Zitat in der Überschrift (Church/Mercer 1993) noch zutrifft oder ob DEREKOs Umfang nicht nunmehr ausreichend ist. Unserer Überzeugung nach gilt, sofern alle anderen Parameter gleich bleiben, nach wie vor: je größer das Korpus, desto verlässlichere Aussagen über mehr verschiedene und seltenere Phänomene. Hinzu kommt, dass es in der Sprache sehr viele seltene Phänomene gibt (*Large Number of Rare Events*). Ein Korpus besteht z.B. unabhängig von seiner Größe typischerweise etwa zur Hälfte aus Hapax Legomena, also aus Wortformen, die nur einmal auftauchen (Perkuhn/Keibel/Kupietz 2012, S. 86), und die Hypothese, dass linguistische Fragestellungen eher seltene Phänomene betreffen, ist sicherlich nicht ganz von der Hand zu weisen (Belica et al. 2011). Seltenes muss zudem nicht auf niedrig-frequente Wörter beschränkt sein, sondern kann auch durch eine Kombination von Bedingungen entstehen (Perkuhn/Keibel/Kupietz 2012, S. 50). Im Falle von DEREKO reduziert sich der Umfang eines virtuellen Korpus gegenüber dem der Ur-Stichprobe auch schon bei seiner Definition durch die Kombination von Eigenschaftsrestriktionen (z.B. Zeitraum 1990–1999, nur aus Österreich, gleichverteilt über Themen). DEREKO muss zudem wachsen, um einerseits reproduzierbare Ergebnisse zu garantieren und andererseits den aktuellen Sprachgebrauch auch morgen noch abzubilden.

4.3 Diversität

Andererseits ist das Größenwachstum eines Korpus nur dann sinnvoll, wenn dabei auch die Streuung zumindest gleich bleibt oder idealerweise wächst. Vorausset-

zung für eine gezielte Verbesserung der Streuung ist, zunächst Dimensionen oder diskrete Strata zu identifizieren, die zum einen potenziell linguistisch relevant und zum anderen auch im Korpus messbar sind, um zunächst überhaupt die Verteilung von Wörtern oder Texten eines Korpus bezüglich dieser quantifizieren zu können (vgl. Kupietz in Vorb.). Recht unproblematisch ist dies im Falle von DEREKO bei Entstehungszeit und Erscheinungsort. Letzterer spielt bei DEREKO insbesondere bei der Akquisition regionaler Zeitungen und Zeitschriften eine wichtige Rolle. Abbildung 1 zeigt z.B., dass die oben angesprochenen Neu-Akquisitionen DEREKOs geographische Abdeckung durch Zeitungen und Zeitschriften insbesondere bezüglich West- und Nordostdeutschland sowie Luxemburg stark verbessert hat, während Norddeutschland nach wie vor kaum repräsentiert ist. Diese geographische Distributionsanalyse ist außerdem auch ein Kriterium bei der Priorisierung der Integration von DEREKO-2014-I in COSMAS II, das aufgrund seiner ausgeschöpften technischen Kapazitäten nicht mehr DEREKO als Ganzes aufnehmen kann.

Abb. 1: Geografische Verteilung alter (grau) und neu-akquirierter (schwarz) Zeitungsquellen von DEREKO

Recht unkompliziert ist zudem auch die (hier aus Platzgründen nicht dargestellte) Distributionsanalyse bezüglich thematischer Domänen – mit der Einschränkung, dass Zuordnungen zu solchen Domänen nur mit Hilfe maschineller Lernverfahren und unvollkommen approximiert werden können (siehe Klosa/Kupietz/Längen 2012; Kupietz in Vorb.).

Sehr viel unklarer ist dagegen, wie Verteilungen bezüglich ebenfalls relevanter Eigenschaften wie Textsorte, Genre, Texttyp und Register zu analysieren sind. Zum einen sind Taxonomien bezüglich dieser Eigenschaften unterschiedlich definierbar und zum anderen ist auch die Operationalisierung gegebener Taxonomien, also eine automatische, zuverlässige Zuordnung von Texten zu diesen, typischerweise nur bedingt möglich (Kupietz in Vorb.). Aus diesen Gründen ist DEREKOs Ansatz zur Erfassung solch schwer definierbarer und operationalisierbarer Eigenschaften, diese zunächst grob und datengeleitet anzugehen und daraus abgeleitete Hypothesen post hoc im Detail zu überprüfen. Eine häufig verwendete Grundlage solcher für DEREKO verwendeter Analysen ist ein von Adam Kilgariff (2001) entwickeltes, ausschließlich in Tokenfrequenzlisten verankertes Maß für die Ähnlichkeit bzw. Unähnlichkeit von Korpora.

Um einen wie oben beschriebenen Überblick über die in 2013 neu akquirierten Zeitungs- und Zeitschriftenquellen zu erhalten, wurde z.B. folgende Vorgehensweise gewählt: Da das χ^2 -basierte Distanzmaß nur für etwa gleichgroße Korpora zuverlässig anwendbar ist, wurden zunächst aus allen bestehenden und neuen Zeitungsquellen sowie aus anderen DEREKO-Quellen Zufallsstichproben der kleinsten Quelle entsprechenden Größe von etwa einer Million Wörtern gezogen. Für die so entstandenen 296 Stichproben (bzw. Subkorpora) wurden entsprechend Kilgariffs Vorgehensweise paarweise alle Distanzen berechnet. Um die so entstandene Distanzmatrix mit über 40.000 Werten besser interpretierbar zu machen, wurde diese dann mit Hilfe nicht-metrischer multidimensionaler Skalierung auf eine Karte abgebildet, in die außerdem auch die Größen der einzelnen Quellen – repräsentiert durch Label-Größen – und die Information DEREKO-Bestand (grau) vs. Neu-Akquisition (schwarz) kodiert wurden. Das in Abbildung 2 dargestellte Ergebnis zeigt zunächst, dass das Verfahren insgesamt ein gewisses Maß an Robustheit und Validität aufweist und sich Korpusähnlichkeit auf der Karte offenbar in Nachbarschaft manifestiert – vor allem da die drei Kontrollstichproben aus DEREKO-Gesamt (D0, D1, D2) fast auf identische Positionen projiziert sind. Außerdem sind plausible Nachbarschaftscluster, wie etwa für Literatur (thm, goe, loz, lit, gri, bio) und gesprochene Sprache (rei, pfe) auszumachen. Bezüglich der konkreten Fragestellung, welche neu akquirierten Quellen am stärksten zur Diversitätssteigerung von DEREKO beitragen, lässt die Karte daher für diese Zwecke vermutlich ausreichend sichere Interpretationen zu. So ist zu erkennen, dass insbesondere die Neu-Akquisitionen zca, zge,

zwi, fom, boz, flt, wwo Eigenschaften haben, die in DEREKO bisher wenig vertreten waren und diese daher gute Kandidaten für eine priorisierte Aufnahme in COSMAS II darstellen.¹

Abb. 2: NMDS-Projektionen der χ^2 -Distanzmatrix von alten (grau) und neuen (schwarz) DEREKO-Quellen basierend auf einer Tokenfrequenzliste

4.4 Big Data?

Die Erweiterung von DEREKO von 7 auf 24 Milliarden Wörter in 2013/2014 legt den Eindruck nahe, dass DEREKO und das IDS nun im ‚Big-Data‘-Zeitalter angekommen sind. Betrachtet man jedoch in Abbildung 3 nur die Entwicklung des Umfangs der primären Textdaten in Relation zu den jeweils gebräuchlichen (portablen) Speichermedien (schwarze Kurve), erkennt man, dass Paul Grebe und seine Kollegen bereits in den 1960er-Jahren viel Erstaunlicheres geleistet haben und dass die Sprachwissenschaft am IDS eine weit zurückreichende ‚Big-Data‘-Tradition hat.

¹ Um bestimmte Fehlerquellen auszuschließen, wurde das Zustandekommen der Distanzen der entsprechenden Stichproben zu den Kontrollstichproben im Einzelnen überprüft. Außerdem wurden weitere Bedingungen getestet (siehe Kupietz/Lüngen 2014, S. 2382).

Abb. 3: Entwicklung des Speicherplatzbedarfs von DEREKO seit 1969 in Relation zur Größe jeweils aktueller portabler Speichermedien – die Werte bis 1993 sind geschätzt, für 1969 wurde eine Box mit 2.000 Lochkarten als Referenz verwendet

Erst wenn man zusätzlich zum Umfang der Primärdaten auch den Umfang von Annotationsdaten berücksichtigt (graue Kurve in Abbildung 3), wird unser Eindruck bestätigt, dass DEREKO auch heute wieder mit potenziell schwer beherrschbaren Datenmengen konfrontiert ist.

Als wesentliche Konsequenz der Größenentwicklung der Annotationsschichten war spätestens 2009 absehbar, dass das bereits Anfang der 1990er Jahre konzipierte COSMAS II nicht ausreichend skalieren würde. Entsprechend wurde 2010 mit den Vorbereitungen des Nachfolgesystems KorAP (Bański et al. 2012; Kupietz/Frick 2013) begonnen, dessen wichtigstes technisches Ziel es ist, eine theoretisch unbegrenzte Menge an Primär- und Annotationsdaten durch horizontale Skalierbarkeit, d.h. Verteilbarkeit auf beliebig viele Rechner, zu erlauben (vgl. Kupietz et al. 2014). Ein inhaltliches Ziel von KorAP bezüglich ‚Big Data‘ ist es außerdem, der Etablierung von korpusgeleiteten explorativen Analysen in der Mainstream-Linguistik, etwa durch das Angebot von Schnittstellen für extern entwickelte Module und Programme, Rechnung zu tragen. Letztlich soll so, frei nach Jim Grays Maxime „put the computation near the data“ (2003), das sprachwissenschaftliche Potenzial von DEREKO optimal ausschöpfbar gemacht werden – trotz der Tatsache, dass sein Angebot zum Download aus rechtlich-wirtschaftlichen Gründen nicht möglich und aufgrund seiner Größe für die meisten Anwender auch nicht sinnvoll ist. Ein wichtiger Aspekt u.a. beim Angebot solcher Schnittstellen wird

es auch sein, eine sich abzeichnende engere Verzahnung und Zyklisierung von quantitativen Analysen und qualitativen Interpretationen, etwa durch die Möglichkeit interaktiver Visualisierung (wie z.B. in Abschnitt 4.3) zu unterstützen.

Auch was diesen Kernaspekt des Begriffs ‚Big Data‘, also die Gewinnung von Erkenntnissen auf Grundlage großer Datenmengen betrifft, kann dabei an eine weit zurückreichende Tradition methodologischer Arbeiten (z.B. Belica 1996) und wissenschaftstheoretischer Arbeiten (z.B. Kupietz/Keibel 2009b, Keibel/Kupietz 2009) am IDS zur Integration quantitativer Befunde in den Prozess des sprachwissenschaftlichen Erkenntnisgewinns angeknüpft werden.

5 Geschichte mündlicher Korpora am IDS

Die Geschichte der mündlichen Korpora am IDS beginnt weit vor der Geburtsstunde des Instituts mit der Gründung des Deutschen Spracharchivs (DSAv) im Jahre 1932 durch Eberhard Zwirner. Unter seiner Leitung wurde in den 1950er Jahren das Korpus „Deutsche Mundarten“ erhoben, das die Dialekte im Gebiet der damaligen Bundesrepublik systematisch dokumentiert. Das Design dieses Korpus war in der Folge Vorlage für weitere Variationskorpora, darunter das Korpus „Deutsche Umgangssprache“ (Pfeffer-Korpus), das Korpus „Deutsche Mundarten: Ehemalige deutsche Ostgebiete“ sowie das Korpus „Deutsche Mundarten: DDR“. Diese und weitere Variationskorpora bilden bis heute einen wesentlichen Bestandteil der Archivbestände am IDS.

Das DSAv wurde 1971 ein Teil des IDS. In diese Zeit fällt mit den Korpora „Grundstrukturen“ (auch: Freiburger Korpus) und „Dialogstrukturen“ auch die Erhebung der ersten Gesprächskorpora des Deutschen, denen in den 1980er und 1990er Jahren weitere – etwa „Stadtsprache Mannheim“ oder „Elizitierte Konfliktgespräche“ – folgten, die heute zu den Beständen des Archivs gehören.

In den 1990er Jahren hat das DSAv den Übergang von einem analogen zu einem digitalen Archiv begonnen, der für die vorhandenen Archivbestände mittlerweile zu großen Teilen abgeschlossen ist. Ein Meilenstein im Übergang zum Digitalen war der erste Online-Zugang zum Archiv, der in Form der Datenbank Gesprochenes Deutsch des DSAv (Fiehler/Wagener 2005) erstmals 2002 zur Verfügung gestellt wurde. Im Jahre 2004 wurde das DSAv in Archiv für Gesprochenes Deutsch (AGD) umbenannt, um die Fokussierung auf das Mündliche auch im Namen deutlich zu machen.

Für eine detailliertere Darstellung der Archiv-Geschichte sei der Leser auf den Überblick in Stift/Schmidt (2014) verwiesen. Die genannten Eckdaten illustrieren aber bereits zwei große Entwicklungen, vor deren Hintergrund auch die aktuellen

Arbeiten im Programmbereich zu sehen sind: zum einen erfolgte eine Entwicklung in der inhaltlichen Ausrichtung des Archivs, in der zum ursprünglich zentralen Datentyp des Variationskorpus später der Datentyp Gesprächskorpus hinzukommt, der gegenwärtig den Schwerpunkt der Arbeiten im Programmbereich bildet. Zum anderen folgt die Entwicklung auch technischen Innovationen, insbesondere was den Beginn des digitalen Zeitalters und die zunehmenden Möglichkeiten der Online-Bereitstellung von Daten betrifft.

6 Aktuelle Arbeiten im Programmbereich „Mündliche Korpora“

Standards und etablierte gute Praktiken für die Erhebung, Erschließung und Bereitstellung mündlicher Korpora haben erst in den letzten zehn Jahren begonnen, Verbreitung in den betreffenden Fachwissenschaften zu finden. Dementsprechend defizitär hinsichtlich rechtlicher Autorisierung, dokumentarischer Erschließung und technischer Aufbereitung sind viele Datenbestände, die das AGD zur Archivierung übernimmt, dementsprechend knapp ist heute allgemein das Angebot an gut (nach)nutzbaren mündlichen Korpora. Der Programmbereich sieht es als eine seiner Aufgaben an, zur Verbesserung dieser Situation beizutragen. Neben der Mitarbeit an Standardisierungsprozessen (z.B. Schmidt 2011) und an der Entwicklung und Etablierung guter Praktiken geschieht dies vor allem durch drei Maßnahmen (siehe 6.1–6.3):

6.1 Datenübernahmen und Datenkuration

Das AGD übernimmt Korpora abgeschlossener Projekte und bereitet diese für eine Nachnutzung durch die wissenschaftliche Öffentlichkeit auf. Je nach Beschaffenheit der Ausgangsdaten umfasst dies die Digitalisierung analoger Ton- oder Videoaufnahmen, die Überführung von Transkript-Dokumenten in strukturierte Formate, ein Text/Ton-Alignment zwischen Aufnahmen und Transkripten und/oder die Erstellung strukturierter Metadaten-Dokumentationen zu Sprechereignissen und Sprechern. Gegenwärtig werden u.a. auf diese Weise mehrere Korpora zum Australien-Deutschen (Clyne 1981), das „Berliner Wendekorpus“ (Dittmar/Bredel 1999), das Korpus „Deutsche Mundarten: DDR“ (siehe oben) sowie das Korpus „Ausbildungssituationen im Bergbau“ (Brünner 2005) mit dem Ziel einer Bereitstellung im Service des AGD und ggf. in der DGD2 bearbeitet.

6.2 Entwicklung der Datenbank für Gesprochenes Deutsch

Mit der Datenbank für Gesprochenes Deutsch (DGD2; Schmidt/Dickgießer/Gasch 2013) entwickelt das AGD eine Plattform, die mündliche Korpora für die computergestützte Analyse erschließt. Dies beinhaltet zum einen Funktionen zum explorativen Browsen (Durchblättern) des Archivbestands. Zum anderen stellt die DGD2 verschiedene Suchfunktionalitäten zur Verfügung, um den Nutzer zu einer gezielten Auswahl aus den Beständen zu befähigen. Gesucht werden kann sowohl auf Ebene der Metadaten, um eine Vorauswahl von Daten gemäß extralinguistischer Kriterien (etwa: „Aufnahmen von Familiengesprächen im nordniederdeutschen Sprachraum“) zu ermöglichen, als auch auf der Ebene von Transkripten und Annotationen, um Belegsammlungen für konkrete linguistische Phänomene (etwa: „alle Vorkommen von Modalverben im Präteritum“) zusammenzustellen und analysieren zu können.

The screenshot displays the DGD2 search interface. At the top, there are tabs for 'SUCHE', 'METADATEN', and 'ANZEIGE'. The search form includes fields for 'Wort:' (containing 'wohl'), 'Normalisiert:', and 'Lemma: wohl'. A 'Suche starten' button and a checkbox for 'Reguläre Ausdrücke' are also present. Below the search bar, a status bar indicates 'Transkriptausschnitt berechnet' and a timer '00:00:01.0'. The main results area shows a table with columns: 'Ergebnisse', 'Sprecher', 'Treffer', 'Ort (Region)', and 'Geschlecht'. The results list various entries with checkboxes, IDs (e.g., 141, 142, 143), and speaker codes (e.g., FOLK_00036, FOLK_00114). A detailed view of a specific result (1224 LS [hm]) is shown, displaying a transcript excerpt: '1224 LS [hm] nich wirklich geklärt aber es scheint wohl tatsächlich so zu sein dass der sprung von der in zimmer a geschoben sie schläft wohl inner da weil die frau patientin ak so arg also gesagt da müssen die schüler wohl auch solche textsorten wie briefe reportagen rezensionen u... es ja zugeben!!! vielen schülern wohl auch tatsächlich gehöflich hat und hilft hingegen die eben fuchzehn sekunde jungs der kloß schüttelt wohl in de kopf do hinne fünf tau und fuchzehn er wusste wohl sehr wohl bescheid hier ja dass die meisten männer da sehr wohl dazu bereit wären zu sagen ich bleib auch zuhause nur sagen dass es aber sehr wohl eben einfach ähm n steppanz ist sehr wohl ne pers jemand ne persönliche fähigkeit die man hat mädeln und die sind auch sehr wohl in der position sich selbst aussuchen zu können weichen gesagt die mädeln sind dort sehr wohl mündig und können sich aussuchen'. The interface also shows a list of other results (152, 153, 154, 155, 156, 157, 158, 159, 160) with their respective speaker codes and regions.

Abb. 4: Screenshot einer DGD2-Suchabfrage auf FOLK

Die DGD2 teilt viele wesentliche Prinzipien und Elemente der Benutzerführung, die im Bereich schriftsprachlicher Korpora schon länger in Form von Desktop-Werkzeugen und Online-Plattformen (z.B. COSMAS und KorAP am IDS) gebräuchlich sind. Dazu gehört als ein zentrales korpuslinguistisches Instrument die Darstellung von Suchergebnissen als Keyword-In-Context-Konkordanz (KWIC) mit entsprechenden Möglichkeiten der Manipulation (Sortieren, Filtern etc.). Um den Besonderheiten mündlicher Korpora Rechnung zu tragen, ergänzt die DGD2 diese Basisfunktionalität durch Funktionen etwa zum Abspielen von Aufnahmeauschnitten zu einem Suchergebnis, zum Anzeigen eines Suchergebnisses im erweiterten Transkriptkontext oder zum Korrelieren von Fundstellen im Transkript mit zugehörigen Sprechermetadaten (vgl. Abb. 4). Da in Transkripten mündlicher Daten – mehr als bei der elektronischen Repräsentation schriftsprachlicher Texte – grundsätzlich nur ein modellhaft verkürzter Ausschnitt des eigentlichen Analysegegenstands abgebildet ist (Ochs 1979; Schmidt 2005a), kommt der Möglichkeit, Suchergebnisse manuell weiterbearbeiten zu können, eine besondere Bedeutung zu. Die DGD2 trägt diesem Bedarf mit zahlreichen Methoden zum nachträglichen Inspizieren und Bearbeiten von Suchergebnissen, mit Funktionen, um bearbeitete Ergebnisse dauerhaft in der Datenbank zu speichern, und mit der Möglichkeit, Ausschnitte aus den Daten auf den lokalen Rechner zu laden und dort mit gängigen Instrumenten (insb. Editoren wie FOLKER, Schmidt 2012) weiterzuverarbeiten, Rechnung.

6.3 Aufbau des Forschungs- und Lehrkorpus Gesprochenes Deutsch

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) ist ein Gesprächskorpus des Deutschen, das mit einem Blick auf den Bedarf in den Fachwissenschaften initiiert und konzipiert wurde. Motivation für den Aufbau von FOLK ist insbesondere der Umstand, dass es für das Deutsche bisher kein öffentlich zugängliches, aktuellen methodischen und technischen Standards genügendes Korpus authentischer Gespräche gibt, obwohl ein solches zur Bearbeitung vielfältiger Fragestellungen in Gesprächsforschung, Korpuslinguistik und weiteren Disziplinen nützlich oder notwendig wäre (Deppermann/Hartung 2011; vgl. auch Schmidt 2005b).

Als Gesprächskorpus setzt FOLK den *Interaktionstyp* als zentralen Stratifikationsparameter, d.h. primäres Ziel bei der Datenauswahl ist es, möglichst viele unterschiedliche kommunikative Gattungen im Korpus zu repräsentieren. Dem nachgeordnet ist eine Stratifizierung anhand regionaler Parameter (Gespräche und Sprecher aus unterschiedlichen Regionen) und anhand verschiedener Eigen-

schaften von Sprechern (insbesondere Geschlecht, Alter, Bildungsstand). Die Ausbaustrategie folgt dem Prinzip „Breite vor Tiefe“, d.h. im Zweifelsfall wird der Erhebung eines neuen Interaktionstyps Vorrang gegeben vor der Erhebung weiterer Instanzen von bereits im Korpus repräsentierten Typen. Dabei sind für FOLK gewisse Qualitätsansprüche maßgeblich, die nicht zu Gunsten größerer Quantitäten preisgegeben werden können: Es kommen nur vollständig erhobene Gespräche in Frage, die nach den im Projekt definierten Vorgaben einheitlich dokumentiert und transkribiert und rechtlich so autorisiert sind, dass sie für die wissenschaftliche Öffentlichkeit verfügbar gemacht werden können. Auf der technischen Ebene schlagen sich diese Qualitätsansprüche in der Verwendung zeitgemäßer texttechnologischer Standards für die Erschließung nieder.

Die erste Phase des Projekts wurde im März 2014 mit der Veröffentlichung von 100 Stunden Material (= ca. 1 Million Wörter) abgeschlossen. Die Auswahl des bisher veröffentlichten Materials folgte bereits dem Prinzip der möglichst breiten Variation der Interaktionstypen, war aber noch insofern opportunistisch, als sie sich zunächst auf die Akquisition leicht zugänglicher Daten konzentrierte. Damit ist nun ein initialer Bestand aufgebaut, der die angestrebte Variationsbreite des Korpus andeuten kann. FOLK umfasst derzeit:

- verschiedene Gespräche aus dem privaten Bereich wie Tischgespräche oder aktivitätsbegleitende Gespräche wie Spielinteraktionen,
- verschiedene Gespräche aus dem institutionellen Bereich, beispielsweise Unterrichtsgespräche aus Berufsschule und Wirtschaftsgymnasium, Prüfungsgespräche aus der Universität oder verschiedene beruflich veranlasste Interaktionen,
- mit den Schlichtungsverhandlungen zu Stuttgart 21 auch eine Interaktion aus dem öffentlichen Bereich.

Die erste Phase diente auch der Etablierung eines Arbeitsablaufs im Projekt. Dieser reicht von der Erhebung über die Erschließung der Daten mittels Transkription und Annotation bis zur Bereitstellung des Korpus in der DGD2. Einige Komponenten des Arbeitsablaufs mussten im Projekt selbst entwickelt bzw. auf dessen konkrete Bedürfnisse hin angepasst und optimiert werden. Dazu gehörten insbesondere die Definition einer für die computergestützte Verarbeitung geeigneten Variante des Gesprächsanalytischen Transkriptionssystems (GAT, Selting et al. 2009; vgl. auch Schmidt 2007) sowie die Entwicklung von Richtlinien für das standard-orthografische Normalisieren von Transkripten in literarischer Umschrift. Auf der technischen Seite wurde dies komplementiert durch die Entwicklung des Transkriptionseditors FOLKER und des Annotations-Tools OrthoNormal (Schmidt 2012). Auf der Grundlage dieses Arbeitsablaufes wird nun der weitere Ausbau von FOLK geplant.

7 Repräsentativität, Ausgewogenheit und Vollständigkeit in FOLK

Als potenzielles Referenzkorpus für das gesprochene Deutsch muss sich FOLK der Frage stellen, inwieweit es in einem methodisch operationalisierbaren Sinne einen „typischen“ Ausschnitt aus der sprachlichen Wirklichkeit darstellt, die anhand von FOLK gewonnenen Erkenntnisse also auf einen über das Korpus hinausgehenden Kontext übertragbar sind.

Theoretisch könnte dabei als Maximalziel eine Repräsentativität im quantitativ-statistischen Sinne angestrebt werden. Dies würde bedeuten, dass FOLK als eine Stichprobe zu behandeln wäre, die bezüglich gewisser (näher zu bestimmender) Parameter dieselbe Verteilung aufweist wie eine (ebenfalls näher zu benennende) Grundgesamtheit. Eine unabdingbare Voraussetzung hierfür wäre, dass die Verteilungen der relevanten Parameter in der Grundgesamtheit zumindest annähernd bekannt sind. Dies ist jedoch allenfalls für einige sekundäre Stratifikationsparameter der Fall – beispielsweise ließe sich anhand von demografischen Statistiken das Verhältnis zwischen männlichen und weiblichen Sprechern oder Sprechern verschiedener Altersgruppen (z.B. in der Bundesrepublik Deutschland) ermitteln und die Verteilung in FOLK an diesen Zahlen ausrichten. Für den primären Stratifikationsparameter – den Interaktionstyp – lassen sich jedoch kaum quantitative Grundannahmen machen. Selbst wenn sich alle für das Korpus relevanten Interaktionstypen exhaustiv aufzählen und in beliebiger Menge erheben ließen, gäbe es immer noch keinerlei plausible Abschätzung darüber, welcher Interaktionstyp in welchen Mengen in der Grundgesamtheit vertreten ist. FOLK strebt daher keine Repräsentativität im quantitativ-statistischen Sinne an.

Da das Konzept der Repräsentativität auch für schriftsprachliche Korpora problematisch ist, wird dort an seine Stelle oft das Konzept der Ausgewogenheit gesetzt. In einem ausgewogenen Korpus sind einzelne Ausprägungen von Variablen oder Variablenkombinationen mit gleichen Mengen an Daten besetzt, wodurch sich ein Bezug auf eine schwer zu bestimmende Grundgesamtheit erübrigt. Beispielsweise arbeitet das DWDS-Kernkorpus mit zwei Hauptvariablen – dem Texttyp mit den Ausprägungen ‚Belletristik‘, ‚Zeitung‘, ‚Wissenschaft‘ und ‚Gebrauchsliteratur‘ sowie der zeitlichen Unterteilung des 20. Jahrhunderts in seine Dekaden – und enthält für jede mögliche Variablenkombination (z.B. Belletristik zwischen 1950 und 1959) annähernd gleiche Mengen an Daten. Um auch FOLK als ausgewogenes Korpus zusammenzustellen, würde es also im Prinzip genügen, auf der primären Stratifikationsebene eine exhaustive Liste von Interaktionstypen zu definieren, diese mit den sekundären Stratifikationsparametern

zu kreuzen und von jeder so entstehenden Variablenkombination (etwa: Tischgespräche zwischen Sprechern über 60 Jahren im moselfränkischen Sprachraum) gleiche Mengen an Daten zu erheben. In der Praxis muss auch eine solche Herangehensweise aber aus mehreren Gründen scheitern: Erstens ist es zweifelhaft, dass sich eine allgemein akzeptierte Taxonomie von Interaktionstypen finden lässt, die – analog zu den im DWDS-Kernkorpus verwendeten „großen“ Textgattungen – exhaustiv und eindeutig auf die zu erhebenden Gespräche anwendbar wäre. Dies ist analog zu einigen weiter oben dargestellten Argumenten, die auch bei schriftsprachlichen Korpora gegen das Ziel der Ausgewogenheit sprechen. Zweitens würde, selbst wenn sich eine solche Taxonomie finden ließe, die Kombination mit sekundären Parametern (z.B. 15 sprachliche Großräume x 3 Altersstufen x 3 Bildungsniveaus x 2 Geschlechter) es entweder notwendig machen, sehr große Datenmengen zu erheben,² was praktisch nicht umsetzbar ist, oder aber den primären Stratifikationsparameter auf eine übermäßig vereinfachte Liste (z.B. Interaktionsdomänen statt -typen, etwa: private/institutionelle/öffentliche Kommunikation) zu reduzieren, was dem Grundgedanken des Korpusdesigns zuwiderlaufen würde.

Wie für DEREKO werden also auch für FOLK weder Repräsentativität noch Ausgewogenheit angestrebt, wohl aber eine breitestmögliche Streuung bezüglich des primären Stratifikationsparameters sowie eine vollständige Abdeckung bezüglich der sekundären Stratifikationsparameter. Konkret bedeutet dies für die aktuelle Ausbauphase nicht nur, dass der Datenbestand um weitere Interaktionstypen erweitert wird, sondern auch, dass das in der ersten Phase aufgetretene Ungleichgewicht, das zu Gunsten von Sprechern aus dem rheinfränkischen Sprachraum zwischen 20 und 30 Jahren ausfällt, durch Erhebungen oder Akquise von Daten in anderen Sprachräumen und mit Sprechern anderer Altersstufen schrittweise ausgeglichen wird. Im Ergebnis sollte das FOLK-Gesamtkorpus dann wie DEREKO auch verwendet werden können, um virtuelle Teilkorpora zu bilden, die nach vom Forscher selbst zu wählenden Parametern durchaus wieder ausgewogen sein können. Eine entsprechende Funktionalität für das Zusammenstellen virtueller Teilkorpora ist bereits Bestandteil der DGD2.

² Das Data Repository of Spoken Czech (Benešová et al. 2014) ist bei einer Zweiteilung der Kategorien „Geschlecht“, „Alter“ und „Bildung“ sowie einer Verteilung auf neun Dialektregionen ein ausgewogenes Korpus mit etwa 3 Millionen transkribierten Wörtern, berücksichtigt den Parameter „Gesprächstyp“ aber nicht in der Stratifikation.

8 Mündliche Korpora und ‚Big Data‘

Da bis vor kurzem im Bereich der mündlichen Korpora das primäre Problem eher war, überhaupt geeignete Daten vorzufinden (Schmidt 2005b), sind Fragen der angemessenen Größe solcher Korpora bislang kaum diskutiert worden. Der Prämisse „More data is better data“ (siehe 4.2) wird angesichts der noch in den Anfängen befindlichen Entwicklungen derzeit wohl kaum widersprochen werden. Dennoch mag es lohnend sein, die Fragen nach dem „Warum“, „Wie viel“ und „Wie“ eingehender zu erörtern und mit den weiter fortgeschrittenen diesbezüglichen Überlegungen zu schriftsprachlichen Korpora in Bezug zu setzen.

8.1 Warum? Gründe für größere Datenmengen im Bereich mündlicher Korpora

Die Gründe, die für größere Datenmengen im Bereich mündlicher Korpora sprechen, sind nicht prinzipiell verschieden von denen, die oben für einen mengenmäßigen Ausbau schriftsprachlicher Korpora angeführt wurden. Wie bereits im vorangehenden Abschnitt dargelegt, läuft zum einen ein Korpusdesign, das wie bei FOLK auf maximale Dispersion ausgelegt ist, von sich aus auf eine Forderung nach großen Datenmengen hinaus, zumal die Prämisse, *vollständige* Gespräche zu erheben, bereits zu relativ umfangreichen einzelnen Datensätzen führen kann.³ Zum anderen gilt auch für mündliche Korpora, dass größere Datenmengen notwendig sein können, um ausreichende Belegzahlen für bestimmte Phänomene und Untersuchungszwecke zu erhalten, was etwa für die lexikografische Untersuchung gesprochener Sprache relevant ist.⁴ Da Worthäufigkeiten auch in mündlichen Korpora einer Zipfschen Verteilung folgen, gilt dies in besonderem Maße, wenn sich das Interesse auf seltene Wörter oder Phänomene richtet.

³ So enthält FOLK z.B. ein Gespräch, das während einer Spielinteraktion („Fußball-Manager“, FOLK_E_00021) aufgezeichnet wurde, die etwa 5:30 h dauerte. Dies entspricht 56.532 transkribierten Wörtern und damit bereits etwa der Größenordnung eines halben Romans.

⁴ Nach einer – sicherlich bestreitbaren – Aussage von Hanks (2012, S. 61) dürften sich anhand von FOLK mit seinem derzeitigen Umfang noch gar keine lexikografisch fruchtbaren Untersuchungen anstellen lassen: „Early electronic corpora [...] had little impact on lexicography [...]. The reason was simple: these were each corpora of only one million words – corpora so small that it was impossible to distinguish statistically significant co-occurrences of words from chance co-occurrences.“

8.2 Wie viel? Größenvergleiche

Dem mengenmäßigen Ausbau mündlicher Korpora sind wegen des Aufwands, den Feldzugang und Transkription bedeuten (siehe 8.3.1 und 8.3.2), enge Grenzen gesetzt. Einen Umfang von Hunderten Millionen oder mehreren Milliarden Wörtern wie bei schriftsprachlichen Korpora anzustreben, ist daher utopisch. Da auch die stark zunehmende Verfügbarkeit digitaler Daten keinen nennenswerten Einfluss auf die Möglichkeiten zum Aufbau mündlicher Korpora hat,⁵ ist weiterhin auch keine qualitative Änderung der Wachstumsraten zu erwarten – dem teilweise exponentiellen Wachstum schriftsprachlicher Korpora wird bis auf Weiteres nur ein annähernd lineares Wachstum mündlicher Korpora entgegenstehen. Mündliche Korpora bewegen sich somit prinzipiell in anderen Größenordnungen als schriftsprachliche.

Als erste Orientierung mögen daher die 10 Millionen Wörter dienen, die das DWDS-Korpus für jede Dekade des 20. Jahrhunderts enthält. Wie das Beispiel einiger existierender Sammlungen zeigt, ist dies eine Dimension, die für mündliche Korpora heute prinzipiell erreichbar ist. So beträgt der Anteil mündlicher Daten am British National Corpus genau 10 Millionen Wörter, das Korpus des gesprochenen Niederländisch ist mit knapp 9 Millionen Wörtern ähnlich umfangreich, und auch am IDS existieren mit den Korpora „Deutsche Mundarten“ und „Deutsch heute“ zwei Datensammlungen von etwa 1.000 Stunden, die gleichfalls auf etwa 10 Millionen Wörter kommen würden, so sie vollständig transkribiert wären. Da diese Datensammlungen allerdings zu größeren Teilen aus elizitierten (z.B. Interviews und Map-Tasks) und/oder monologischen (z.B. Ansprachen) Daten bestehen, ist der Aufwand für ihre Erstellung geringer anzusetzen als bei einem Gesprächskorpus wie FOLK. „Echte“ Gesprächskorpora sind daher kleiner – z.B. umfasst das Repository of Spontaneous Spoken Czech ca. 3 Millionen Wörter und das mehrsprachige GeWiss-Korpus (Fandrych/Meißner/Slavcheva 2012) ca. 1.2 Millionen Wörter. In den IDS-Beständen ist das Freiburger Korpus mit ca. 600.000 Wörtern das größte historische Gesprächskorpus. Da das FOLK-Projekt als Dauervorhaben des IDS eine im Vergleich zu anderen Korpuserhebungen relativ langfristige Perspektive hat, scheint eine Zielvorgabe von 10 Millionen Wörtern hier aber sinnvoll. Ob, wie, und in welchem Zeitraum bzw. mit welchen Kosten sich diese 10 Millionen Wörter erreichen lassen, wird im Folgenden diskutiert.

⁵ Allerdings bietet sich durch die Möglichkeit der Vernetzung verschiedener Ressourcen im digitalen Medium, wie sie beispielsweise in CLARIN angestrebt wird, durchaus eine Gelegenheit, die nutzbare Datenbasis deutlich zu vergrößern.

8.3 Wie? Flaschenhalse beim Aufbau von FOLK

Was also hindert FOLK daran, schnell zu einem Korpus von 10 Millionen Wörtern anzuwachsen? Die etwa 1.000 Stunden Aufnahmen, die hierfür nötig wären, könnten mit moderner digitaler Aufnahmetechnik relativ einfach und kostengünstig erhoben werden. Die Probleme stellen sich also nicht bei der Akquise des Rohmaterials „Gesprächsaufzeichnung“, sondern hängen zum einen mit dem Ziel einer breiten Streuung von Gesprächstypen, zum anderen mit der Notwendigkeit einer adäquaten Erschließung der Daten zusammen. Die beiden entscheidenden Flaschenhälse sind dabei der Feldzugang und die Transkription.

8.3.1 Flaschenhals „Feldzugang“

Die Aufzeichnung authentischer Gespräche im Feld setzt einen geeigneten Feldzugang, voraus. In den allermeisten Fällen werden sich Personen nur dann darauf einlassen, ihre sprachlichen Aktivitäten für wissenschaftliche Zwecke aufzeichnen zu lassen, wenn ein entsprechendes Vertrauensverhältnis zur Aufnahmeleitung gegeben ist. Dieses wiederum ist am einfachsten durch private, persönliche Kontakte herzustellen, zumal in einigen Fällen schon eine enge Vertrautheit mit den Gegebenheiten des privaten oder institutionellen Umfelds notwendig ist, um überhaupt für FOLK geeignete Gesprächsanlässe identifizieren zu können. Der in vielen variationslinguistischen Projekten verfolgte effiziente Ansatz, Sprecher zentral gesteuert (z.B. über ausgewählte Schulen im Bundesgebiet wie bei „Deutsch heute“) zu akquirieren und Aufnahmen mit diesen Sprechern dann von einigen wenigen Projektmitarbeitern konzertiert durchführen zu lassen, ist daher für FOLK nicht praktikabel.

In der ersten Erhebungsphase wurden stattdessen private Kontakte von Projektmitarbeitern sowie von Teilnehmern einer regelmäßig abgehaltenen Lehrveranstaltung genutzt, um geeignete Gesprächsanlässe zu identifizieren und mit dem Einverständnis der Gesprächsteilnehmer aufzeichnen zu lassen. Ergänzt wurde dies durch Aufnahmen aus dem IDS-eigenen Projekt „Variation des gesprochenen Deutsch“ sowie durch Datenspenden externer Projekte (insbesondere universitäre Prüfungsgespräche aus dem GeWiss-Projekt). Um die regionale Verteilung der Aufnahmen auszugleichen, werden in der derzeitigen Ausbauphase verstärkt Kooperationen mit Institutionen in anderen Teilen der Bundesrepublik gesucht, die Expertise in der Erhebung mündlicher Daten haben und persönliche Kontakte vor Ort für einen geeigneten Feldzugang nutzen. Daten aus einer solchen Kooperation mit dem „Hamburger Zentrum für Sprachkorpora“ und dem Projekt „Gesprochene Sprache im Ruhrgebiet“ werden derzeit aufbereitet. Hinzu kommen

weitere Datenspenden, etwa aus dem Projekt „Sprachvariation in Norddeutschland“ und aus dem Dissertationsprojekt von Hee (2012).

Diese Form des schrittweisen Ausbaus eines Netzwerks von Feldkontaktpersonen, Aufnahmeleitern und Kooperationspartnern (vielleicht als eine bescheidene Form des „Crowd-Sourcing“ aufzufassen) ist zwar aufwändiger als eine zentral geplante Datenerhebung, hat sich für FOLK aber bislang als praktikabel und adäquat erwiesen. Es steht zu hoffen, dass sich mit einer fortschreitenden Routinisierung dieser Abläufe die Erschließung neuer Feldzugänge bzw. auch die Akquise weiterer Datenspenden mittelfristig noch effizienter gestalten lässt, der Flaschenhals „Feldzugang“ sich mithin also erweitern lassen sollte.

8.3.2 Flaschenhals „Transkription“

Dass der Aufwand, authentische Gespräche adäquat zu transkribieren, enorm ist, ist jedem mit der Materie vertrauten Forscher bewusst. Je nach Qualität und Beschaffenheit der Aufnahme und Anforderungen des Transkriptionssystem sind im seltenen Idealfall (gute Qualität, diszipliniertes standardnahes Gespräch ohne Simultanpassagen und Störungen, einfache Verbal-Transkription) pro Stunde Aufnahme mindestens 20 Stunden Transkriptionsarbeit zu veranschlagen. Suboptimale Fälle (Störgeräusche in der Aufnahme, Sprecherüberlappungen, dialektale Passagen etc.) sind häufiger und führen zu Transkriptionsverhältnissen bis zu 1:100. Rechnet man die Erfahrungswerte aus dem bisherigen Projektverlauf hoch, gelangt man auch unter Annahme einer deutlichen Beschleunigung durch Routinisierung und Optimierung der Abläufe zu dem Schluss, dass 15 weitere Jahre mit gleichbleibender Ausstattung notwendig wären, um in FOLK die oben veranschlagten 10 Millionen Wörter zu erreichen. Der Frage, wie sich die Transkription effektivieren lässt, kommt damit für den mengenmäßigen Ausbau mündlicher Korpora eine zentrale Bedeutung zu. Sie lässt sich unter methodischen, technologischen oder organisatorischen Gesichtspunkten betrachten.

Unter den methodischen Aspekt fällt zum einen die Materialauswahl. Das in anderen Zusammenhängen verfolgte Konzept, die Transkription durch eine Ausparung „schwieriger“ Passagen (also z.B. solchen mit vielen Überlappungen oder Störgeräuschen) zu beschleunigen bzw. von vorneherein kein solches Material zu erheben, kommt für FOLK mit seinem Anspruch, vollständige authentische Gespräche in ihrer gesamten Variationsbreite zu erfassen, nicht in Frage. Zum anderen ist auch die methodisch motivierte Wahl des Transkriptionssystems geeignet, zur Reduzierung des Transkriptionsaufwandes beizutragen. In FOLK ist dies durch eine Entscheidung für das GAT-Minimaltranskript, in dem der Transkribent von besonders interpretationshaltigen und damit zeitaufwändigen Entscheidun-

gen (wie z.B. der Markierung von Fokusakzenten oder der Unterteilung von Beiträgen in Intonationsphrasen) befreit wird,⁶ jedoch bereits geschehen. Weitere Vereinfachungen (etwa eine weniger präzise Markierung von Simultanpassagen) sind mit Blick auf die angestrebte Qualität der Transkriptionen und ihre Verwendung zu gesprächsanalytischen Zwecken kaum zu verantworten.

Mehr Spielraum bestünde grundsätzlich in organisatorischen Aspekten der Transkription. Die Koordination der Arbeit von 12–15 studentischen Hilfskräften, die im Umfang von 20–40 Stunden pro Monat mit der Transkription betraut sind, nimmt derzeit einen erheblichen Anteil der Projektressourcen in Anspruch. Da diese Arbeitsverhältnisse selten länger als zwei Jahre dauern, müssen permanent neue Kräfte angeworben, in der Transkriptionstätigkeit geschult und zumindest anfänglich intensiv betreut werden. Da die Kompetenz von Transkribenten einen ganz wesentlichen Faktor im Transkriptionsaufwand darstellt – ein geübter und routinierter Transkribent ist nach Erfahrungen im Projekt bis zu viermal schneller als ein ungeübter – wäre wahrscheinlich ein erheblicher Effizienzgewinn zu erzielen, wenn Transkribenten auf andere Weise – z.B. durch längerfristige Arbeitsverträge mit höheren Stundenzahlen – beschäftigt werden könnten. Dies scheitert derzeit allerdings noch an administrativen Vorgaben und der Lebensrealität von Studierenden im Bachelor-/Master-System.

Es verbleiben technische Maßnahmen zur Effektivierung der Transkription, wie sie z.B. auch schon in Brinckmann (2009) diskutiert werden. Eine vollständige Automatisierung ist zwar außer Reichweite; die Hoffnung, durch geeigneten vorbereitenden Einsatz von Sprachtechnologie zu substanziellen Zeiteinsparungen zu gelangen, scheint aber nicht mehr vermessen. So kann etwa eine Pausenerkennung („silence detection“) zu einer Vorsegmentierung des Transkripts genutzt werden, über eine Sprechererkennung („speaker diarization“) eine Vorsegmentierung mit einer Sprecherzuweisung verbunden werden und eine Spracherkennung („speech recognition“) gar eine erste Rohversion des Transkripts liefern. Untersuchungen im AvaTech-Projekt⁷ gelangen so zu einer Effizienzsteigerung um ca. 50%, und erste diesbezügliche Experimente in FOLK mit einem kommerziellen Spracherkennung deuten auf ein ähnliches Potenzial hin. Allerdings wurde dieses konkret nur in einem einzigen Fall – einer relativ standardnahen und qualitativ hochwertigen Aufnahme der Schlichtungsverhandlungen zu Stuttgart 21 – realisiert, wäh-

⁶ Die Zeitersparnis beim Transkribieren ist allerdings nicht der einzige Grund für diesen Verzicht. Solche prosodischen Annotationen sind zudem nur für einen Teil des Zielpublikums von FOLK relevant und außerdem wegen ihrer Interpretationsabhängigkeit wenig reliabel.

⁷ „The resulting efficiency for audio segmentation with recognizers increased by 49%.“ (https://tla.mpi.nl/projects_info/avatech/avatech-results/, Stand: 15.8.2014).

rend in allen anderen Fällen die Ergebnisse der Spracherkennung schlicht unbrauchbar waren. Dennoch ist vorstellbar, dass durch geeignete Adaption und gezielten Einsatz solcher sprachtechnologischer Methoden mittelfristig ein Weg zur Überwindung des Transkriptions-Flaschenhalses gefunden werden kann.

9 Gemeinsamkeiten und Unterschiede schriftlicher und mündlicher Korpora

Schriftliche und mündliche Korpora am IDS blicken auf eine lange Tradition zurück. Ihre gegenwärtige Situation ist geprägt von der fortschreitenden Entwicklung digitaler Methoden, die sowohl neue Möglichkeiten als auch neue Herausforderungen mit sich bringt. Obwohl schriftliche und mündliche Korpora im Detail sehr unterschiedliche Arbeitsweisen erfordern, bestehen einige wichtige Gemeinsamkeiten, die insbesondere Grundannahmen und daraus folgende methodologische Konsequenzen betreffen. So herrscht z.B. Einigkeit darüber, dass es keine allgemeingültig repräsentativen oder ausgewogenen Korpora gibt und daher beim Korpusaufbau neben Diversität nur absolute und keine relativen Größen und Strata-Besetzungen relevant sind. Die Komposition fragestellungsspezifisch proportionierter Stichproben sollte entsprechend in der Nutzungsphase mittels virtueller Korpora erfolgen. Diesen Gemeinsamkeiten stehen erhebliche Unterschiede entgegen, was die derzeit größten und dringlichsten Herausforderungen betrifft. Während bei den schriftlichen Korpora diese vor allem bei der Akquisition bestimmter Strata bestehen und bei der Entwicklung eines Rechtersystems, das mit Daten im Petabyte-Bereich umgehen kann, liegen sie bei den mündlichen Korpora vor allem bei Erhebung (Feldzugang) und Erschließung (Transkription) von Primärdaten, denen mit der Entwicklung geeigneter technischer, methodischer und organisatorischer Praktiken begegnet werden kann.

Literatur

- al-Wadi, Doris (1994): COSMAS – Ein Computersystem für den Zugriff auf Textkorpora. Benutzerhandbuch. Mannheim.
- Aston, Guy/Burnard, Lou (1998): The BNC Handbook: Exploring the British National Corpus with SARA. (= Edinburgh Textbooks in Empirical Linguistics). Edinburgh.
- Bański, Piotr/Fischer, Peter M./Frick, Elena/Ketzan, Erik/Kupietz, Marc/Schnober, Carsten/Schonefeld, Oliver/Witt, Andreas (2012): The new IDS Corpus Analysis Platform: Challenges and prospects. In: Calzolari (Hg.), S. 2905–2911.

- Belica, Cyril (1996): Analysis of temporal changes in corpora. In: *International Journal of Corpus Linguistics* 1, S. 61–73.
- Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lüngen, Harald (2011): The morphosyntactic annotation of DeReKo: Interpretation, opportunities, and pitfalls. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich Hermann (Hg.): *Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.–24.9.2009.* (= *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache* 1). Tübingen, S. 451–469.
- Benešová, Lucie et al. (2014): Building a data repository of spontaneous spoken Czech. In: Ruhi, Şükriye et al. (Hg.): *Best practices for spoken corpora in Linguistic Research.* Cambridge, S. 129–142.
- Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. In: *Sprachreport* 3/2005, S. 2–5.
- Brinckmann, Caren (2009): Transcription bottleneck of speech corpus exploitation. In: Lyding, Verena (Hg.): *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics.* (= *Arbeitshefte Europäische Akademie Bozen* 54). Bozen.
- Brünner, Gisela (2005): Kommunikation in institutionellen Lehr-Lern-Prozessen. Diskursanalytische Untersuchungen zu Instruktionen in der betrieblichen Ausbildung. Radolfzell am Bodensee.
- Calzolari, Nicoletta et al. (Hg.) (2012): *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 12).* Istanbul.
- Church, Kenneth/Mercer, Robert L. (1994): Introduction to the special issue on computational linguistics using large corpora. In: *Computational Linguistics* 19, S. 1–24.
- Clyne, Michael (1981): Deutsch als Muttersprache in Australien. Zur Ökologie einer Einwanderersprache. (= *Deutsche Sprache in Europa und Übersee* 8). Wiesbaden.
- Davies, Mark (2010): The corpus of contemporary American English as the first reliable monitor corpus of English. In: *Literary & Linguistic Computing* 25, S. 447–464.
- Deppermann, Arnulf/Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch“ (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard et al. (Hg.): *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen.* (= *Linguistik – Impulse & Tendenzen* 44). Berlin/New York, S. 414–450.
- Dittmar, Norbert/Bredel, Ursula (1999): *Die Sprachmauer. Die Verarbeitung der Wende und ihre Folgen in Gesprächen mit Ost- und WestberlinerInnen.* Berlin.
- Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012): The GeWiss corpus: Comparing spoken academic German, English and Polish. In: Schmidt, Thomas/Wörner, Kai (Hg.): *Multilingual corpora and multilingual corpus analysis.* S. 319–337. Amsterdam.
- Fiehler, Reinhard/Wagener, Peter (2005): Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 6, S. 136–147. Internet: www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2005/px-fiehler.pdf (Stand: 5.8.2014).
- Gray, Jim (2003): *Distributed computing economics.* Technical Report MSR-TR-2003-24, Redmond.
- Hanks, Patrick (2012): Corpus evidence and electronic lexicography. In: Granger, Sylviane/Paquot, Magali (Hg.): *Electronic lexicography.* Oxford, S. 57–82.
- Hee, Katrin (2012): *Polizeivernehmungen von Migranten. Eine gesprächsanalytische Studie interkultureller Interaktionen in Institutionen.* (= *OraLingua* 3). Heidelberg.

- Hellmann, Manfred W. (Hg.) (1984): Ost-West-Wortschatzvergleiche. Maschinell gestützte Untersuchungen zum Vokabular von Zeitungstexten aus der BRD und der DDR. (= Forschungsberichte des Instituts für deutsche Sprache 48). Tübingen.
- IDS (2014): Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2014-I, Institut für Deutsche Sprache, Release vom 15.4.2014. Internet: www.ids-mannheim.de/kl/projekte/korpora/archiv.html (Stand 15.4.2014).
- Institut für Deutsche Sprache (Hg.) (2014): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim.
- Jakubíček, Miloš/Kilgarriř, Adam/Kovář, Vojtěch/Rychlý, Pavel/Suchomel, Vít (2013): The TenTen corpus family. In: Abstract Book of the 7th International Corpus Linguistics Conference CL2013. Lancaster, S. 125–127.
- Keibel, Holger/Kupietz, Marc (2009): Approaching grammar: Towards an empirical linguistic research programme. In: Minegishi/Kawaguchi (Hg.), S. 61–76.
- Kilgarriř, Adam (2001): Comparing corpora. In: International Journal of Corpus Linguistics 6, S. 97–133.
- Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28, S. 71–97.
- Kupietz, Marc (2014): Der Programmbereich Korpuslinguistik am IDS: Gegenwart und Zukunft. In: Institut für Deutsche Sprache (Hg.), S. 320–328.
- Kupietz (in Vorb.): Constructing a Corpus. In: Durkin, Philip (Hg.): Oxford handbook of lexicography. Oxford.
- Kupietz, Marc/Bankhardt, Christina (Hg.) (2009): D-SPIN Report R7.1 – Legal Aspects in the Provision of Language Resources: The German Context. Internet: http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.1.pdf (Stand: 15.6.2014).
- Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). Valetta, S. 1848–1854.
- Kupietz, Marc/Frick, Elena (2013): Korpusanalyseplattform der nächsten Generation. In: Kratochvílová, Iva/Wolf, Norbert Richard (Hg.): Grundlagen einer sprachwissenschaftlichen Quellenkunde. (= Studien zur Deutschen Sprache 66). Tübingen, S. 27–36.
- Kupietz, Marc/Keibel, Holger (2009a): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi/Kawaguchi (Hg.), S. 53–59.
- Kupietz, Marc/Keibel, Holger (2009b): Gebrauchsbasierte Grammatik: Statistische Regelmäßigkeit. In: Konopka, Marek/Strecker, Bruno (Hg.): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. (= Jahrbuch des Instituts für Deutsche Sprache 2008). Berlin/New York, S. 33–50.
- Kupietz, Marc/Lüngen, Harald (2014): Recent developments in DeReKo. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14). Reykjavik, S. 2378–2385.
- Kupietz, Marc/Lüngen, Harald/Baňski, Piotr/Belica, Cyril (2014): Maximizing the potential of very large corpora. In: Kupietz, Marc et al. (Hg.): Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2). Reykjavik, S. 1–6.
- Lüngen, Harald/Kupietz, Marc/Belica, Cyril/Perkuhn, Rainer (2011): Strategien zur Weiterentwicklung von DeReKo 2012 bis 2020. Tech. Rep. KT-2011-01. Mannheim.
- Minegishi, Makoto/Kawaguchi, Yuji (Hg.) (2009): Working papers in corpus-based linguistics and language education 3. Tokio.

- Ochs, Elinor (1979): Transcription as theory. In: Ochs, Elinor/Schieffelin, Bambi B. (Hg.): *Developmental pragmatics*. New York, S. 43–72.
- Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): *Korpuslinguistik*. Paderborn.
- Schmidt, Thomas (2005a): Computergestützte Transkription. Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. (= *Sprache, Sprechen und Computer* 7). Frankfurt a.M.
- Schmidt, Thomas (2005b): Datenarchive für die Gesprächsforschung. Perspektiven, Probleme und Lösungsansätze. In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 6, S. 103–126. Internet: www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2005/px-schmidt.pdf (Stand: 5.8.2014).
- Schmidt, Thomas (2007): Transkriptionskonventionen für die computergestützte gesprächsanalytische Transkription. In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 8, S. 229–241. Internet: www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2007/px-schmidt.pdf (Stand: 5.8.2014).
- Schmidt, Thomas (2011): A TEI-based approach to standardising spoken language transcription. In: *Journal of the Text Encoding Initiative* 1. Internet: <http://jtei.revues.org/142> (Stand: 5.8.2014).
- Schmidt, Thomas (2012): EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In: Calzolari (Hg.), S. 236–240.
- Schmidt, Thomas/Dickgießer, Sylvia/Gasch, Joachim (2013): *Die Datenbank für Gesprochenes Deutsch – DGD2*. Mannheim 2013. URN: urn:nbn:de:bsz:mh39-12747.
- Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, S. 353–402. Internet: www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2009/px-gat2.pdf (Stand: 5.8.2014).
- Stift, Ulf-Michael/Schmidt, Thomas (2014): *Das Archiv für Gesprochenes Deutsch*. In: *Institut für Deutsche Sprache* (Hg.), S. 360–375.
- Teubert, Wolfgang/Belica, Cyril (2014): Von der linguistischen Datenverarbeitung am IDS zur „Mannheimer Schule der Korpuslinguistik“. In: *Institut für Deutsche Sprache* (Hg.), S. 298–319.