

Thomas Schmidt

9 Gesprächskorpora

Aktuelle Herausforderungen für einen besonderen Korpusstyp

Abstract: Dieser Beitrag setzt sich mit Gesprächskorpora als einem besonderen Typus von Korpora gesprochener Sprache auseinander. Es werden zunächst wesentliche Eigenschaften solcher Korpora herausgearbeitet und einige der wichtigsten deutschsprachigen Gesprächskorpora vorgestellt. Der zweite Teil des Beitrags setzt sich dann mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) auseinander. FOLK hat sich zum Ziel gesetzt, ein wissenschaftsöffentliches Korpus von Interaktionsdaten aufzubauen, das methodisch und technisch dem aktuellen Forschungsstand entspricht. Die Herausforderungen, die sich beim Aufbau von FOLK in methodischer und korpustechnologischer Hinsicht stellen, werden in abschließenden Abschnitt diskutiert.

Keywords: Gesprächsforschung, gesprochene Sprache, Interaktion, Korpuslinguistik

1 Einleitung

Wenn gesprochene Daten in der Korpuslinguistik generell schon eine Sonderrolle einnehmen (siehe Mair in diesem Band), so stellen Gesprächskorpora noch einmal einen besonderen Fall unter den mündlichen Korpora dar, der ganz eigene Forschungsperspektiven und methodisch-technische Herausforderungen mit sich bringt. Um diese Perspektiven und Herausforderungen soll es im vorliegenden Beitrag gehen. Ich arbeite in Abschnitt 2 zunächst wesentliche Eigenschaften solcher Korpora heraus, die auch dazu dienen, sie von anderen Korpora mündlichen Sprachgebrauchs zu unterscheiden. Abschnitt 3 stellt dann einige der wichtigsten deutschsprachigen Korpora vor und geht auf die Problematik ein, dass ältere Sammlungen von Gesprächsdaten oft kaum für eine korpuslinguistische Nachnutzung zur Verfügung stehen. Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), das in Abschnitt 4 vorgestellt

Thomas Schmidt, Institut für Deutsche Sprache, R5, 6–13, D-68161 Mannheim,
E-Mail: thomas.schmidt@ids-mannheim.de

 Open Access. © 2018 Thomas Schmidt, publiziert von De Gruyter.  Dieses Werk ist lizenziert unter der Creative Commons Attribution 4.0 Lizenz.
<https://doi.org/10.1515/9783110538649-010>

wird, hat sich vor diesem Hintergrund zum Ziel gesetzt, der Forschergemeinschaft ein großes, breit diversifiziertes Korpus von Interaktionsdaten zur Verfügung zu stellen. Aktuelle Herausforderungen, die sich beim Aufbau von FOLK in methodischer und korpustechnologischer Hinsicht stellen, werden dann in Abschnitt 5 diskutiert.

2 Gesprächskorpora als besondere mündliche Korpora

Unter einem Gesprächskorpus soll hier ein Korpus verstanden werden, das folgende Eigenschaften besitzt:

1. Die Primärdaten sind Audio- und/oder Videoaufzeichnungen von *Gesprächen*, also von verbaler Interaktion zwischen zwei oder mehr Teilnehmer(innen).
2. Die aufgezeichneten Gespräche sind *authentische* (natürliche) Interaktionen in dem Sinne, dass sie nicht eigens vom Forscher veranlasst wurden (wie es etwa bei einem Sprachexperiment, einem Interview oder laborphonetischen Daten der Fall ist).¹
3. Die aufgezeichneten Äußerungen bestehen weitestgehend aus *spontanen* Äußerungen in dem Sinne, dass sie in ihrer konkreten Form nicht umfassend vorgeplant wurden (wie es etwa bei einer abgelesenen Rede oder einem geskripteten Dialog der Fall wäre).
4. Die Aufzeichnungen sind *vollständig*, erstens in dem Sinne, dass nicht nur ein zeitlicher Ausschnitt aus der Interaktion aufgezeichnet wird, zweitens auch in dem Sinne, dass die Aufzeichnung die Beiträge *aller* Interaktionsteilnehmer(innen) gleichberechtigt umfasst (und nicht etwa durch die Aufnahmetechnik nur ausgewählte Teilnehmer fokussiert werden).

In der Gesamtheit dieser Eigenschaften unterscheiden sich Gesprächskorpora von anderen mündlichen Korpora, insbesondere von den meisten Variations-

¹ Die Authentizität steht dabei immer in einem Spannungsverhältnis zum Beobachter-Paradoxon, das Labov (1972: 209) wie folgt beschreibt: „[T]he aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation.“ Die Gesprächsforschung hat sich mit diesem Paradoxon eingehend auseinandergesetzt und Methoden entwickelt, die negativen Auswirkungen des Paradoxons auf die Authentizität von Gesprächsdaten zu minimieren (z. B. Lalouchek & Menz 2002).

korpora (siehe Kehrein & Vorberger i. d. Bd. und Boas & Fingerhuth i. d. Bd.) und den meisten Sammlungen mündlicher Daten, die in der Phonetik oder Sprachtechnologie zum Einsatz kommen (siehe Draxler & Schiel in diesem Band). Die Grenzen zwischen einem Gesprächskorpus und anderen Typen mündlicher Korpora sind im Einzelfall allerdings nicht eindeutig zu ziehen: Einerseits weisen auch als Gesprächskorpora konzipierte Datensammlungen Merkmale auf, die es erlauben, variationslinguistische, phonetische oder sprachtechnologische Untersuchungen an ihnen auszuführen – beispielweise beinhaltet ein Korpus wie FOLK natürlich auch regionalsprachliche Variation. Andererseits lassen sich auch anders konzipierte Korpora teilweise unter gesprächsanalytischer Perspektive betrachten – beispielsweise können auch biographische oder narrative Interviews, wie sie für die dialektologische Tradition (etwa im Korpus „Deutsche Mundarten“, siehe auch Kehrein & Vorberger in diesem Band) charakteristisch sind, als spezielle Formen von Gesprächen untersucht werden. Es geht hier daher nicht um eine absolute taxonomische Abgrenzung, sondern darum, den Begriff „Gesprächskorpus“ für solche Korpora zu reservieren, bei deren Design, Umsetzung und Anwendung der Gedanke von Sprache im interaktiven Handeln leitend ist.

In ihrem Datenverständnis sind Gesprächskorpora vor allem von gesprächsanalytischen Forschungsansätzen wie der Konversationsanalyse, der Interaktionalen Linguistik oder der Funktionalen Pragmatik geprägt. Dies zieht zum einen eine besondere Aufmerksamkeit seitens der Korpusersteller und -nutzer für die sozialen Hintergründe und Zusammenhänge, denen die Daten entstammen und die sich in ihnen widerspiegeln, nach sich. Zum anderen ergibt sich daraus auch eine gewisse Zurückhaltung in der apriorischen Kategorisierung und „Kontrolle“ von Variablen des Korpus-Designs, die sich einer korpuslinguistischen Methodik – zumindest in einer stärker quantitativen Orientierung – zunächst entgegenstellt. Da bislang weder die Gesprächsanalyse über ein fortgeschrittenes Instrumentarium zur Korpusanalyse verfügt noch die Korpuslinguistik sich umfassend mit dem Datentyp „Gespräch“ auseinandergesetzt hat, erfordern der Aufbau und die Analyse von Gesprächskorpora also auch methodische Innovationen (siehe dazu auch Schmidt 2014a).

Zu den ganz konkreten methodischen Fragen, die sich für den Typus „Gesprächskorpus“ exklusiv oder zumindest in besonderem Maße stellen, gehören:

- Welche Kontextinformationen oder Zusatzmaterialien sind zu einem möglichst vollständigen Verständnis eines Gesprächs und zu seiner Interpretation notwendig? Hierzu gehören zum einen Metadaten, die beispielsweise die institutionelle Einbettung eines Gesprächs oder die sozialen Rollen der Teilnehmer(innen) hinreichend genau beschreiben. Zum anderen können hier auch Objekte, die von der Aufnahme nicht (vollständig) erfasst wer-

den, eine Rolle spielen, wie z. B. PowerPoint-Folien bei einem mündlichen Vortrag, Zeichnungen oder Notizen, die im Laufe einer Besprechung angefertigt werden.

- Auf welche Art und Weise werden die Audio- oder Video-Daten auf eine schriftliche und damit automatisch durchsuchbare Form abgebildet? Die theorie- und forschungsfragenabhängige, modellhafte Relation zwischen den Primärdaten (den Aufnahmen) und ihrer Erschließung in Form von Transkription ist vielfach thematisiert worden (dazu Ochs 1979 und bspw. Schmidt 2005a). Sie spielt auch im Kontext korpuslinguistischer Herangehensweisen eine fundamentale Rolle, denn letztendlich bilden ja die schriftlichen Abbilder, nicht die Aufnahmen selbst, den Ausgangspunkt jeder Analyse. Die folgenden Fragen können auch als Teilaspekte dieser übergeordneten Frage aufgefasst werden (siehe auch Abb. 9.1, aus der deutlich wird, in welcher Dichte diese Fragen auftreten).
- In welcher Form sollen auch nonverbale Elemente der Interaktion berücksichtigt werden? Es geht dabei um hörbare (z. B. Räuspern, Lachen, Geräusche) als auch sichtbare (Mimik, Gestik, Handlungen) Bestandteile der Aufzeichnungen, die hinsichtlich ihrer kommunikativen Relevanz beurteilt und mit geeigneten, idealerweise auch konsistent recherchierbaren, Beschreibungen versehen werden müssen.
- Wie soll die zeitliche Struktur sprachlicher Interaktionen repräsentiert werden? Hierbei geht es zum einen generell um die Frage, wie (z. B. in welcher Granularität) zeitliche Bezüge zwischen Transkription und Aufnahme („Text-Ton-Alignment“) in den Daten festgehalten werden, zum anderen ganz besonders auch um den Umgang mit zeitlicher Parallelität, die in authentischen Gesprächen allgegenwärtig ist („Überlappungen“, „Backchannelling“).
- Wie sollen Phänomene mündlicher „Performanz“ wie Pausen, Häitationen, Verschleifungen, Elisionen, Abbrüche oder Reparaturen, insbesondere auch in ihrer Relevanz für die Interaktion, in den Daten festgehalten und bei Analysen berücksichtigt werden? Wie ist mit dem Umstand umzugehen, dass die Untersuchung von authentischer gesprochener Sprache in der Regel nur unter nicht-optimalen akustischen Bedingungen erfolgen kann? Hintergrundgeräusche, schwankende Aufnahmequalität u. Ä. und die daraus resultierende schwere Verständlichkeit mancher Äußerungen können dazu führen, dass einige Teile der Interaktion weniger genau und zuverlässig beschrieben werden können als andere.

Dass diese Aspekte Gesprächskorpora zu einem besonderen Korпустyp machen, wird deutlich, wenn man ihnen den Prototyp eines schriftsprachlichen Korpus

≡	0011	AM	was
≡	0012	PB	hier bitteschön
≡	0013	AM	oh (.) danke schon zucker rein getan
≡	0014	PB	nee aber ich hab den dir hier mitgebracht
≡	0015		(0.72)
≡	0016	PB	das_s ungesüßt das schme[ckt dir]
≡	0017	AM	[das is aber kein] latte macchiato
≡	0018		(0.67)
≡	0019	PB	[[[lacht]]] das_s cappuccinopulver mit viel milch (.) das wird (dafür/das ja) ausreichen ((Störgeräusch))
≡	0020	AM	[des is n cappuccino]
≡	0021		(0.58)
≡	0022	AM	würdst du das bitte da liegen lassen

Abb. 9.1: Transkriptausschnitt aus FOLK_E_00043_SE_01_T_01 in der DGD (Alltagsgespräch – Paargespräch), mit Beispielen für die Transkription von Pausen (Zeilen 13, 15, 18, 21), Überlappungen (Zeilen 16/17 und 19/20), Verschleifungen (Zeilen 16 und 19), Elisionen (Zeilen 20 und 22), Unsicherheit des Transkribenten (Zeile 19).

redigierter, veröffentlichter Texte gegenüberstellt. Dieser „Default Case“ zeichnet sich nämlich gerade dadurch aus, dass die darin enthaltene Sprache weitestgehend losgelöst von ihrem außersprachlichen Kontext (als „Sprache der Distanz“ nach Koch & Oesterreicher 1985), frei von äußeren „Störungen“ und als einfache lineare Abfolge eindeutiger, nicht interpretationsbedürftiger sprachlicher Zeichen analysiert werden kann, die genannten Aspekte dort mit hin kaum eine Rolle spielen. Dementsprechend können Methoden, Werkzeuge und Datenmodelle, die für schriftsprachliche Korpora entwickelt und angewendet werden, in der Regel nicht direkt auf die Arbeit mit Gesprächskorpora übertragen werden, sondern es sind hierfür eigene Arbeitsabläufe notwendig (siehe z. B. Schmidt 2016b).

3 Gesprächskorpora des Deutschen

Der Ausrichtung dieses Bandes folgend beschränke ich mich hier weitestgehend auf Gesprächskorpora des Deutschen. Der Vollständigkeit halber sei jedoch eingangs festgestellt, dass auch für viele andere Sprachen Datensammlungen existieren, die nach den oben genannten Kriterien als Gesprächskorpora zu klassifizieren sind. Für das amerikanische Englisch können das *Newport Beach Corpus* (Jefferson) und das *Santa Barbara Corpus of Spoken American English* (Du Bois et al. 2000–2005), für das australische Englisch das *Griffith Corpus of Spoken Australian English* (Haugh & Chang 2013) als prototypische Vertreter gelten. Das *British National Corpus* (BNC) hat insbesondere der Erhebung des „context-governed part“ seiner „spoken component“ eine

Systematik zugrunde gelegt, aufgrund derer man das resultierende Teilkorpus als Gesprächskorpus betrachten kann. Ähnliches gilt für Teile der Neuerhebungen im *Spoken BNC2014* (Love et al. 2017) oder auch für die betreffenden Subkorpora des *Corpus Gesprochen Nederlands* (CGN, Oostdijk 2002), wobei letzteren drei Korpora gemein ist, dass sie gerade nicht als Gesprächskorpora konzipiert sind, sondern Gesprächsdaten lediglich als ein Datentypus unter anderen in das Design des Gesamtkorpus aufgenommen wurden. Für das Französische weisen das *Corpus International Écologique de la Langue Française* (CIEL-F, Dister et al. 2008) wie auch mehrere Korpora, die in der Plattform *Corpus de Langue Parlée en Interaction* (CLAPI, Groupe ICOR 2010) verfügbar sind, recht eindeutig die wesentlichen Eigenschaften eines Gesprächskorpus auf, und auch beim Aufbau des ESLO-Korpus (Eshkol-Taravella et al. 2012) werden nach anfänglicher Konzentration auf soziolinguistische Interviews neuerdings vermehrt authentische Interaktionsdaten berücksichtigt.

Für das Deutsche kann das Korpus „Grundstrukturen“ – oft besser bekannt unter dem Namen „Freiburger Korpus“ –, das als „Korpus der alltäglichen, übergruppal und überregional verstandenen und akzeptierten gesprochenen deutschen Standardsprache“ (vgl. Schröder 1975: 12) konzipiert wurde, als Wegbereiter für das Feld der Gesprächskorpora gelten. In einem Folgeprojekt wurde mit dem Korpus „Dialogstrukturen“ (Berens et al. 1976) eine weitere Datensammlung aufgebaut, die noch dezidierter auf den Interaktionsaspekt mündlichen Sprachgebrauchs fokussiert ist. Beide Korpora entstanden in Projekten des Instituts für Deutsche Sprache (IDS) und stehen am Anfang einer Hinwendung des Instituts zu soziolinguistisch und pragmatisch ausgerichteter Forschung, die in den 1980er und 1990er Jahren ihren Niederschlag auch im Aufbau weiterer Gesprächskorpora fand. Zu nennen sind hier mindestens die Korpora „Stadtsprache Mannheim“ (Kallmeyer 1994), „Beratungsgespräche“ (Nothdurft, Reitemeier & Schröder 1994), „Gespräche im Fernsehen“ (Schütte 1996), „Schlichtungs- und Gerichtsverhandlungen“ (Schröder 1997) und „Deutsch-Türkische Powergirls“ (Keim 2008), die sich jeweils als Gesprächskorpora verstehen lassen, in denen ein bestimmter Lebensraum, ein bestimmtes soziales Milieu bzw. ein bestimmter Interaktionstyp fokussiert wird.

Auch im Kontext funktional-pragmatischer Diskursanalysen entstanden in dieser Zeit größere Datensammlungen, die sich als Gesprächskorpora – hier oft mit Ausrichtung auf einen bestimmten institutionellen Kontext – qualifizieren lassen, z. B. zur „Analyse von Unterrichtskommunikation“ (Ehlich & Rehbein 1986), zur „Ausbildung im Bergbau“ (Brünner 1987) oder zu „Sprachlichen Verständigungsprozessen in der Arzt-Patienten-Kommunikation“ (Rehbein & Löning 1993), später auch mit einem speziellen Blick auf Mehrsprachigkeit wie in „Die Entwicklung narrativer Diskursfähigkeiten im Deutschen und Tür-

kischen in Familie und Schule“ (ENDFAS/SKOBI, Herkenrath & Rehbein 2012), zur „Sprache der Höflichkeit in der interkulturellen Kommunikation“ (SHiK, Rehbein et al. 2001), zu „Japanischen und deutschen Expertendiskursen in ein- und mehrsprachigen Konstellationen“ (JadEx, Hohenstein 2006) oder zum „Dolmetschen im Krankenhaus“ (DiK, Bührig et al. 2012).

An jüngeren Initiativen, die den Aufbau von Gesprächskorpora zum Gegenstand haben, sind insbesondere das „Kiezdeutschkorpus“ (KiDKo, Wiese et al. 2012) und das Korpus „Gesprochene Wissenschaftssprache Kontrastiv“ (GeWiss, Fandrych, Meißner & Slavcheva 2012) erwähnenswert, außerdem die Datenbank „Gesprochenes Deutsch für die Auslandsgermanistik“ (Imo & Weidner in diesem Band).

Mit den bislang genannten Korpora ist allerdings nur ein kleiner Teil der Gesprächsdaten erfasst, die seit der „pragmatischen Wende“ in Forschungsprojekten erhoben wurden und theoretisch für Korpusanalysen verwendet werden könnten. Ein erschöpfender Überblick scheint zum einen mangels vollständiger Informationsquellen kaum möglich (siehe aber Wagener & Bausch 1997 und Glas & Ehlich 2000). Zum anderen offenbart schon ein zweiter Blick auf die genannten Korpora das grundlegende Problem, dass viele Gesprächskorpora, die als empirische Grundlage für Einzelprojekte aufwändig erhoben und erschlossen wurden, nach Abschluss dieser Projekte oft nicht für weitere Analysen zur Verfügung stehen.² Die Gründe hierfür sind vielfältig: neben ungelösten technischen Herausforderungen und dem Mangel an Bereitschaft zum Teilen von Daten (siehe dazu auch Schmidt 2005b) fehlt bei vielen Korpora die rechtliche Grundlage für eine Weitergabe der Daten, weil bei ihrer Erhebung keine geeignete Einwilligung der Gesprächsteilnehmer eingeholt wurde. Dies ist zum Teil dem Umstand geschuldet, dass rechtliche Fragen einer Datenweitergabe bei der Erhebung nicht oder nicht ausreichend bedacht wurden bzw. – mit Blick auf die Distribution von Daten in elektronischer Form über das

² Die Korpora „Grundstrukturen“ und „Dialogstrukturen“ sind vollständig über das Archiv für Gesprochenes Deutsch (AGD) und die Datenbank für Gesprochenes Deutsch (DGD) verfügbar, die weiteren im IDS-Kontext entstandenen Korpora jedoch nur in kleineren Auszügen oder gar nicht. Das Korpus „Ausbildung im Bergbau“ wurde 2014 ins AGD integriert, die Audio- und Videoaufnahmen sind über den persönlichen Archivservice erhältlich. Die Korpora ENDFAS/SKOBI und DiK werden über das Hamburger Zentrum für Sprachkorpora (<https://corpora.uni-hamburg.de/hzsk/> [letzter Zugriff: 26. 9. 2017]) archiviert und weitergegeben, allerdings ohne das zugrunde liegende Audio. Gleiches gilt für das KiDKo, das über die ANNIS-Plattform an der HU Berlin (<https://korpling.german.hu-berlin.de/annis3/> [letzter Zugriff: 26. 9. 2017]) verfügbar gemacht wird. Für das GeWiss-Korpus wird an der Universität Leipzig eine eigene Zugriffs-Plattform (<https://gewiss.uni-leipzig.de/> [letzter Zugriff: 26. 9. 2017]) betrieben, zusätzlich wird das Korpus aktuell in die Bestände des AGD integriert. Für alle anderen genannten Korpora müssen Bedingungen der Archivierung und möglichen Weitergabe aktuell als ungeklärt gelten.

WWW – gar nicht bedacht werden konnten. Darüber hinaus fragt sich jedoch, ob die betreffenden Datensammlungen bei ihrer Entstehung überhaupt als „Korpora“ konzipiert worden waren: die Erwartung, dass sie als empirische Datengrundlage über den eigentlichen Projektkontext hinaus verwendet werden können oder gar sollten, mag aus heutiger Perspektive selbstverständlich erscheinen; für ein ethnographisch oder soziolinguistisch orientiertes Forschungsprojekt in den 1970er oder 1980er Jahren kann sich dies jedoch grundlegend anders dargestellt haben.

4 FOLK

Obwohl die empirische Arbeit mit Gesprächsdaten also in Deutschland auf eine mehrere Jahrzehnte umfassende Tradition zurückblickt, standen der germanistischen Sprachwissenschaft noch zu Beginn des Jahrtausends kaum allgemein nutzbare Gesprächskorpora zur Verfügung. Dieser Missstand war für das IDS der Anlass, ein Projekt zum Aufbau eines „nationalen Gesprächskorpus“ (Deppermann & Hartung 2012) zu initiieren:

Aufgrund dieser unbefriedigenden Situation haben wir im Jahre 2008 am IDS damit begonnen, ein nationales Gesprächskorpus aufzubauen, das den „kommunikativen Haushalt“ (Luckmann 1986) der deutschsprachigen mündlichen Kommunikationspraxis in seinen wesentlichen Ausprägungen repräsentieren soll [...]. Das regulative Ziel ist es, das volle Spektrum der privaten, institutionellen, öffentlichen und massenmedialen Anlässe und Typen mündlicher Kommunikation nach und nach durch Audio- und Videoaufnahmen zu dokumentieren, zu transkribieren und soweit als möglich der wissenschaftlichen Gemeinschaft zur Verwendung für Forschungs- und Lehrzwecke zur Verfügung zu stellen. Dementsprechend nennen wir dieses Korpus „Forschungs- und Lehrkorpus gesprochenes Deutsch“ (FOLK [...]). (Deppermann & Hartung 2012: 418)

FOLK wurde Ende 2012 erstmalig mit Version 2.0 der *Datenbank für Gesprochenes Deutsch* (DGD, Schmidt 2014b) veröffentlicht und wird seitdem kontinuierlich ausgebaut. Die aktuelle Version (vom April 2017) umfasst 259 Gespräche im Umfang von etwas mehr als 202 Stunden und ca. 2 Millionen transkribierten Tokens, die verschiedenste Interaktionstypen aus den Bereichen privater (z. B. Tischgespräche, Telefongespräche, Spielinteraktionen, Gespräche bei privaten Aktivitäten), institutioneller (z. B. schulischer Unterricht, Verkaufsgespräche, Fahrtschulstunden, berufliche Gespräche, universitäre Prüfungsgespräche) und öffentlicher Kommunikation (z. B. Podiumsdiskussion, Schlichtungsgespräch) abdecken.

Die korpustechnologischen Werkzeuge und Verfahren, die beim Aufbau von FOLK zum Einsatz kommen und zu einem nicht unerheblichen Teil eigens

für dieses Projekt entwickelt oder optimiert wurden, sind in Schmidt (2016b) näher beschrieben. Sie dienen alle dem übergeordneten Ziel, die Erstellung eines Gesprächskorpus nicht nur praktisch und technisch handhabbar zu machen, sondern die entstehenden Daten darüber hinaus anschlussfähig an gute Praktiken (dazu Schmidt 2016a) und anderweitig etablierte digitale Verfahren in der Korpuslinguistik zu machen. Somit soll sich FOLK als Gesprächskorpus bei allen Besonderheiten und Unterschieden zum „Default Case“ des schriftsprachlichen Korpus (siehe dazu auch Kupietz & Schmidt 2015) mittelfristig in ein Gesamtgefüge einordnen, in dem auch gemeinsame oder kontrastierende Untersuchungen über verschiedene Korpusstypen hinweg ermöglicht werden.

Wie aus mittlerweile über 7.000 Registrierungen für die DGD, in der FOLK das mit Abstand am meisten genutzte Korpus ist, ersichtlich ist, wird mit der Bereitstellung dieses Gesprächskorpus ein realer und großer Bedarf von Forschenden, Lehrenden und Studierenden adressiert. Eine systematische Nutzerstudie (Fandrych et al. 2016) hat gezeigt, dass die Daten in unterschiedlichsten Anwendungsszenarien zum Einsatz kommen. FOLK wird demnach außer als Basis-Ressource in der sprachwissenschaftlichen universitären Ausbildung und als Datengrundlage für gesprächsanalytische Arbeiten insbesondere auch für variationslinguistische Untersuchungen, für vergleichende Korpusanalysen, als Ressource für die Sprachvermittlung im Bereich DaF/DaZ und als Quelle für die Entwicklung sprachtechnologischer Anwendungen fruchtbar gemacht. Auffällig ist darüber hinaus ein großes Interesse an FOLK in der Auslandsgermanistik, das sich nicht zuletzt dadurch erklären lässt, dass mit FOLK Studierenden im Ausland ein einfacher Zugriff auf authentische und aktuelle Gesprächsdaten des Deutschen in größerer Vielfalt ermöglicht wird (siehe dazu auch Imo & Weidner in diesem Band).

Exemplarische Analysen, die anhand von FOLK gesprächsanalytische Methodik mit korpuslinguistischen Verfahren kombinieren, finden sich beispielsweise in Deppermann & Schmidt (2014), Schmidt (2014a) und Kaiser (2017), wo jeweils einzelne Diskursmarker (*das heißt, ich sag mal, bzw. sprich*) untersucht werden. Mehrere aus dem Projekt „Verbkomplemente im gesprochenen Deutsch“ hervorgegangene Arbeiten (vgl. Deppermann et al. 2017) widmen sich der deskriptiven und funktionalen Beschreibung von Argumentstrukturen und Verbkomplementen und stützen sich sowohl bei der Kontrastierung von Mündlichkeit und Schriftlichkeit als auch bei der Untersuchung interaktionsspezifischer Besonderheiten wesentlich auf Korpus-Evidenz aus FOLK. In ähnlicher Weise hat das Projekt „Lexik des Gesprochenen Deutsch“ (LeGeDe, Meliss & Möhrs 2017) ausgehend von FOLK begonnen, erstmalig den Wortschatz des gesprochenen Deutsch in der Interaktion mit korpuslexikographischen Methoden zu untersuchen und zu beschreiben.

5 Aktuelle Herausforderungen

Die bisherigen Arbeiten mit Daten aus FOLK machen bereits das große Potenzial deutlich, das ein wissenschaftsöffentlich verfügbares Gesprächskorpus für die sprachwissenschaftliche Forschung und Lehre birgt. Beim Aufbau von FOLK offenbaren sich aber auch eine ganze Reihe von Herausforderungen, die dieser spezielle Korpusstyp mit sich bringt und deren Bearbeitung noch lange nicht als abgeschlossen betrachtet werden kann. Diese sind zum Teil eher theoretisch-methodischer, zum Teil eher praktisch-technologischer Natur, sie interagieren aber auch vielfältig miteinander. Exemplarisch seien im Folgenden die Bereiche des Korpus-Designs und der Korpus-Technologie diskutiert.

5.1 Korpus-Design und Stratifikation

Als Referenzkorpus muss FOLK anstreben, seinen Gegenstand – Gesprächsinteraktionen im Deutschen – in möglichst großer Breite und Differenziertheit und nach einer nachvollziehbaren Systematik abzubilden. Leitend für das Korpus-Design ist dabei zunächst der Begriff des Gesprächstyps, d. h. vor allen anderen Eigenschaften sind es Unterschiede in Interaktionsanlässen, -konstellationen, -kontexten und -inhalten (i. w. S. „Situational Parameters“ nach Biber 1993: 245), die im Korpus angemessen abgebildet werden müssen. Bei schriftsprachlichen Korpora können zumindest allgemeinere Kategorien wie „Zeitungstext“, „Belletristik“, „Gebrauchstext“, „wissenschaftlicher Text“ (vgl. das „Kernkorpus“ des Digitalen Wörterbuchs der Deutschen Sprache, DWDS)³ insofern als robust und etabliert gelten, als sie in dieser oder ähnlicher Benennung und Systematik beim Design mehrerer Referenzkorpora zur Anwendung kommen. Für die Binnendifferenzierung solcher übergeordneten Kategorien stehen außerdem oft zusätzliche externe Systematiken (wie Fachsystematik für wissenschaftliche Texte, Ressorts für Zeitungstexte, literarische Gattungen für Belletristik) zur Verfügung, die für eine detaillierte Korpus-Stratifikation fruchtbar gemacht werden können. Für die Klassifizierung mündlicher Interaktionen existiert keine vergleichbar stabile Ausgangslage. Als weitestgehend einfach operationalisierbar kann allenfalls eine erste Unterscheidung in Interaktionsdomänen gelten, die ein gegebenes Gespräch z. B. dem privaten, dem institutionellen oder dem öffentlichen Bereich (so in FOLK)⁴

³ <https://www.dwds.de/> [letzter Zugriff: 07. 11. 2017].

⁴ Ähnlich z. B. bei Biber (1993: 245), wo unter dem Stichwort „Setting“ zwischen „Institutional, other public, private-personal“ unterschieden wird, oder beim slowenischen GOS-Korpus (Verdonik et al. 2013), dessen Bestandteile jeweils einer der drei Kategorien „Public, Non-pub-

zuordnet. Für eine weitere Binnendifferenzierung können bei institutionellen Gesprächen ggf. noch die betreffenden Institutionen selbst (z. B. Schule, Universität, Verein, Kirche usw.) und eventuell diesen eigene (quasi „institutionalisierte“) Typisierungen (z. B. nach Schulfach oder Klassenstufe in der Schule, Seminare/Übungen vs. Prüfungen an der Universität, Vorstandssitzung vs. Mitgliederversammlung im Verein) herangezogen werden; insbesondere im privaten Bereich ist die eindeutige Zuordnung eines gegebenen Gesprächs innerhalb einer eindeutigen Typen-Hierarchie aber oft nicht möglich, gerade weil sich private Alltagsgespräche dadurch auszeichnen, dass ihre Form nicht oder nur in geringem Maße äußerlich vorgegeben ist. Deppermann & Hartung (2012: 423 f.) schlagen daher für Gesprächskorpora eine „parametrisierte Systematik“ vor, die ein Gesprächsereignis statt durch eine einfache Zuordnung zu einem Typ durch ein Bündel von Merkmalen in Form von Attribut-Wert-Paaren charakterisiert. Angeführt werden z. B. Parameter wie „Teilnehmerzahl“ (z. B. mit Werten „dyadisch“ vs. „Mehrpersonengespräch“), „Vertrautheit der Teilnehmer“ („unbekannt“, „bekannt“, „vertraut“), „Publikum“ („ja“, „nein“) oder „Zugang“ („geschlossen“, „halb-öffentlich“, „öffentlich“). Im Hinblick auf Korpus-Design und -Ausbau (aber auch bei der Analyse) ist ein solcher Ansatz prinzipiell praktikabler, weil er im Gegensatz zu einer fixierten Gattungssystematik weniger (evtl. theoretisch strittige) Festlegungen erfordert und auch Raum lässt, Gesprächsereignisse ins Korpus-Design zu integrieren, die bei der Planung nicht vorhergesehen wurden. Allerdings ist es alles andere als trivial, eine solche Systematik für die Anwendung auf reale Gesprächsaufnahmen zu operationalisieren, denn sie erfordert u. a. die Klärung von Grenzfällen (z. B. „Ist ein Prüfungsgespräch zwischen Student und Prüfer, bei dem ein Beisitzer anwesend ist, der aber nicht aktiv am Gespräch teilnimmt, dyadisch oder ein Mehrpersonengespräch?“) und Definitionen oder Leitlinien für interpretative Entscheidungen (z. B. „Ab wann gelten Gesprächsteilnehmer als vertraut und nicht mehr als nur bekannt?“). Letztendlich bieten sich für die Operationalisierung daher korpuslinguistische Methoden an, in denen diese Parameter als globale Annotationen verstanden werden, deren Intersubjektivität durch explizite Leitlinien und Inter-Rater-Agreement-Messungen abgesichert werden kann. Somit können Design und Stratifikation eines Gesprächskorpus wie FOLK also nicht vollständig à priori „am Reißbrett“ erfolgen, sondern müssen begleitend zum Aufbau anhand des jeweils schon vorliegenden Materials em-

lic non-private, Private“ zugeordnet sind. Das CGN unterscheidet hingegen zunächst nur zwischen „Private“ und „Public“, beim „context-governed part“ des BNC kommt eine Kategorie „Public/Institutional“ zur Anwendung, die mit den Kategorien „Educational/Informative“, „Business“ und „Leisure“ kontrastiert.

pirisch entwickelt und fortwährend verifiziert werden. Nachdem FOLK in einer ersten Phase zunächst überwiegend opportunistisch (also mit Aufnahmen leicht erreichbarer Gesprächsereignisse) aufgebaut und in einer zweiten Phase mit Blick auf eine möglichst breite Streuung über (vorläufig angenommene) Gesprächstypen ausgebaut wurde, ist mit den nun vorliegenden 259 Gesprächsereignissen eine Basis gegeben, um eine parametrisierte Systematik in dieser Weise korpuslinguistisch zu fundieren, d. h. an konkretem empirischen Material zu entwickeln und zu erproben.

Neben dem Gesprächstyp sind für das Design und die Stratifikation eines „nationalen Gesprächskorpus“ jedoch auch demographische Kriterien, also Eigenschaften der aufgenommenen Sprecher, relevant. Als Mindestanforderung für FOLK kann in dieser Hinsicht gelten, dass weibliche und männliche Sprecher in vergleichbaren Mengen berücksichtigt werden, dass das Korpus regionale Variation in ausreichendem Maße abbildet und dass auch bezüglich Alter und Bildungshintergrund der Sprecher eine Ausgewogenheit – oder zumindest vollständige Abdeckung – angestrebt wird. Diese „sekundären“ Stratifikationsparameter sind zwar in der Theorie einfacher zu handhaben als die zuvor genannten, da sie weitestgehend objektiv feststellbar (und in diesem Sinne „echte“ Metadaten, keine Annotationen) sind. Allerdings stellt sich bei der praktischen Umsetzung das Problem einer kombinatorischen Explosion: Ein Korpus-Design, das alle dann festgelegten Parameter miteinander kreuzt und für jede mögliche Kombination eine Mindestmenge an Daten vorsieht (also z. B. mindestens eine Aufnahme eines dyadischen, geschlossenen Gesprächs zwischen einander vertrauten Teilnehmern aus dem norddeutschen Sprachraum mit männlichen Sprechern unter 30 Jahren mit höherem Bildungsabschluss, und desgl. für alle anderen Kombinationen), führt zwangsläufig zu Datenmengen, die praktisch nicht mehr erheb- und verarbeitbar sind. Für Design und Stratifikation von FOLK muss also ein Kompromiss gefunden werden, der demographische Kriterien nicht ignoriert, sich aber am organisatorisch Machbaren orientiert. Der aktuell in FOLK (und teilweise auch in anderen Gesprächskorpora, die sich dieser Frage stellen) favorisierte Ansatz sieht vor, zum einen die Zahl der Attribut-Wert-Kombinationen für die demographische Stratifikation möglichst gering zu halten (indem z. B. nur zwei oder drei Altersspannen oder nur vier bis sechs sprachliche Großregionen unterschieden werden), zum anderen eine systematische Streuung über solche Parameter nur für ausgewählte, möglichst alltägliche Gesprächstypen (wie privates Telefongespräch, Tischgespräch, berufliches Meeting) anzustreben.

Kompromisse bei Korpus-Design und Stratifikation werden aber nicht nur auf Grund begrenzter Kapazitäten für Datenerhebung und -verarbeitung notwendig; nicht wenige Gesprächsereignisse sind auch wegen schwieriger akus-

tischer Bedingungen (z. B. Gespräch in der Disko) oder mangelnder Vorhersehbarkeit (z. B. Unterhaltung bei einer zufälligen Begegnung) kaum erhebbar oder aufgrund erhöhter Sensibilität (z. B. psychotherapeutisches Gespräch) für ein wissenschaftsöffentliches Korpus nicht autorisierbar. Der Anspruch, in einem Gesprächskorpus „das volle Spektrum [von Gesprächen] zu dokumentieren“ (Deppermann & Hartung 2012: 418) kann daher nur als ein Ideal verstanden werden, dem man sich bestenfalls soweit annähern kann, dass die Variation der Stratifikationsparameter in einer für den Korpusnutzer nachvollziehbaren Weise maximiert wird. Da in FOLK in diesem Sinne das Prinzip „Breite vor Tiefe“ angewendet wird – der Aufnahme von Gesprächen mit bislang nicht besetzten Parameterkombinationen also üblicherweise der Vorzug vor der Erhebung weiterer Instanzen bereits vorhandener Typen gegeben wird –, ist das Korpus dann auch weniger zur Bearbeitungen solcher Fragestellungen geeignet, die ganz spezielle Interaktionspraktiken oder Sprechertypen in den Blick nehmen. FOLK versteht sich auch in dieser Hinsicht als ein Referenzkorpus, das als Vergleichsbasis für vorhandene (z. B. zur Hochschulkommunikation wie in GeWiss, zur Arzt-Patienten-Kommunikation wie in DiK, zur Kommunikation zwischen multi-ethnischen Sprecherinnen wie in KiDKo) oder als Orientierungspunkt für zukünftig zu erstellende spezialisierte Gesprächskorpora dienen kann. Dies gilt auch für Daten, die in FOLK aus organisatorischen oder prinzipiellen Erwägungen bis auf Weiteres unberücksichtigt bleiben werden, wie insbesondere Gesprächsdaten aus Österreich und der deutschsprachigen Schweiz und mehrsprachigen Kommunikationssituationen.⁵

5.2 Korpustechnologie

Wie oben angesprochen, können korpuslinguistische Verfahren zur Annotation und Analyse generell nicht einfach vom schriftsprachlichen auf den mündlichen Fall übertragen werden, und Gesprächsdaten weisen in dieser Hinsicht gegenüber anderen gesprochen sprachlichen Daten noch einmal besonders komplexe Eigenschaften auf.

Aus korpustechnologischer Sicht zentral ist zunächst die Tatsache, dass die Erhebung und Grunderschließung von Gesprächsdaten kaum durch auto-

⁵ Deren offensichtliche Relevanz soll damit in keiner Weise in Frage gestellt werden – ihre Berücksichtigung würde aber die Komplexität des Korpus-Aufbaus um zusätzliche Dimensionen erweitern. Gesprächsdaten des Österreichischen werden in einigen Teilprojekten des Spezialforschungsbereichs „Deutsch in Österreich“ erhoben. Der Aufbau eines Referenzkorpus zu Gesprächen in mehrsprachigen Konstellationen unter Beteiligung des Deutschen bleibt ein Desiderat, siehe dazu aber mehrere Beiträge in Schmidt & Wörner (2012).

matische sprach- oder texttechnologische Verfahren unterstützt werden kann (wohingegen die Akquise schriftsprachlicher Daten über geeignete Harvesting-Methoden oft fast vollständig automatisiert ist). Die Erhebung einer Gesprächsaufnahme erfordert einen geeigneten Feldzugang, der in aller Regel nur über persönliche Kontakte herzustellen ist, und auch die Aufnahme selbst muss i. d. R. von einer technisch verständigen Person vorbereitet, den Bedingungen der jeweiligen Situation angepasst und durchgeführt werden.⁶ Für die anschließende Basiserschließung, d. h. Transkription, einer Aufnahme ist prinzipiell der Einsatz von Spracherkennungstechnologie denkbar und wurde exemplarisch auch schon erprobt (siehe z. B. Moore 2015). Erste Experimente in FOLK haben aber ergeben, dass beim dort vorliegenden Material in aller Regel (d. h. von wenigen Ausnahmen abgesehen) nur Worterkennungsraten von deutlich unter 50 % erreicht werden, so dass der resultierende Korrekturbedarf letztendlich mindestens den gleichen Aufwand verursacht wie eine rein manuelle Transkription. Ähnliches gilt für einfachere sprachtechnologische Verfahren wie „silence detection“ (die Erkennung von Pausen zur Vorsegmentierung einer Aufnahme) oder „speaker diarization“ (die Erkennung von Sprechern und Sprecherwechseln in der Aufnahme), die prinzipiell geeignet erscheinen, den manuellen Transkriptionsaufwand deutlich zu reduzieren, aber in der Anwendung auf Gesprächsdaten⁷ zu fehlerhaft sind, um dieses Potenzial zu realisieren. Die Überwindung des „Transkriptionsflaschenhalses“ („transcription bottleneck“, Brinckmann 2009) mittels Sprachtechnologie bleibt daher bis auf Weiteres ein Wunschtraum.

Liegt zu einer Aufnahme erst einmal eine Transkription vor, ist es möglich, diese mit Annotationsverfahren, die ursprünglich für schriftsprachliche Daten entwickelt wurden, automatisch anzureichern. Im Falle von FOLK umfasst dies derzeit eine orthographische Normalisierung (also die Abbildung literarisch transkribierter Formen wie *zwohunmert* auf ihre standardorthographische Entsprechung *zweihundert*) und, darauf aufbauend, eine Lemmatisierung und ein

⁶ Ein gewisses Potenzial zur Zentralisierung (wenn auch nicht Automatisierung) besteht immerhin bei medial vermittelten Gesprächen, also z. B. Telefon- oder Skypegesprächen, die über geeignete Software aufgezeichnet werden können, ohne dass eine Anwesenheit des Forschers „vor Ort“ notwendig wäre. Zentral akquiriert werden können außerdem Aufnahmen aus Rundfunk und Fernsehen, sofern geeignete Abmachungen mit den Sendeanstalten vorliegen.

⁷ Dies gilt nicht unbedingt in gleichem Maße für andere Typen mündlicher Daten. Wo immer die Aufnahmebedingungen soweit kontrolliert werden können, dass durchgängig hochwertige Audiodaten mit gleichen technischen Parametern und ohne größere Störungen entstehen, erhöhen sich die Erfolgsaussichten beim Einsatz von Spracherkennungstechnologie. Am AGD wird solche Technologie daher zunächst – und teilweise bereits erfolgreich – in der Anwendung auf Variationskorpora erprobt.

Part-of-Speech-Tagging. Die einzelnen Verfahren sollen hier nicht im Detail diskutiert werden (siehe dazu Westpfahl & Schmidt 2016 und Schmidt 2016b). Wichtig ist, dass sie zwar einerseits (teil-)automatisiert werden können, andererseits aber erst dann zufriedenstellende Ergebnisse liefern, wenn sie – mit nicht unerheblichem Aufwand – an die Eigenheiten mündlicher Daten im Allgemeinen und von Gesprächsdaten im Besonderen angepasst wurden. Für FOLK und die genannten Annotationstypen ist diese Anpassung mittlerweile erfolgt, für andere automatische Annotationsverfahren, die für die korpuslinguistische Analyse schriftsprachlicher Daten fruchtbar gemacht werden (z. B. Parsing, morphologische Annotation) steht sie noch aus.

Ähnliches gilt für Verfahren der automatisierten Auswertung, die sich in der Korpuslinguistik etabliert haben, beispielsweise Kookkurrenzprofile oder Kollokationsmaße. Exemplarisch zeigen etwa Batinic & Schmidt (2017) am Beispiel der Rekonstruktion separabler Partikelverben, dass automatisierte Verfahren nicht ohne Modifikation vom schriftsprachlichen auf den mündlichen Fall übertragen werden können, z. B. weil diesen Verfahren die vermeintliche Selbstverständlichkeit zugrunde liegt, dass das zu annotierende Material aus Sätzen bestehe – was für die Gesprächsdaten in FOLK aber nicht gilt.

Schließlich erfordern Gesprächskorpora auch vom Kernstück der korpuslinguistischen Analyse – der Korpus-Query, also der gezielten und systematischen Suche nach sprachlichen Formen im Korpus – weitreichende Anpassungen. Fast allen gängigen Recherchesystemen (wie dem Corpus Query Processor CQP, dem Corpus Search, Management and Analysis System COSMAS oder der Korpusanalyseplattform KorAP) liegt ein Modell zugrunde, das Korpora als eine Menge von Texten, ggf. mit diesen zugeordneten Metadaten, behandelt und die Texte selbst als „Stream of Tokens“ (Menke et al. 2015) – also als lineare Abfolge von Wort- und Interpunktions-Tokens, ggf. mit Annotationen, die einzelne Tokens oder Token-Folgen referenzieren – betrachtet. In der Anwendung auf Gesprächsdaten greift ein solches Modell in mehrfacher Hinsicht zu kurz:

- Die Transkripte, auf denen eine Korpusrecherche im Falle von Gesprächskorpora ausgeführt wird, sind Sekundärdaten, die abschnittsweise den Primärdaten – also den Audio- oder Videoaufnahmen – zugeordnet sind. Bei der Recherche selbst kann dieser Umstand zunächst in den Hintergrund gerückt werden, bei der Präsentation des Rechercheergebnisses muss die Zuordnung aber nutzbar gemacht werden, indem ein Zugriff auf den betreffenden Abschnitt der Aufnahme ermöglicht wird.
- Nicht alle Metadaten beziehen sich auf den „Text“ (d. h. das Transkript bzw. die zugrunde liegende Gesprächsaufnahme) als Ganzes. Soziobiographische Daten der Sprecher, die für viele Analyse Zwecke sehr wichtig sein

- können, müssen jeweils nur den Beiträgen des betreffenden Sprechers – und damit nur ausgewählten „Text“-Teilen – zugeordnet werden. In einer Korpusrecherche muss diese Zuordnung nutzbar sein, z. B. indem eine Datenabfrage auf die Beiträge männlicher Sprecher aus dem norddeutschen Raum beschränkt wird.
- Neben vollwertigen Worttokens enthalten Transkripte auch andersartige Elemente, z. B. Beschreibungen nonverbalen Verhaltens, Pausen oder unvollständige Wörter (Abbrüche u. Ä.). Je nach Recherche-Interesse kann es sinnvoll sein, solche Elemente bei einer Query zu berücksichtigen oder unbeachtet zu lassen. Dies hat z. B. Auswirkungen auf die Berechnung von Token-Abständen in kontextsensitiven Suchen oder auf die Berechnung von Token-Frequenzen.
 - Wort- und andere Tokens sind in Gesprächskorpora nicht durchgängig linear angeordnet – die Reihenfolge zweier Tokens aus überlappenden Redebestandteilen verschiedener Sprecher ist z. B. nicht immer eindeutig festgelegt, oder zwei solcher Tokens können identische Positionen haben. Ein echter „Stream of Tokens“ kann daher immer nur lokal (für einzelne Sprecherbeiträge) oder für eine Teilmenge der Daten (alle Beiträge eines Sprechers) angenommen werden. Auch dies hat Konsequenzen z. B. für die Berechnung von Token-Abständen.
 - Im selben Sinne kann der Begriff „Kontext“ im Falle von Gesprächskorpora nicht auf die Tokens reduziert werden, die einem Recherchetreffer unmittelbar vorausgehen und folgen. Diese können zwar innerhalb einer Keyword-in-Context (KWIC)-Darstellung eines Rechercheergebnisses für einen kompakten Überblick genutzt werden. Zusätzlich muss aber die Möglichkeit gegeben sein, vorausgehende und folgende Beiträge des gleichen oder eines anderen Sprechers mit einzubeziehen, indem z. B. der gesamte zugehörige Transkriptausschnitt angezeigt werden kann.

Da sich insbesondere die Gesprächsanalyse vornehmlich für Strukturen und Mechanismen interaktiven Handelns interessiert, ist mit der „Textstruktur“ (d. h. der in den Transkripten abgebildeten Struktur von Turn-Organisation, Sprecherwechseln etc.) schließlich auch eine weitere Dimension in der Korpusrecherche von Interesse, die theoretisch zwar auch für schriftsprachliche Texte von Belang ist, von den meisten gängigen Korpusssystemen aber unberücksichtigt gelassen wird. Es geht hierbei um die Einschränkung einer Suche auf bestimmte strukturelle Positionen, beispielsweise „unmittelbar nach einem Sprecherwechsel“, „innerhalb einer Überlappung“ oder „in einem Beitrag mit weniger als drei Tokens“.

In der Summe führen diese zusätzlichen Anforderungen wiederum dazu, dass korpuslinguistische Recherchesysteme für die Anwendung auf Gesprächs-

The image shows four sequential screenshots of the FOLK search interface:

- Step 1:** The 'POSITION' tab is active. The 'Vorlage' dropdown is set to '(2) höchstens N Wörter nach Beginn eines Beitrags'. The 'Parameter' field contains 'N=1'. A note below states: 'Die Position wird berücksichtigt, wenn eine Tokensuche ausgeführt wird.'
- Step 2:** The 'TOKEN' tab is active. 'Transkribiert:' is 'z.B. 'kannstsch'' and 'Normalisiert:' is 'nein'. 'Lemma:' is 'z.B. 'können'' and 'POS:' is 'z.B. \VMFIN'. There is a search button 'Suche starten' and a checkbox for 'Reguläre Ausdrücke'.
- Step 3:** The 'KONTEXT' tab is active. 'Transkribiert:' is 'z.B. 'kannstsch'' and 'Normalisiert:' is 'aber'. 'Lemma:' is 'z.B. 'können'' and 'POS:' is 'z.B. \VMINF'. Context settings are 'Kontext: 1 Token' and 'rechts'. There is a 'Kontext filtern' button and a checkbox for 'Reguläre Ausdrücke'.
- Step 4:** The 'METADATEN' tab is active. The 'Deskriptor:' dropdown is set to 'S: Geschlecht' and the selected value is 'Männlich'. A button 'Metadaten anzeigen / Filter anwenden' is visible.

Abb. 9.2: Formulierung einer schrittweisen Suche auf FOLK in der DGD nach Vorkommen von „nein“, geäußert von männlichen Sprechern im unmittelbaren Kontext von „aber“ und am Beginn eines Sprecherbeitrags.

korpora umfassend angepasst oder eigens entwickelt werden müssen. Transkripte als einfache „Texte“ mit den Mechanismen der Systeme zu verarbeiten, die auf schriftsprachliche Texte ausgelegt sind, ist zwar möglich und wird auch praktiziert, z. B. bei der Integration des BNC in das Korpusportal der Brigham Young University (<https://corpus.byu.edu> [letzter Zugriff: 26. 9. 2017]). Ohne die Möglichkeiten eines Rückgriffs auf Audio- und Videoaufnahmen, einer Einschränkung von Suchen auf sprecherspezifische Metadaten oder der Berücksichtigung der Besonderheiten von Gesprächs-Tokens und deren Kontexteigenschaften bleibt aber ein Großteil des besonderen Potenzials von Gesprächskorpora ungenutzt. Ansätze, die diese Bedarfe adressieren, finden sich z. B. bei KonText, der Query Engine des *Czech National Corpus* (<https://kontext.korpus.cz> [letzter Zugriff: 26. 9. 2017]), im polnischen Portal SPOKES (<http://spokes.clarin-pl.eu/> [letzter Zugriff: 26. 9. 2017]) oder in der französischen CLAPI-Plattform (<http://clapi.ish-lyon.cnrs.fr> [letzter Zugriff: 26. 9. 2017]). Für die Recherche auf FOLK in der DGD (<http://dgd.ids-mannheim.de> [letzter Zugriff: 26. 9. 2017]) wird ein schrittweise verfeinerbares Verfahren der Korpusrecherche angeboten, in dem gesprächsstrukturelle Constraints, Eigenschaften von Tokens und Tokens im Kontext sowie sprecherspezifische Metadaten einbezogen und miteinander kombiniert werden können (Abb. 9.2).

Bei der Präsentation der Ergebnisse als KWIC-Konkordanz gibt es dann die Möglichkeit, einzelne Treffer im Kontext des Transkriptausschnittes anzuzeigen und das zugrunde liegende Audio oder Video abzuspielen (Abb. 9.3).

The screenshot shows a KWIC-Konkordanz search interface. The main table lists search results with columns for 'Ergebnis', 'Sprecher', 'Treffer', and 'Geschlecht'. A detailed view of hit #9 is shown in a pop-up window, displaying the original text and the KWIC extraction.

Ergebnis	Sprecher	Treffer	Geschlecht
1	FOLK_00001 LB	nee aber sie ham s verstanden denk iach	Männlich
2	FOLK_00007 JK	nee aber man kann s ja kontrollieren ja un bevor ich	Männlich
3	FOLK_00011 VK	nein aber weil ihr hier die ganze zeit so rum	Männlich
4	FOLK_00012 VK	nee aber hier fängt ma an eins zwei drei zum beispiel	Männlich
5	FOLK_00021 SK	nee aber dass die den direkt nach	Männlich
6	FOLK_00021 CH	nee aber beim ha es vau bald	Männlich
7	FOLK_00021 MT	nee aber	Männlich
8	FOLK_00021 SK	nö aber des könnt ja sein	Männlich
9	FOLK_00021 CH	nö aber ich könnt ja dann	Männlich

Ergebnis	Sprecher	Treffer	Geschlecht
10	FOLK_00030 PB	nee aber mit internetsellen meils manchmal	Männlich
11	FOLK_00039 NO	nee aber is schon	Männlich
12	FOLK_00039 NO	nee aber dass dass schon	Männlich
13	FOLK_00039 NO	nee aber dann suchen wa uns en foto en foto suchen	Männlich
14	FOLK_00042 LK	nee aber die	Männlich
15	FOLK_00042 LK	nein aber ich nein nein nein nein so hab ich	Männlich
16	FOLK_00043 PB	nee aber ich hab den dir hier mitgebracht	Männlich
17	FOLK_00043 PB	nee aber ich mag diese russpampe net so	Männlich
18	FOLK_00047 PB	nee aber das ja ich will das ja auch f bisschen	Männlich
19	FOLK_00047 PB	nein aber ku ma die sind alle reserviert	Männlich
20	FOLK_00066 PA	nee aber ähm	Männlich

Abb. 9.3: KWIC-Konkordanz als Ergebnis zur Recherche aus Abbildung 9.2 mit eingeblenndetem Transkriptausschnitt zu Treffer #9. Durch Doppelklick auf ein beliebiges Wort im Transkript oder Klick auf den „Play“-Button einer beliebigen KWIC-Zeile wird das zugehörige Audio abgespielt.

Zu den Herausforderungen für die Arbeit mit Gesprächskorpora gehört auch, solche und weitere Verfahren, die speziell den Interaktions-Aspekt der Sprache korpuslinguistisch zugreifbar machen, weiter zu entwickeln und zu verfeinern.

6 Schlussbemerkung

Gesprächskorpora sind in diesem Beitrag als ein besonderer Korпустyp definiert und beschrieben worden, dessen Potenzial für die Korpuslinguistik sich erst in jüngerer Zeit zu erschließen begonnen hat und für den sich methodische und technische Herausforderungen stellen, die über den korpuslinguistischen „Mainstream“ hinausweisen. Wenn der „eklatante[n] Unterrepräsentation spontansprachlicher Daten in Korpora“ (Mair in diesem Band) entgegengewirkt werden soll, können und sollten Gesprächskorpora dabei eine wichtige Rolle spielen. Die Korpuslinguistik (als linguistische Teildisziplin oder teildisziplinen-übergreifende methodische Ausrichtung) wäre dann gefordert, die Eigenheiten dieses „besonderen“ Korпустyps in ihre Methodik einzubeziehen und ihre technologischen Lösungen so zu gestalten, dass Text-, Gesprächs- und ggf. weitere Korпустypen auf einer gemeinsamen Basis verarbeitet und analysiert werden können. Umgekehrt müssten angesichts

des großen Aufwandes, den die Erstellung von Gesprächskorpora mit sich bringt, Fragen der technischen Aufbereitung, der Standardisierung und des Teilens und Nachnutzens von Gesprächsdaten im Rahmen der gesprächsanalytischen Disziplinen verstärkte Aufmerksamkeit finden.

Literatur

- Batinic, Dolores & Thomas Schmidt (2017): Reconstruction of separable particle verbs in a corpus of spoken German. Erscheint in: Proceedings der GSCL-Tagung 2017, Berlin.
- Berens, Franz-Josef, Karl-Heinz Jäger, Gerd Schank & Johannes Schwitalla (1976): *Projekt Dialogstrukturen. Ein Arbeitsbericht* (Heutiges Deutsch 1/12). München: Hueber.
- Biber, Douglas (1993): Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4), 243–257.
- Brinckmann, Caren (2009): Transcription bottleneck of speech corpus exploitation. In Verena Lyding (Hrsg.), *LULCL II 2008 – Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics. Bozen-Bolzano, 13th–14th November 2008*, 165–179. Bozen-Bolzano: EURAC.
- Brünner, Gisela (1987/2005): *Kommunikation in institutionellen Lehr-Lern-Prozessen. Diskursanalytische Untersuchungen zu Instruktionen in der betrieblichen Ausbildung*. Tübingen: Narr; Neuauflage: Radolfzell: Verlag für Gesprächsforschung, 2005. <http://www.verlag-gespraechsforschung.de/2005/bruenner.htm> (letzter Zugriff: 26. 9. 2017).
- Bührig, Kristin, Ortrun Kliche, Bernd Meyer & Birte Pawlack (2012): The corpus “Interpreting in Hospitals”: Possible approaches for research and communication training. In Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14), 305–314. Amsterdam: Benjamins.
- Deppermann, Arnulf & Martin Hartung (2012): 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 Ekkehard Felder, Marcus Müller & Friedemann Vogel (Hrsg.), *Korpuspragmatik*, 414–450. Berlin, Boston: de Gruyter.
- Deppermann, Arnulf & Thomas Schmidt (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik – Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). *Mitteilungen des Deutschen Germanistenverbandes* 61 (1), 4–17.
- Deppermann, Arnulf, Nadine Proske & Arne Zeschel (Hrsg.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr.
- Dister, Anne, Françoise Gadet, Ralph Ludwig, Chantal Lyche, Lorenza Mondada, Stefan Pfänder, Anne Catherine Simon & Ingse Skattum (2008): Deux nouveaux corpus internationaux du français: CIEL-F (Corpus International et Écologique de la Langue Française) et CFA (Français contemporain en Afrique et dans l’Océan Indien). *Revue de Linguistique Romane* 285/286, 295–314.
- Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, & Nii Martey (2000–2005). *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia, PA: Linguistic Data Consortium.

- Ehlich, Konrad & Jochen Rehbein (1986): *Muster und Institution: Untersuchungen zur schulischen Kommunikation*. Tübingen: Narr.
- Eshkol-Taravella Iris, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua & Isabelle Tellier (2012): Un grand corpus oral «disponible»: le corpus d'Orléans 1968–2012. *Ressources Linguistiques Libres, TAL* 52 (3), 17–46.
- Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (2012): The GeWiss corpus: Comparing spoken academic German, English and Polish. In Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14), 319–338. Amsterdam: Benjamins.
- Fandrych, Christian, Elena Frick, Hanna Hedeland, Anna Iliash, Daniel Jettka, Cordula Meißner, Thomas Schmidt, Franziska Wallner, Kathrin Weigert & Swantje Westpfahl (2016): User, who art thou? User profiling for oral corpus platforms. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, 280–287. Paris: European Language Resources Association (ELRA).
- Glas, Reinhold & Konrad Ehlich (2000): Deutsche Transkripte 1950 bis 1995. Ein Repertorium (Arbeiten zur Mehrsprachigkeit, 63). Hamburg: Institut für Germanistik I/Arbeitsbereich Deutsch als Fremdsprache, Arbeitsstelle Mehrsprachigkeit/Research Center for Multilingualism, Universität Hamburg.
- Groupe ICOR (Michel Bert, Sylvie Bruxelles, Carole Etienne, Lorenza Mondada, Véronique Traverso) (2010): Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). *Pratiques – Interactions et corpus oraux* 147–148, 17–34.
- Haugh, Michael & Wei-Lin M. Chang (2013): Collaborative creation of spoken language corpora. In Tim Greer, Donna Tatsuki & Carsten Roeveer (Hrsg.), *Pragmatics and Language Learning, Volume 13*, 133–159. Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.
- Herkenrath, Annette & Jochen Rehbein (2012): Pragmatic corpus analysis, exemplified by Turkish-German bilingual and monolingual data. In: Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14) 123–152. Amsterdam: Benjamins.
- Hohenstein, Christiane (2006): *Erklärendes Handeln im Wissenschaftlichen Vortrag. Ein Vergleich des Deutschen mit dem Japanischen* (Studien Deutsch 36). München: iudicium.
- Kaiser, Julia (2016): Reformulierungsindikatoren im gesprochenen Deutsch: Die Benutzung der Ressourcen DGD und FOLK für gesprächsanalytische Zwecke. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 17, 196–230.
- Kallmeyer, Werner (Hrsg.) (1994): *Exemplarische Analysen des Sprachverhaltens in Mannheim. Kommunikation in der Stadt* (Schriften des Instituts für deutsche Sprache 4.1). Berlin, New York: de Gruyter.
- Keim, Inken (2008): *Die „türkischen Powergirls“ – Lebenswelt und kommunikativer Stil einer Migrantinnengruppe in Mannheim* (Studien zur deutschen Sprache 39). 2. korrig. Aufl. Tübingen: Narr.
- Koch, Peter & Wulf Oesterreicher (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15–43.
- Kupietz, Marc & Thomas Schmidt (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In Ludwig M. Eichinger (Hrsg.):

- Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven* (Jahrbuch des Instituts für Deutsche Sprache 2014), 297–322. Berlin, Boston: de Gruyter.
- Labov, William (1972): *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania.
- Lalouschek, Johanna & Florian Menz (2002): Empirische Datenerhebung und Authentizität von Gesprächen – am Beispiel medizinischer Kommunikation. In Gisela Brünner, Reinhard Fiehler & Walther Kindt (Hrsg.), *Angewandte Diskursforschung*, Band 1: *Grundlagen und Beispielanalysen*, 46–68. Radolfzell: Verlag für Gesprächsforschung.
- Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery (2017): The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22 (3), 319–344.
- Luckmann, Thomas (1986): Grundformen der gesellschaftlichen Vermittlung des Wissens: Kommunikative Gattungen. *Zeitschrift für Soziologie* 27, 191–211.
- Menke, Peter, Farina Freigang, Thomas Kronenberg, Sören Klett & Kirsten Bergmann (2015): First steps towards a tool chain for automatic processing of multimodal corpora. *Journal of Multimodal Communication Studies* 2, 30–43. http://jmcs.home.amu.edu.pl/wp-content/uploads/2015/09/Menke_et_al_2014_JMCS.pdf.
- Meliss, Meike & Christine Möhrs (2017): Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. *Sprachreport* 4/2017, 42–53. <http://nbn-resolving.de/urn:nbn:de:bsz:mh39-68549>
- Moore, Robert J. (2015): Automated transcription and conversation analysis. *Research on Language and Social Interaction* 48 (3), 253–270.
- Nothdurft, Werner, Ulrich Reitemeier & Peter Schröder (1994): *Beratungsgespräche. Analyse asymmetrischer Dialoge* (Forschungsberichte des Instituts für deutsche Sprache 61) Tübingen: Narr.
- Ochs, Elinor (1979): Transcription as theory. In Elinor Ochs & Bambi Schieffelin (Hrsg.) (1979), *Developmental pragmatics*, 43–72. New York u. a.: Academic Press.
- Oostdijk, Nelleke (2002): The design of the Spoken Dutch Corpus. In Pam Peters, Peter Collins & Adam Smith (Hrsg.), *New frontiers of corpus research*, 105–112. Amsterdam: Rodopi.
- Rehbein, Jochen & Petra Löning (1993): *Arzt-Patienten-Kommunikation: Analysen zu interdisziplinären Problemen des medizinischen Diskurses*. Amsterdam: Benjamins.
- Rehbein, Jochen, Jutta Fienemann, Sören Ohlhus & Christine Oldörp (2001): Nonverbale Kommunikation im Videotranskript. Zu nonverbalen Aspekten höflichen Handelns in interkulturellen Konstellationen und ihre Darstellung in computergestützten Videotranskriptionen. In Dieter Möhn, Dieter Roß & Marita Tjarks-Sobhani (Hrsg.), *Mediensprache und Medienlinguistik. Festschrift für Jörg Hennig*, 167–198. Frankfurt am Main: Peter Lang.
- Schmidt, Thomas (2005a): *Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit text-technologischen Mitteln*. Frankfurt am Main: Peter Lang.
- Schmidt, Thomas (2005b): Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 6, 103–126.
- Schmidt, Thomas (2014a): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 15, 196–233.
- Schmidt, Thomas (2014b): The database for spoken German – DGD2. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Ninth International Conference on Language Resources*

- and Evaluation (LREC'14)*, Reykjavik, Iceland, 1451–1457. Reykiavik: European Language Resources Association (ELRA).
- Schmidt, Thomas (2016a): Good practices in the compilation of FOLK (Research and Teaching Corpus of Spoken German). *International Journal of Corpus Linguistics* 21 (3), 396–418.
- Schmidt, Thomas (2016b): Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics* 31 (1), 127–154.
- Schmidt, Thomas & Kai Wörner (Hrsg.) (2012): *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies on Multilingualism 14). Amsterdam: Benjamins.
- Schröder, Peter (1975): Die Untersuchung gesprochener Sprache im Projekt ‚Grundstrukturen der deutschen Sprache‘ – Planungen, Probleme, Durchführung. In Ulrich Engel & Irmtraud Vogel (Hrsg.), *Gesprochene Sprache: Bericht der Forschungsstelle Freiburg* (Forschungsberichte des Instituts für Deutsche Sprache 7), 5–46. 2. Aufl. Tübingen: Narr.
- Schröder, Peter (Hrsg.) (1997): *Schlichtung*, Band 3: *Schlichtungsgespräche. Ein Textband mit einer exemplarischen Analyse* (Schriften des Instituts für Deutsche Sprache 5.3). Berlin, New York: de Gruyter.
- Schütte, Wilfried (1996): Boulevardisierung von Information: Streitgespräche und Streitkultur im Fernsehen. In Bernd Ulrich Biere & Rudolf Hoberg, (Hrsg.), *Mündlichkeit und Schriftlichkeit im Fernsehen* (Studien zur deutschen Sprache 5), 101–134. Tübingen: Narr.
- Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek & Marko Stabej (2013): Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation* 47 (4), 1031–1048, doi: 10.1007/s10579-013-9216-5.
- Wagener, Peter & Karl-Heinz Bausch (Hrsg.) (1997): *Tonaufnahmen des gesprochenen Deutsch. Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven*. Berlin, New York: de Gruyter.
- Westpfahl, Swantje & Thomas Schmidt (2016): FOLK-Gold – A GOLD standard for part-of-speech-tagging of spoken German. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 1493–1499. Paris: European Language Resources Association (ELRA).
- Wiese, Heike, Ulrike Freywald, Sören Schalowski & Katharina Mayr (2012): Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache* 2, 97–123.