

## **Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik – Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2)**

Arnulf Deppermann / Thomas Schmidt

### **1. Was ist ein Gesprächskorpus?**

Unter einem Gesprächskorpus verstehen wir eine Sammlung von Aufzeichnungen (Audio- und/oder Videoaufnahmen) authentischer Gespräche (i. e. konzeptionell und medial mündlicher, i. d. R. spontaner, Interaktion von zwei oder mehr Teilnehmern), die nach einer wissenschaftlich begründeten und explizit dargelegten Systematik zusammengestellt und über eine Transkription, ggf. zusätzliche Annotationen und die Dokumentation von Metadaten (zu Gesprächsumständen und beteiligten Sprechern) für eine (sprach-)wissenschaftliche Analyse erschlossen wird. Gesprächskorpora sind zum einen abzugrenzen von schriftsprachlichen Korpora, bei denen die Analyse unmittelbar auf den Primärdaten (i. e. den schriftsprachlichen Texten) aufsetzt, ohne auf die Erschließung mittels theoriegeleiteter Abstraktionen wie Transkriptionen angewiesen zu sein. Zum anderen sind sie zu unterscheiden von anderen Typen mündlicher Korpora, die Gespräche beinhalten, welche eigens für Forschungszwecke geführt wurden, d. h. Interviewkorpora, die der Bearbeitung soziolinguistischer (z. B. Labov 1972) oder dialektologischer (z. B. Zwirner-Korpus, Bethge/Bonnin 1969) Fragestellungen dienen, oder sog. »Speech Corpora« zur Konstruktion sprachtechnologischer Anwendungen (z. B. Hinrichs et al. 2000). Neben FOLK existieren für das Deutsche weitere Gesprächskorpora – beispielsweise das »Freiburger Korpus« (Steger/Engel/Moser 1971), das Korpus »Dialogstrukturen« (Berens et al 1976, beide ebenfalls über die DGD2 zugänglich), das Korpus »Gesprochene Wissenschaftssprache Kontrastiv« (GeWiss, Fandrych et al. 2012) oder das »Kiezdeutsch«-Korpus (KidKo, Wiese 2012).

Gesprächskorpora können die Analysegrundlage für sehr unterschiedliche Fragestellungen aus verschiedenen Teilgebieten der Linguistik bilden – beispielsweise ermöglichen sie einem Lexikographen eine korpusbasierte Untersuchung des mündlichen Gebrauchs eines Lexems, einem Phonetiker (eine ausreichende Aufnahmequalität vorausgesetzt) lautliche Analysen von Spontansprache oder einem Computerlinguisten die Studie von Dialogstrukturen als Grundlage von Systemen zur automatischen Sprachverarbeitung. Das zentrale Anwendungsfeld für Gesprächskorpora sind aber sicherlich diejenigen Teilgebiete der Linguistik, die die Gesprächsstrukturen und -praktiken und die Verwendung sprachlicher Formen in der sozialen Interaktion zum zentralen For-

### Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

schungsgegenstand haben, insbesondere also verschiedene Ausprägungen der Gesprächsforschung wie die Konversationsanalyse, die interaktionale Linguistik, die funktionale Pragmatik oder die Gesprochene-Sprache-Forschung.

## 2. Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

Ort	Region: Rheinfränkische Sprachregion
Institution / Räumlichkeiten	Nicht vorhanden / Wohnzimmer
Aufnahmebedingungen	Pokerchips machen Geräusche auf dem Tisch
3 dokumentierte Sprecher	FOLK_S_00054 ▶ (Mitspieler in FOLK_E_00040_SE_01) FOLK_S_00055 ▶ (Mitspieler in FOLK_E_00040_SE_01) FOLK_S_00056 ▶ (Mitspieler in FOLK_E_00040_SE_01)

Abbildung 1: Ereignismetadaten in FOLK (Ausschnitt)

Das »Forschungs- und Lehrkorpus Gesprochenes Deutsch« (FOLK) wird seit 2008 im Archiv für Gesprochenes Deutsch am IDS Mannheim aufgebaut. Es ist ein Gesprächskorpus mit den folgenden Eigenschaften:

- es beinhaltet ganzer Gespräche;
- Metadateninformationen über Gesprächsereignisse und Sprecher werden nach einem einheitlichen Schema dokumentiert (siehe Abbildung 1);
- die Gesprächsaufnahmen werden vollständig in literarischer Umschrift nach den Vorgaben des GAT2-Systems (Selting et al. 2009) für Minimaltranskripte mit dem Editor FOLKER (Schmidt/Schütte 2010) transkribiert, Aufnahmen und Transkripte werden dabei in Segmenten von etwa 3 bis 5 Sekunden Länge miteinander aligniert;
- zusätzlich zur literarischen Umschrift erfolgen eine standardorthographische Normalisierung der transkribierten Formen, sowie eine Lemmatisierung und ein Part-Of-Speech-Tagging auf zusätzlichen Annotationsebenen (siehe Abbildung 2 und Westpfahl/Schmidt 2013).

Transkription	da	gehst	de	jetz	einfach	über	dem	bild
Normalisierung	da	gehst	du	jetzt	einfach	über	dem	Bil
Lemmatisierung	da	gehen	du	jetzt	einfach	über	d	Bil
POS	ADV	VFIN	PPER	ADV	ADJD	APPR	ART	NN

Abbildung 2: Annotationsebenen in FOLK

FOLK ist als dauerhaft wachsendes Korpus konzipiert, d. h. es soll in den kommenden Jahren kontinuierlich um neue Aufnahmen ergänzt werden. Die Strategie zur Auswahl von Gesprächen in der ersten Ausbauphase beschreiben Deppermann/Hartung (2011, 445 f.) wie folgt:

»Priorität für die Korpuserstellung ist es dabei zunächst, die grundlegenden Parameter der ›Merkmale der Sprechsituation‹ [...] möglichst breit zu variieren. Das erste Ziel besteht darin, eine möglichst breite Spannweite unterschiedlicher kommunikativer Anlässe und Formen (Gattungen) im Korpus abzudecken.«

Diese erste Ausbauphase ist mit der Veröffentlichung von ca. 100 h Aufnahmen bzw. ca. 1 Millionen transkribierter Wort-Tokens Anfang 2014 abgeschlossen. Die bereits verfügbaren Daten umfassen eine breite Spannweite von unterschiedlichen Typen von Alltagskommunikation (z. B. Tischgespräche, Spielinteraktionen zwischen Erwachsenen, verschiedene Formen der Eltern/Kind-Interaktion) und institutioneller Kommunikation (z. B. Unterrichtskommunikation, Prüfungsgespräche, verschiedene Formen beruflicher Kommunikation), dazu kommen Aufnahmen öffentlich-medialer Kommunikation (Schlichtungsgespräche zu Stuttgart 21). Weiteres Ziel für die zukünftige Korpuserweiterung ist eine stärkere regionale und altersbezogene Variationsbreite des Korpus, in dem zur Zeit Aufnahmen von Sprechern mit höherem Bildungsabschluss im Alter zwischen 20 und 30 aus der rheinfränkischen Sprachregion dominieren.

Ziel kann dabei allerdings nicht sein, ein vollständig ausgewogenes Korpus zu erstellen. Dafür ist die Zahl der interessierenden Variablen, bezüglich derer eine Ausgewogenheit herzustellen wäre, zu groß. Eine ausgewogene Stratifizierung, beispielsweise nach Ort der Erhebung, Alter, Geschlecht und Bildungsstand der Sprecher, die zusätzlich auch noch eine grobe Gesprächstypenklassifizierung mit einbezieht (etwa Alltags- vs. institutionelle Kommunikation), würde bedeuten, dass für jede Kombination von Variablenausprägungen (bspw. ›Alltagsgespräche aus dem bairischen Sprachraum mit älteren männlichen Sprechern mit niedrigerem Bildungsabschluss‹) ausreichend und gleich große Datenmengen in das Korpus einfließen müssten. Dies wäre angesichts des Aufwandes, der mit der Datenerhebung und -aufbereitung verbunden ist, eine utopische Anforderung. Statt also die Variablen in Kombination zu betrachten, soll daher künftig versucht

### Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

werden, zumindest zu jeder einzelnen Variablen-Ausprägung ausreichend (nicht aber unbedingt gleiche Mengen an) Daten im Korpus zu haben – also etwa bei den Erhebungsorten keine sprachliche Großregion auszulassen, und Sprecher aller Altersklassen und Bildungstufen zu berücksichtigen. Damit wird zwar keine Ausgewogenheit des Korpus als Ganzes erreicht, es ist aber immerhin möglich, aus dem Gesamtbestand Teilkorpora zu bilden, die bezüglich einer ausgewählten Variablen ausgewogen sind. Die technische Voraussetzung hierfür ist eine Funktionalität, die es dem Benutzer erlaubt, durch systematische Suchanfragen auf Metadaten und/oder manuelle Auswahl einzelner Gesprächsereignisse virtuelle Korpora zu bilden und die weitere Analyse dann auf ein solches virtuelles Korpus zu beschränken. Die DGD2 stellt eine solche Funktionalität bereits zur Verfügung (siehe Abbildung 3).

		Ergebnis	Ort (Region)	Sprecher
<input checked="" type="checkbox"/>	1	FOLK_E_00189	Nordniederdeutsche Sprachregion	FOLK_S_00431 Männlich
<input checked="" type="checkbox"/>	2	FOLK_E_00100	Nordniederdeutsche Sprachregion	FOLK_S_00301 Männlich
<input checked="" type="checkbox"/>	3	FOLK_E_00107	Nordniederdeutsche Sprachregion	FOLK_S_00276 Männlich
<input checked="" type="checkbox"/>	4	FOLK_E_00178	Bairische Sprachregion	FOLK_S_00420 Männlich FOLK_S_00433 Männlich
<input checked="" type="checkbox"/>	5	FOLK_E_00186	Bairische Sprachregion	FOLK_S_00426 Männlich FOLK_S_00433 Männlich
<input checked="" type="checkbox"/>	6	FOLK_E_00184	Nordniederdeutsche Sprachregion	FOLK_S_00424 Männlich FOLK_S_00432 Männlich

Abbildung 3: Zusammenstellen eines virtuellen Korpus (Männliche Sprecher aus drei ausgewählten Sprachregionen)

### 3. Gesprächsdatenbanken: Die Datenbank für Gesprochenes Deutsch (DGD2)

Außer durch die Daten selbst werden die Nutzungsmöglichkeiten eines Gesprächskorpus ganz wesentlich mitbestimmt durch die Instrumente, die für deren Analyse zur Verfügung stehen. Erst durch eine geeignete informationstechnische Erschließung der Korpusdaten können diese ihr volles Potential entfalten. Das zentrale informationstechnologische Instrument für die Nutzung von FOLK ist die Datenbank für Gesprochenes Deutsch (DGD2, <http://dgd.ids-mannheim.de> Schmidt et al. 2013). Diese stellt FOLK (und derzeit 18 weitere mündliche Korpora, darunter große Korpora zu binnen- und auslandsdeutschen Varietäten und weitere Gesprächskorpora, s. o.) in Form einer Website – und somit ohne größere praktische Hürden wie die Notwendigkeit, eine spezialisierte Software zu installieren oder

umfangreiche Datenbestände auf den lokalen Rechner herunterzuladen – der wissenschaftlichen Öffentlichkeit kostenlos zur Verfügung.

Die DGD2 ermöglicht sowohl ein exploratives Browsen (Durchblättern) als auch ein gezieltes Durchsuchen der FOLK-Daten. In beiden Funktionalitäten wird sichergestellt, dass die Beziehungen zwischen den verschiedenen Datentypen für den Anwender unmittelbar verfü- und nutzbar sind. So können beispielsweise aus der Darstellung eines Transkripts heraus auch Metadaten zu den beteiligten Sprechern abgerufen und zu einer gegebenen Transkriptstelle die zugehörige Stelle in der Audio-Aufnahme abgespielt werden. Ebenso bietet die Suchfunktionalität Möglichkeiten, ausgehend von jeder Fundstelle, die in der Übersicht zunächst als Keyword-In-Context (KWIC) angezeigt wird, zugehörige Sprecher- oder Ereignismetadaten, die zugehörige Transkriptstelle sowie die korrespondierende Stelle der Aufnahme abzurufen.

Auch wenn das Ziel ist, mit diesen (in Zukunft noch auszubauenden) Funktionalitäten möglichst viele Szenarien qualitativer und quantitativer Analyse zu bedienen, so erheben FOLK und die DGD2 nicht den Anspruch, die Analyseerfordernisse, die an Gesprächskorpora gestellt werden können, in ihrer ganzen Breite und vollständig abzudecken. So mögen insbesondere viele gesprächsanalytische Fragestellungen eine vertiefte ethnografische Untersuchung, verfeinerte Transkriptionen oder Annotationen notwendig machen. Die DGD2 ist daher nicht als geschlossenes System konzipiert, sondern erlaubt dem Nutzer auch, Transkriptions- und Audioausschnitte zu Einzelbelegen aus der webbasierten Umgebung heraus auf den lokalen Rechner herunterzuladen und dort mit weiteren Analyseinstrumenten (insbesondere mit einem Transkriptionseditor wie FOLKER<sup>1</sup> oder einem Tool zur phonetischen Analyse wie Praat<sup>2</sup>) weiter zu bearbeiten.

#### **4. Verwendung von Gesprächsdatenbanken in der interaktionalen Linguistik**

Gesprächsdatenbanken lassen sich besonders im Rahmen des Paradigmas der interaktionalen Linguistik vielversprechend verwenden. Wie die Funktionalitäten der Datenbank genutzt werden können, hängt vom Typ der Fragestellung ab. Formbestimmte Fragestellungen, bei denen die Verwendung einer sprachlichen Form den Untersuchungsfokus bildet, sind ideale Kandidaten. Dies gilt für solche Formeigenschaften, die im Transkript oder durch Kodierungen und Annotate erfasst sind. Im Korpus FOLK sind dies Lautvarianten von Ausdrücken, soweit sie in der literarischen Umschrift des Transkriptionssystems GAT mithilfe des Standardalphabets repräsentiert werden können, die orthografisch normalisierten Wortformen, auf die die transkribierten Lautvarianten zu beziehen sind, und die

---

1 siehe <http://agd.ids-mannheim.de/folker.shtml>

2 siehe <http://www.fon.hum.uva.nl/praat/>

### Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

---

Lemmata, auf die als lexikalische Nennform unterschiedliche morphologische Varianten zu beziehen sind. Alle Vorkommen von Lautvarianten, lexikalischen Formen und grammatischen Ausdrücken, die über eine dieser drei Möglichkeiten als Zeichenkette definiert werden können, können in FOLK per Suchmaschine gefunden werden. Dabei ermöglicht das Part-of-Speech-Tagging, die Suche in Bezug auf Wortarten weiter einzugrenzen.

Die maschinelle Suche ist der Ausgangspunkt für die Erstellung einer Kollektion von Fällen. Diese wird dann anhand der Sichtung und der detaillierten Form- und Sequenzanalyse der einzelnen Fälle verfeinert (s. Abschn. 5; vgl. Deppermann 2008, Kap. 6.3). Die Möglichkeit, aufgrund von ausgewählten Metadatenparametern virtuelle Korpora zu definieren, erlaubt es außerdem, die Suche auf bestimmte Kontextbedingungen wie Sprechergruppen, Interaktionstypen oder Aufnahme-regionen zu beschränken. Es ist geplant, auch formale gesprächsstrukturelle Merkmale als Suchbedingung einbeziehen zu können, so etwa das Vorkommen einer Form vor, innerhalb oder nach einer Überlappung mehrerer Sprecher oder die Position des Phänomens am Turnanfang, in der Mitte oder am Ende des Turns (wie z. B. in der französischen Datenbank CLAPI; vgl. Groupe ICOR 2008a, b).

Weniger geeignet für maschinelle Suchen sind formal-abstrakte und interpretative Phänomene in Gesprächen. Gemeint sind damit abstrakte, nicht-formdefinierte grammatische und semantische Klassen (wie z. B. Transitivkonstruktion, Bewegungsverb), Diskurstopiks (Sprechen über Politik) und pragmatisch-interaktionale Phänomene, die eine Interpretation des Interaktionsgeschehens erfordern (z. B. sprachliche Handlungen wie Fragen oder Beleidigungen; längere Diskursabschnitte, die kommunikative Gattungen wie Klatsch oder Belehrungen ausmachen; inferenzielle Bedeutungen wie Anspielungen oder Ironie; Phänomene der Beziehungs- und Identitätskonstitution wie ›Selbst-/Fremdpositionierung als glaubwürdig‹ oder ›Inszenierung von epistemischer Überlegenheit‹). Diese Untersuchungsgegenstände können nicht direkt als solche zum Gegenstand einer Suchanfrage gemacht werden. Dies erforderte, dass die entsprechenden Phänomene in den Daten annotiert werden. Solche Annotationen werden in FOLK nicht vorgenommen, da sie enorm aufwändig (und deshalb den Korpusausbau verlangsamend), immer nur für spezifische Forschungszwecke in einer speziellen Granularität (z. B. ›Erzählung‹ vs. ›autobiographische Erzählung‹ vs. ›Opfererzählung‹) brauchbar und manchmal auch theorieabhängig sind (z. B. ›Narration‹ als genereller Terminus vs. ›Berichten‹ vs. ›Erzählen‹). Ein datenbankbasiertes Korpus wie FOLK kann dennoch zur Untersuchung solcher Phänomene benutzt werden. Zum einen bietet es die Möglichkeit, die einzelnen Gesprächstranskripte und -aufnahmen daraufhin durchzusehen, wo Fälle der interessierenden Gesprächsphänomene vorkommen, und diese dann »von Hand« zu sampeln und auszuwerten. Da sie oft unter besonderen Beteiligungskonstellationen interessieren, können mit Hilfe der Metadaten die potenziell interessierenden Gespräche als virtuelles Korpus eingegrenzt werden. Zum anderen gibt es manchmal typische formale Indikatoren für bzw. Realisierungen von interpretativ konstituierten Gesprächsphänomenen (z. B.

pejorative Personenbezeichnungen in Beschimpfungen, die Diskursmarker *also* oder *das heißt* als Reformulierungsindikatoren), die man maschinell suchen kann. Solche Phänomene können den Ausgangspunkt für Beobachtungen und Hypothesen bilden, die dann anhand von einzelfallbasierten, sequenzanalytisch-interpretativen Untersuchungen weiterer Vorkommen, die aufgrund der Durchsicht und Analyse der Transkripte und Audiodaten gesucht werden müssen, geprüft und vertieft werden. Das Gleiche gilt auch für phonetische und prosodische Phänomene. Sie werden in FOLK nicht transkribiert bzw. annotiert, können aber durch Abhören der Audiodaten erschlossen werden.

## 5. Eine exemplarische Untersuchung: der Diskursmarker *das heißt*

Im Folgenden zeigen wir, wie das Korpus FOLK für eine interaktionslinguistische Untersuchung genutzt werden kann. Als Beispiel wählen wir den Diskursmarker *das heißt* (in Bezug auf das Französische s. Kotschi 1990). Diskursmarker sind eine häufige Erscheinung in der verbalen Interaktion. Sie sind charakteristisch für die Grammatik der gesprochenen Sprache (Schiffrin 1987; Heritage 1984; Auer/Günthner 2005; Fiehler et al. 2004). Mit Diskursmarkern zeigen Sprecher, in welcher pragmatischen Relation ein Turn (bzw. eine Turnkonstruktionseinheit) zu einem anderen, (meist) vorangehenden Turn (bzw. Turnkonstruktionseinheit) steht. Diskursmarker im Deutschen zeichnen sich durch ihre Stellung außerhalb des Satzrahmens (im Vorvorfeld) aus, sie sind weder in Rektionsbeziehungen eingebunden noch üben sie Rektion aus. Diskursmarker sind homonym zu anderen Konnektoren, Adverbien und Adjektiven (z. B. *obwohl*, *klar*) oder deverbale (*ich mein*, *siehste*). Eine primäre Funktion von Diskursmarkern ist es, Verstehensanweisungen für das Folgende, den Turnabschnitt, der in ihrem Skopus liegt, zu geben (Fiehler et al. 2004).

Da sie anhand ihrer morphologischen Form zu suchen sind und häufig vorkommen, sind Diskursmarker ein idealer Gegenstand für eine datenbankgestützte Untersuchung. Wir nehmen in der DGD2 zunächst eine Suche nach der normalisierten Form *heißt* vor. Sie ergibt 630 Treffer. Da für eine interaktionslinguistische Untersuchung die Analyse jedes Einzelbelegs erforderlich ist, grenzen wir die Trefferanzahl auf eine manuell handhabbare Menge ein. Wir wählen eine Zufallsstichprobe von 200 aus den 630 gefundenen Instanzen von *heißt*. Zumeist, und so auch in unserem Fall, sind Suchen überinklusiv, d. h. sie ergeben viele falsche Positive, die gar nicht Fälle der interessierenden Form bzw. Praktik sind. Daher filtern wir aus der Zufallsstichprobe diejenigen Verwendungen von *heißt* aus, die nicht den Konnektor *das heißt* betreffen (siehe Abbildung 4).<sup>3</sup>

3 Dies sind alle Fälle, in denen *heißt* mit einem anderen Subjekt als *das* vorkommt.

## Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

Ergebnisse 1 bis 20 von 200 (2 ausgefiltert)

✓	ID	Code	Text	Suchbegriff
✓	1	FOLK_00001 LB	denk des kennen sie die schätzung des	heißt
✓	2	FOLK_00001 LB		das heißt
✓	3	FOLK_00001 LB		des heißt isch bekäme en zündimpuls
✓	4	FOLK_00001 LB		das heißt was müsse mer mache immer wenn mer prüfen mer
✗	5	FOLK_00093 CJ	zahlung hier kommt benni	heißt es
✓	6	FOLK_00004 GS		des heißt sie brauchen mehrere von dem karteikarte
✗	7	FOLK_00004 GS	genau danke was	heißt koga tiv
✓	8	FOLK_00005 LB		des heißt dass es zu kompliziert werd
✓	9	FOLK_00006 LB	signal des sieht ganz anders aus das	heißt wenn isch des

Abbildung 4: Manuell gefilterte Liste von Suchergebnissen für ›heißt‹

Die uns interessierende Diskursmarkerverwendung von *das heißt* hat die nicht-referenzielle, anaphorische Bedeutung als Reformulierungsindikator: ›mit der vorangegangenen Äußerung (Wort, Phrase, Satz, längerer Turn) will der Sprecher das Folgende zum Ausdruck bringen‹. Dies trifft für 95 der 200 zufällig aus der Gesamtmenge der *heißt*-Treffer ausgewählten Fälle zu. Dabei zeigt sich, dass die Berücksichtigung der lautlichen Formvarianten wichtig war: In 32 der 95 Fälle ist das Pronomen *das* als *des* (süddeutsche regionalsprachliche Variante) realisiert, einmal klitisiert als »s\_heißt«.

Die einfache KWIC-Ansicht reicht zumeist nicht aus, um eine syntaktische Analyse der gefundenen Fälle vornehmen zu lassen. Wir müssen dazu den angezeigten Kontext der Belege auf mindestens fünf, manchmal aber auch mehr Segmente erweitern (siehe Abbildung 5).

201 FOLK\_50223 FM des (104) 20

0332	S2	(0.45)	schon nicht
0333	HM	die misse_mer jetzt auswerte	
0334		(0.6)	
0335	HM	des heißt ah	
0336		(1.56)	
0337	HM	wir brauchen ah_n termin wann wer des (.) wann wer die bänder abböre un des ganze mache	
0338		(2.5)	

Abbildung 5: Anzeigen des erweiterten Kontexts eines Belegs in der KWIC

Meist hat *heißt* dann die Semantik ›einen Namen/eine Bezeichnung tragen‹ (z. B. »wie heißt die«), ›in einer anderen Sprache bedeuten‹ (»weißt du was freunde auf türkisch heißt«), ›die Angemessenheit einer Formulierung in Frage stellen‹ (»also was heißt nachteil«) oder ›Sachverhalt A ist Ursache für Sachverhalt B‹ (»nur weil ich en mann bin (.) heißt des ja noch lange nicht dass der mich anbagert«). Doch auch *das heißt* kann in einer dieser Bedeutungen vorkommen, z. B. in »das heißt (.) die selektive prüfung des steuergerätes (0.6) der begriff wird ihne (0.35) noch öfters ah begegnen«.



Es ergibt sich ein klarer Befund: In 85 von 95 Fällen folgt *das heißt* ein syntaktisch unverbundener Hauptsatz – *das heißt* ist hier, wie in den Segmenten 0976 und 0984 im folgenden Ausschnitt, Diskursmarker.

0972	LB	[dieser] kleine messstrom (.) reischt leider nischt aus
0973		(1.09)
0974	LB	ja
0975		(0.39)
0976	LB	<b>des heißt</b> sie ham hier en problem
0977		(0.68)
0978	LB	sie haben
0979		(0.42)
0980	LB	er hat_s ja gsagt der herr grauberg fünf mal die null sieben
0981		(0.27)
0982	LB	kommt ihr messstrom nisch dursch
0983		(0.71)
0984	LB	<b>des heißt</b> sie können zunächst mal diese sekundärspule
0985		(0.36)
0986	LB	sekundärmäßig nischt prüfen

Transkriptausschnitt 1: FOLK\_E\_00009\_SE\_01\_T\_01/00:20:12.81

In drei Fällen ist der folgende Satz ein Fragesatz, z. B. »**das heißt** was müsse mer mache immer wenn mer prüfen hier (1.06) wenn\_s steuergerät was machen soll« (FOLK\_E\_00001\_SE\_01\_T\_02/00:41:53.55<sup>4</sup>). Wir sehen also, dass *das heißt* nicht in jedem Fall eine folgende Reformulierung ankündigt, sondern auch benutzt werden kann, um anzuzeigen, dass eine Aufforderung zu einer Reformulierung folgt.

Nur in 7 Fällen wird *das heißt* als Matrixsatz mit folgendem abhängigen *dass*-Nebensatz verwendet. Diese Fälle haben wir auf einen Blick, wenn wir die ausgewählten Treffer nach ihrem rechten Kontext, d. h. dem Wort, das dem gesuchten Element folgt, alphabetisch sortieren (siehe Abbildung 6):

In drei Fällen wird die Matrixsatzstruktur, in keinem aber der Diskursmarker negiert, so in »ich hab jetzt hier mal\_n paar mögliche ergebnisse **des heißt** net dass des perfekt is was jetzt da so dran is aber« (FOLK\_E\_00120\_SE\_01\_T\_02/

4 Wir zitieren Einzelbelege aus dem FOLK-Korpus in der Form FOLK\_E\_XXXXX\_SE\_01\_T\_yy/hh:mm:ss.ss. Dabei entspricht der erste Bestandteil der DGD2-Kennung des betreffenden Transkripts (XXXXX ist die Nummer des Gesprächsereignisses, yy die Nummer des zugehörigen Transkripts), der zweite Bestandteil gibt den Zeitpunkt in der zugehörigen Aufnahme an. Diese Form der Zitierung bleibt auch dann gültig, wenn – was von einer Version des Korpus zur nächsten der Fall sein kann – Details des Transkripts korrigiert oder ergänzt werden.

## Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

	Beispiel	Sprecher							
✓ 1	FOLK_00005	LB					des	heißt	dass es zu kompliziert wird
✓ 2	FOLK_00007	HM					des	heißt	dass ma halt net nie geht un äh
✓ 3	FOLK_00024	AW				auch für vier	des	heißt	dass die dann
✓ 4	FOLK_00032	AK			Lesekompetenz dauerhaft verbessern	das	heißt	sie ähm	
✓ 5	FOLK_00033	ST			ne art transferprozess	das	heißt	dass man fähigkeiten erwirbt die man auf andere	

Abbildung 6: Treffer in der KWIC, nach rechtem Kontext sortiert

00:02:20.15). Dies ist ein weiterer Hinweis auf die grammatikalisierte und lexematisch verfestigte Natur der Diskursmarkerverwendung.

Nur in zwei der 95 Fälle von *das heißt* als Reformulierungsindikator folgt auf *das heißt* eine subsentenzielle Struktur (d. h. nur eine NP oder eine AdJP) wie hier: »als ein [...] größtenteils konventionell gewordenes system °hh das heißt sprachbewusstheit (.) und wozu is denn das wichtig« (FOLK\_E\_00033\_SE\_01\_T\_02/00:26:47.07). Wenn wir den Kontext über 5 Segmente hinaus erweitern, sehen wir, dass die Sätze, die auf den Diskursmarker *das heißt* folgen, in 30 Fällen komplexe Sätze mit mindestens einem Nebensatz sind. Wenn wir nach dem Erreichen der maximalen Entfaltung der KWIC-Ansicht in die Volltranskripte und Aufnahmen der Belege wechseln, stellen wir außerdem fest, dass der Skopus von *das heißt* sich oftmals über weit mehr als den folgenden Satz erstreckt und komplexe Beschreibungen, Argumentationen und Abfolgen von Selbstreformulierungen (mehrere *das heißt*-Konstruktionen nacheinander) umfasst (vgl. Transkriptausschnitt 1 oben). Unsere Untersuchung legt einen interessanten Befund nahe: Es scheint eine Arbeitsteilung zwischen *also* und *das heißt* zu geben. Während nach *also* Reformulierungen folgen, die fast immer kürzer als ein Satz sind und komplexe Aussagengefüge (Erzählungen, Argumentationen) auf einen Begriff bringen (Deppermann 2011), verwenden Sprecher *das heißt* zu expandierenden Reformulierungen, die eher erläutern und komplexe Schlussfolgerungen ziehen als verbegrifflichen (vgl. a. Bührig 1996, 191 f.).

Wenn wir Treffer nach linkem bzw. rechtem Kontext sortieren, sehen wir, dass es erstaunlich wenig ausgeprägte Kookkurrenzpartner von *das heißt* gibt. Unmittelbar vorangehend finden wir 9 mal *also*. Unmittelbar folgend finden wir je 6 mal *also* und *äh(m)*. *Also* als Kookkurrenzpartner von *das heißt* bekräftigt weiter, dass *das heißt* eine inferenzielle Reformulierung ankündigt.

*Das heißt* kann sowohl zur Selbst- wie zur Fremdreformulierung benutzt werden. Im Korpus überwiegen die Selbstreformulierungen (82/95 Fälle), *das heißt* erscheint dann in der Mitte eines Sprecherbeitrags und wird zur Verdeutlichung oder Spezifikation der Redeintention benutzt, wie im Transkriptausschnitt 1.

In 13 der 95 Fälle zeigt *das heißt* eine Fremdreformulierung einer unmittelbar vorangehenden Partneräußerung an und wird als Verstehensdokumentation (vgl. Deppermann/Schmitt 2009) eingesetzt. Es wird dabei am Turnbeginn produziert:

0064	HY	er unternimmt zwar was aber er wartet so wie er (.) auf_m stuhl
0066		(0.25)
0067	HK	ja
0068		(0.59)
0069	HY	und lebt au nicht sein leben richt[ig ]
0070	HK	[das heißt] er bleibt passiv
0071	HY	ja
0072		(0.31)
0073	HK	anstatt (.) aktiv zu werden so wie du auch gesagt hast rene (.) ja

Transkriptausschnitt 2: FOLK\_E\_00120\_SE\_01\_T\_02/00:41:52.77

Um dies festzustellen, können wir uns nicht auf die KWIC-Ansicht verlassen: Die DGD zeigt 37 Fälle an, in denen vor *das heißt* kein linker Kontext erscheint. Dabei handelt es sich aber nicht immer um Turnanfänge. In 29 der Fälle liegen turninterne Pausen vor, die, weil sie separat transkribiert werden, dazu führen, dass sich der Turn auf mehrere Transkriptionssegmente verteilt. Da Transkriptionssegmente somit nicht Turns entsprechen und letztere nicht als solche kodiert werden, kann der Turnbeginn nicht automatisch identifiziert werden. Umgekehrt gibt es auch Fälle, in denen am Turnbeginn Vorlaufelemente vor *das heißt* erscheinen (»äh«, »also«), ohne dass aber die vorangehende Partneräußerung innerhalb des Turns paraphrasiert, pronominalisiert oder wiederholt wird. Entscheidend für die Verwendung als Fremdreformulierung am Turnbeginn ist also nicht der Turnbeginn im Sinne des Beginns der vokalen Turnproduktion, sondern ob eine vorangehende Partneräußerung oder ein vorangehender Teil des eigenen Turns im gleichen Turn vor *das heißt* reformuliert wurde.

Aufgrund dieser Befunde ist es nun möglich, unsere Kollektion weiter zu spezifizieren. Wenn wir uns beispielsweise ausschließlich für die Fälle als Fremdreformulierung interessieren, könnten wir in den aus der anfänglichen Zufallsstichprobe ausgeschlossenen 430 Fälle die weiteren Verwendungen zur Fremdreformulierung suchen und zu den bereits gefundenen 13 Fällen als Kollektion hinzufügen. Wir könnten dann zum Beispiel weiter untersuchen, in welchen Interaktionstypen und welchen Sprechern diese Verwendung vorgenommen wird.

## 6. Fazit und Reflexion

In diesem Beitrag haben wir zu zeigen versucht, wie die im Korpus FOLK verfügbaren Gesprächsdaten für interaktionslinguistische Forschungen genutzt werden können. Das Korpus FOLK und die datentechnische Infrastruktur der DGD unterstützen die verschiedenen methodologischen Anforderungen, die sich einer interaktionslinguistischen Untersuchung stellen: Sie erlauben die extensive und detaillierte sequenzielle Analyse von Einzelfällen durch den Zugriff auf vollständige Transkripte und Aufnahmen, die Bildung von Kollektionen eines sprachlichen

### Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

Phänomens durch systematische, maschinell unterstützte Suchoperationen und die Abschätzung der Bereichsspezifität und Generalisierungsreichweite von Befunden durch ausführliche Metadateninformationen über Sprecher und Gesprächsereignisse. Mithilfe der Bildung virtueller Korpora können bereichsspezifische Verwendungen identifiziert und miteinander kontrastiert werden. Die Möglichkeit, Ausschnitte von Tondaten und Transkripten zu exportieren, erlaubt dem Nutzer, einschlägige Daten weiter gemäß den eigenen Analyseinteressen aufzuarbeiten (z.B. prosodisch zu transkribieren, mit PRAAT auszumessen etc.). Das Korpus FOLK kann als alleinige Basis für eine Untersuchung benutzt oder aber als Kontrastkorpus verwendet werden, um die Spezifität bzw. Generalität der Befunde, die an einem anderen Korpus gewonnen wurden, zu testen.

Mangels hinreichend umfangreicher Korpora, der Tatsache, dass brauchbare Korpora oft nicht wissenschaftsöffentlich verfügbar sind und dass leistungsfähige Erschließungsinstrumente fehlen, wurden Gesprächsdatenbanken in der interaktionalen Linguistik (wie generell zur Untersuchung gesprochener Sprache) nur selten als Basis für Untersuchungen verwendet. Hinzu kommen die für die qualitative Forschung typischen, methodologischen Vorbehalte gegen die Untersuchung von nicht vom Forscher selbst erhobenen Daten, die Brauchbarkeit maschineller Instrumente der Korpuserschließung, welche eine formale und mehr oder weniger kontextfreie Definition von zu suchenden Strukturen erfordern, und die Quantifizierung von Befunden. Dies sind u.E. keineswegs grundsätzliche Einwände gegen die Verwendung von Gesprächsdatenbanken als Datengrundlage und methodisches Hilfsmittel, sondern vielmehr Hinweise auf Probleme, Aufgaben und methodische Vorsichtsmaßnahmen, die im Forschungsprozess zu beachten sind. Wir haben verschiedentlich darauf hingewiesen, dass maschinell gewonnene Suchergebnisse, in bestimmten Ansichten bereitgestellte Repräsentationen der Daten und statistische Kennwerte auf keinen Fall unbesonnen als Grundlage linguistischer Erkenntnisse genommen werden können und dass sie bei naiver Nutzung zu Aussagen führen, die Methodenartefakte sind. Auch sind reine Verteilungsaussagen keine linguistisch relevanten Erkenntnisse, sondern nur Befunde, die interpretiert und erklärt werden müssen. Maschinelle Suchen helfen die Korpuserschließung zu systematisieren. Sie ersetzen aber nicht den Zugang zu den Primärdaten, die detaillierte linguistische Analyse und die Interpretation der Daten; dies umso mehr, je interpretativer das Untersuchungsphänomen ist.

Wenn eine Gesprächsdatenbank entsprechend methodologisch reflektiert eingesetzt wird, bietet sie u.E. unschätzbare, noch kaum genutzte Erkenntnismöglichkeiten, die mit kleineren Korpora und ohne maschinelle Erschließungsinstrumente nicht zu haben sind. Unsere kleine exemplarische Anwendung hat gezeigt, dass eine Gesprächsdatenbank wie FOLK zum einen die induktive Entdeckung von Mustern und Verteilungen (in Statistiken, KWIC-Ansichten, Sortierungen kookurrenter Kontexte) unterstützt. Gerade der größere Umfang und die breitere Stratifikation von FOLK im Vergleich zu den viel kleineren Forschungsprojektkorpora lässt Muster hervortreten, die in kleinen Samples und erst recht in Ein-

zelfallanalysen unbemerkt bleiben, über deren Systematizität und deren Abhängigkeit von spezifischen Kontextbedingungen nicht entschieden werden kann. Genauso unterstützt FOLK aber auch die Varianten- und Bedingungssuche und das Testen von Hypothesen durch die Bereitstellung aussagekräftiger Datenmengen und die Möglichkeit systematischer, maschinell unterstützter Korpuserschließung. Im Forschungsprozess verbinden sich induktive und deduktiv-testende Erkenntnisbildung in einer Bewegung der zunehmenden Verfeinerung von Hypothesen und der korrespondierenden Belegkollektionen. Das Ziel besteht dann in einer Darstellung sprachlicher Praktiken, die sowohl ihrer kontextsensitiven Adaptivität als auch ihren generischen, kontextfreien bzw. -typspezifischen Potenzialen gerecht wird (vgl. Heritage 2010) und dabei das Generalisierungsniveau der wissenschaftlichen Aussagen korpusbasiert empirisch fundiert.

## Literatur

- Auer, Peter/Günthner, Susanne: Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In: Grammatikalisierung im Deutschen. Hg. von Torsten Leuschner und Tanja Mortelsmans. Berlin 2005, S. 335–362.
- Bethge, Wolfgang/Bonnin, Gunther M.: Proben deutscher Mundarten. 123 S. und 1 Karte. Tübingen 1969.
- Berens, Franz-Josef u.a.: Projekt Dialogstrukturen. Ein Arbeitsbericht. Heutiges Deutsch I/12. München 1976.
- Bührig, Kristin: Reformulierende Handlungen. Zur Analyse sprachlicher Adaptierungsprozesse in institutioneller Kommunikation. Tübingen 1996.
- Deppermann, Arnulf: Gespräche analysieren. Wiesbaden 2008.
- Deppermann, Arnulf: Notionalizations: The transformation of descriptions into categorizations. In: *Human Studies* 34, 2, (2011), S. 155–181.
- Deppermann, Arnulf/Hartung, Martin: 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: *Korpuspragmatik*. Hg. von Ekkehard Felder, Marcus Müller und Friedemann Vogel. Berlin, S. 414–450.
- Deppermann, Arnulf/Schmitt, Reinhold: Verstehensdokumentation: Zur Phänomenologie von Verstehen in der Interaktion. In: *Deutsche Sprache* 36, 3, (2009), S. 220–245.
- Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana: The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In: *Multilingual corpora and multilingual corpus analysis*. Hg. von Thomas Schmidt und Kai Wörner. Amsterdam 2012, S. 319–337.
- Fiehler, Reinhard u.a.: Eigenschaften gesprochener Sprache. Theoretische und empirische Untersuchungen zur Spezifik mündlicher Kommunikation. Tübingen 2004.
- Groupe ICOR (M. Bert, S. Bruxelles, C. Etienne, L. Mondada, V. Traverso): Tool-assisted analysis of interactional corpora: »voilà« in the CLAPI database. In: *Journal of French Language Studies* 18 (2008), S. 121–145.

Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik

- Groupe ICOR ( M. Bert, S. Bruxelles, C. Etienne, L. Mondada, S. Teston, V. Traverso):  
 ›oh::, oh là là, oh ben...‹ : les usages du marqueur »oh« en français parlé en interaction.  
 In: Actes du Congrès Mondial de Linguistique Française (CMLF08). Hg. von Durand,  
 J., Habert, B., Laks, B. Paris 2008, 17 pages (CD-ROM). ( DOI: 10.1051/cmlf08099).
- Heritage, John: A change-of-state token and aspects of its sequential placement. In:  
 Structures of Social Action. Hg. von John M. Atkinson und John Heritage. Cambridge  
 1984, S. 299–345.
- Heritage, John: Conversation Analysis: Practices and Methods. In: Qualitative Sociology  
 (3rd Edition). Hg. von David Silverman. London 2010, S. 208–230.
- Hinrichs, Erhard et al.: The Tübingen Treebanks for Spoken German, English, and  
 Japanese. In: Verbmobil: Foundations of Speech-to-Speech-Translation. Hg. von  
 Wolfgang Wahlster. Berlin 2000, S. 552–576.
- Kotschi, Thomas: Reformulierungshandlungen und Textstruktur. Untersuchungen zu  
 frz. *c'est-à-dire*. In: Sprache und Pragmatik 19 (1990), S. 1–27.
- Labov, William: Sociolinguistic Patterns. Philadelphia 1972.
- Wiese, Heike: Kiezdeutsch. Ein neuer Dialekt entsteht. München 2012.
- Schiffrin, Deborah: Discourse Markers. Cambridge 1987.
- Selting, Margret u. a.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Ge-  
 sprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10 (2009), S. 353–402.
- Schmidt, Thomas/Schütte, Wilfried: FOLKER: An Annotation Tool for Efficient  
 Transcription of Natural, Multi-party Interaction. In: Proceedings of the Seventh  
 conference on International Language Resources and Evaluation (LREC'10). Hg.  
 von Nicoletta Calzolari et al. Valletta, Malta 2010.
- Schmidt, Thomas/Dickgießer, Sylvia/Gasch, Joachim: Die Datenbank für Gesprochenes  
 Deutsch – DGD2. Mannheim 2013. URN: urn:nbn:de:bsz:mh39–12747.
- Steger, Hugo/Engel, Ulrich/Moser, Hugo (Hgg.): Texte gesprochener deutscher Stan-  
 dardsprache I. Erarbeitet vom Institut für deutsche Sprache, Forschungsstelle Frei-  
 burg. München 1971.
- Westpfahl, Swantje/Schmidt, Thomas: POS für(s) FOLK– Part of Speech Tagging des  
 Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language  
 Technology and Computational Linguistics, erscheint 2014.

<http://agd.ids-mannheim.de/folker.shtml> [letzter Zugriff 1/14].

<http://dgs.ids-mannheim.de> [letzter Zugriff 1/14].

<http://www.fon.hum.uva.nl/praat/> [letzter Zugriff 1/14].

Prof. Dr. Arnulf Deppermann, Institut für Deutsche Sprache Mannheim,  
 deppermann@ids-mannheim.de

Dr. Thomas Schmidt, Institut für Deutsche Sprache Mannheim, thomas.schmidt@ids-  
 mannheim.de