Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 19 (2018), Seite 515-552 (www.gespraechsforschung-ozs.de)

Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien

Julia Kaiser

Abstract
Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), zugänglich über die Datenbank für Gesprochenes Deutsch (DGD), strebt den Status eines Referenzkorpus für den aktuellen mündlichen Sprachgebrauch im deutschen Sprachraum an. Es enthält einen wachsenden Bestand von Audio- und Videoaufnahmen authentischer Gespräche aus verschiedenen Bereichen des gesellschaftlichen Lebens. Die Dokumentation und Repräsentation von Interaktions- und Sprecherinformationen sind bereits seit den Anfängen des Korpusaufbaus integrale Bestandteile von FOLK. Allerdings lag bislang kein ausgearbeitetes, empirisch erprobtes und vollständig in die Korpusinfrastruktur integrierbares Stratifikationskonzept vor. Mit dem vorliegenden Artikel wird ein solches Konzept vorgeschlagen. Es knüpft an frühere Konzeptionen an und wurde anhand der vorhandenen Daten überprüft, korrigiert und erweitert. Dieser Prozess verlief parallel zur Überarbeitung des XML-Schemas zur Metadatendokumentation, um die konkrete Implementierung vorzubereiten. Im Anschluss an eine Skizzierung genereller Aspekte des Korpusdesigns werden die stratifikationsleitenden und ergänzenden Parameter vorgestellt und erläutert. Abschließend werden Ansätze und Strategien zum Korpusausbau diskutiert.

Keywords: Korpusdesign – Metadaten – Parametersystematik – Gattungssystematik – Konversationsanalyse.

English Abstract
The Research and Teaching Corpus of Spoken German (FOLK), accessible via the Database for Spoken German (DGD), aims for the status of a reference corpus for spoken language in the German-speaking area. It contains a growing inventory of audio and video recordings of authentic conversations from various areas of social life. Since the beginning of the corpus construction, documentation and representation of information about interactions and participants have been integral components. A systematic stratification concept which is explicitly elaborated, tested and capable of being fully integrated is still lacking, though. The present paper will propose such a concept. It builds on previous conceptualizations and has been validated, corrected and expanded on the basis of the existing corpus data. This process took course in parallel with the revision of the XML-schema for metadata documentation in order to prepare for the concrete implementation. After a sketch of general aspects of the corpus design, stratification-leading and additional parameters will be presented and explained. Finally, approaches and strategies for further corpus developments are discussed.

Keywords: Corpus design – metadata – parameter systematics – genre systematics – conversation analysis.

1. Einleitung: übergeordnete Aspekte

Für das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) definieren Deppermann/Hartung (2011:418) das übergeordnete Ziel, den "kommunikativen Haushalt" (nach Luckmann 1986, 1988) der "deutschsprachigen mündlichen Kommunikationspraxis in seinen wesentlichen Ausprägungen" zu repräsentieren. Ausgehend von dem bisherigen Desiderat großer, systematisierter und (wissenschafts-)öffentlich zugänglicher Korpora für die Gesprächsforschung und verwandte Disziplinen strebt das Projekt am Institut für Deutsche Sprache (IDS) Mannheim die Dokumentation (mittels Audio und Video), Transkription und Verfügbarmachung des "vollen Spektrum[s] der privaten, institutionellen, öffentlichen und massenmedialen Anlässe und Typen mündlicher Kommunikation im Deutschen" (ebd.) für Forschungs- und Lehrzwecke in der wissenschaftlichen Gemeinschaft an.

Nach welchen qualitativen und quantitativen Kriterien diese Datensammlung systematisiert und repräsentiert werden soll, muss durch ein detailliertes und explizites Stratifikationskonzept geregelt werden, welches die Grundlage für die Gesprächsbeschreibung, Metadatendokumentation, Suchfunktionalitäten und Ausbauplanung bildet. Bereits zu Projektbeginn lag zur Stratifikation von FOLK ein theoretisch-konzeptioneller Ansatz von Deppermann/Hartung (2011) vor.[1] Eine projektinterne Ausbauplan-Skizze von Schmidt (2017a) ergänzt und konkretisiert

[1] Vgl. auch die Ausführungen in dieser Publikation zu den Abgrenzungskriterien für die grundlegende qualitative Zusammensetzung der FOLK-Daten (421f.).

zentrale, aktuell relevante Aspekte. Auch ein nicht publiziertes Papier von Winterscheid (2016) ist zu nennen und darüber hinaus vor allem mehrere Veröffentlichungen von Schmidt (2014a, b, c, 2017b, c, 2018) und Deppermann/Schmidt (2014) zu konzeptionellen, inhaltlichen, technischen und weiteren Aspekten des Korpus und seiner Einbettung in die Datenbank für Gesprochenes Deutsch. Die folgenden Darstellungen und Überlegungen orientieren sich hauptsächlich an der zuerst genannten Arbeit sowie an Schmidt (2018) und betrachten sie als wesentlichen Ausgangspunkt für weitere Ausarbeitungen von Konzeptionen sowie Auf- und Ausbaustrategien für FOLK. Im Folgenden werden daher zunächst einige der dort angestellten Überlegungen aufgeführt, um im Anschluss auf die jeweiligen Hauptaspekte eingehen zu können.

Deppermann/Hartung (2011:422ff.) erklären, dass für die Entwicklung und Umsetzung eines taxonomischen, stratifikationsleitenden Konzeptes die in der Forschung gewonnenen Erkenntnisse über die Konstitution von verbaler Interaktion und über die Verteilung und den Gebrauch von Sprachvariation als wesentliche theoretische Basis einbezogen werden sollten. Dafür sind a) Fragen zu konkreten inhaltlichen Kriterien und b) zum Ideal der Ausgewogenheit bzw. Repräsentativität zu klären (im Folgenden nacheinander diskutiert).

a) Die Parametersystematik und die Gattungssystematik[2] werden als die beiden zentralen Konzeptionsmöglichkeiten für die Erfassung kommunikativer Variation gegenübergestellt: Beim ersten Ansatz werden forscherseitig Parameter bestimmt, deren Werte bzw. Ausprägungen als grundlegend "für die Konstitution unterschiedlicher Formen von Kommunikationsereignissen" angesehen werden (423). Dabei wird zunächst von einem eher statischen Model der Gesprächssituation ausgegangen, die sich apriorisch durch 'äußere' Merkmale bestimmen lässt. Deppermann/Hartung (2011) führen hierzu drei zentrale Klassen von Parametern auf, welche in einem ausgewogenen und qualitativ repräsentativen Korpus systematisch variiert werden sollten: Merkmale des Sprechereignisses, der Sprecher und der Sprache der Interaktion. Dagegen wird die Gattungssystematik durch "emische Orientierungskategorien" (427) bestimmt, da Gattungen nicht nur durch außenstrukturelle und situative Parameter, sondern auch durch ihre Binnenstruktur, also nicht vorhersagbare oder planbare sprachlich-kommunikative Verfahren, und durch ihre Zwecke oder emergente Themen wesentlich mitkonstituiert werden. Diese sind also nicht als Planungskriterien, sondern erst im Nachhinein erfassbar und bilden häufig kleinere Einheiten innerhalb der Sprechereignisse, sodass eine taxonomische und hierarchische Systematisierung sehr problematisch wird.

Die Autoren plädieren dennoch für eine Kombination beider Ansätze als sinnvoll und umsetzbar: Sofern die konstitutionstheoretisch relevanten Parameter unter Anwendung eines dynamischen, flexiblen Kontextbegriffs identifiziert und kategorial erfasst werden, können einige Merkmalskombinationen auch unterspezifiziert bleiben. Andere, idiosynkratische Aspekte müssten dann für bestimmte spezifische und/oder emergente, gesellschaftlich relevante Gattungen punktuell ergänzt werden. Wichtig ist insgesamt eine klare Definition der oberen hierarchischen Ebenen einer Taxonomie des "kommunikativen Haushalts", sodass auch

[2] Vgl. für den ersten Ansatz das Freiburger Redekonstellationsmodell von Steger et al. (1974) und später Henne/Rehbock (1982/1995) oder das SPEAKING-Modell nach Hymes (1968), für den zweiten Ansatz Luckmann (1986, 1988), Bergman (1987); Günthner (1995, 2000), Günthner/Knoblauch (1994) u.a.

Gattungen – ansonsten häufig über ihren Zweck bzw. ihre Funktion oder auch über thematische, mediale o.a. Kriterien definiert[3] – erhoben werden können, mit denen eine Variation hinsichtlich dieser Oberkategorien repräsentiert wird.

Schmidt (2018) erklärt, dass "leitend für das Korpusdesign […] dabei zunächst der Begriff des Gesprächstyps" sei, es also vor allem um "Unterschiede in Interaktionsanlässen, -konstellationen, -kontexten und -inhalten (i.w.S. 'Situational Parameters' nach Biber 1993:245)" gehe, die angemessen abgebildet werden müssen (218). Als weitgehend einfach operationalisierbar beurteilt er dabei eine erste Unterscheidung in Interaktionsdomänen (218) (wie Privat, Institution, Öffentlichkeit). Als weitere Binnendifferenzierung werden bei institutionellen Gesprächen die Institutionen selbst und eventuell diesen eigene Typisierungen genannt. Im privaten Bereich seien eindeutige Typen-Hierarchien dagegen aufgrund der geringen oder ganz fehlenden äußeren Vorgaben oft nicht möglich (219).

Laut Schmidt (2018) erweist sich die Anwendung einer parametrisierten Systematik auch aufgrund ihrer höheren Flexibilität als praktikabler als eine Gattungssystematik. Er weist allerdings darauf hin, dass eine entsprechende Operationalisierung für reale Gesprächsaufnahmen ebenfalls nicht einfach ist, insofern Grenzfälle geklärt und Definitionen oder Leitlinien für interpretative Entscheidungen bei schwer zu standardisierenden, auswertungsintensiven Kategorien getroffen werden müssen (219). Wie Deppermann/Hartung (2011) plädiert er dementsprechend dafür, Parameter vor allem als globale Annotationen zu behandeln und sich vornehmlich an möglichst groben Situationsparametern zu orientieren.[4] Auch Mehrfachkodierungen müssen angesichts der Variabilität und Hybridität vieler konkreter Interaktionsformen – also eher auf "Token-" denn auf "Type"-Ebene – erlaubt sein, selbst wenn dies die Metadatendeskription und auf dieser aufsetzende automatische Recherchen natürlich deutlich komplexer macht. Die idealiter bei einer Stratifikationssystematik zu beachtenden Kriterien Monotypisierung, Exhaustivität, Homogeneität und Teilnehmerrelevanz erweisen sich nach Deppermann und Hartung (2011:429f.) bei einem Gesprächskorpus somit letztlich als kaum erfüllbar.

b) Die Kriterien für ein "ausgewogenes Korpus" (nach Lemnitzer/Zinsmeister 2006) müssen, so Deppermann/Hartung (2011), an gesprochene Sprache und soziale Interaktionen angepasst werden. Deppermann/Schmidt (2014:6f.) räumen bezüglich der systematischen Variation der oben genannten Parameter im Hinblick auf Ausgewogenheitsbestrebungen ein:

> 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 miteinbezieht (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 Datener-

[3] Vgl. Deppermann/Hartung (2011:428) auch ausführlicher zur Spezifik von Gattungen.

[4] Die Intersubjektivität bei der Kodierung von Interaktionsparametern, die interpretationsintensiver sind, kann dabei entsprechend korpuslinguistischer Methoden erstens durch explizite Leitlinien und zweitens durch Inter-Rater-Agreement-Messungen oder auch, wie aktuell im Projekt, durch mehrfache Test-Kodierungsdurchläufe mit Revisionen abgesichert werden.

> hebung und -aufbereitung verbunden ist, eine utopische Anforderung. Statt also die Variablen in Kombination zu betrachten, soll daher künftig versucht 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 Bildungsstufen 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.

Deppermann/Hartung (2011:434) führen weiter aus:

> Erst zu einem späteren Zeitpunkt kann man darüber nachdenken, wie man statistisch relevante Stratifikationsparameter gewinnt. Dazu müssten makrosoziologische Parameter berücksichtigt werden, wie z.B. gesellschaftliche Zeitbudgets (bestimmter sozialer Gruppen für bestimmte kommunikative Aktivitäten), soziodemographische Verteilungen, die sektorielle Logik der gesellschaftlichen Praxis usw.

Sowohl zu soziodemographischen Verteilungen als auch zu gesellschaftlichen Zeitbudgets gibt es öffentliche Erhebungen des Statistischen Bundesamtes,[5] die zukünftig zumindest selektiv und schrittweise für einen Abgleich der relativen Anteile von Sprechergruppen und Gesprächstypen herangezogen werden, sobald noch mehr Daten von bislang unterrepräsentierten Sprechergruppen vorhanden sind.

Bezüglich des Problems der kombinatorischen Explosion von Merkmalen vor allem bei den demographischen, also "sekundären" Parametern schlussfolgert Schmidt (2018:220) die Empfehlung, die Zahl der Attribut-Wert-Kombinationen für die demographische Stratifikation möglichst gering zu halten (z.B. etwa nur zwei oder drei Altersspannen oder nur vier bis sechs Sprachregionen anzugeben). Eine systematische Streuung über sekundäre Parameter sei nur für einzelne, möglichst alltägliche Gesprächstypen wie z.B. privates Telefongespräch, Tischgespräch oder auch berufliches Meeting anzustreben (vgl. die Ausführungen im letzten Abschnitt).

In Bezug auf die im Zitat oben ebenfalls anklingenden Problematik des Repräsentativitätsbegriffs ergeben sich, so Deppermann/Hartung (2011), also ganz grundlegende Probleme und Fragen, die innerhalb des Projekts schwerlich endgültig und umfassend zu klären sein werden (438ff.): Was ist überhaupt die als Referenz anzusetzende Grundgesamtheit? Wie müsste eine vollständige Liste aller Gattungen der kommunikativen Praxis aussehen (welche sich zudem ständig verändert) und wie werden ihre Bezeichnungen, Relationen und Grenzen definiert? Wie lässt sich das quantitative Vorkommen im Verhältnis ermitteln, also die globale Zusammensetzung in der Gesprächswirklichkeit, und nach welchen Kriterien wird die Auswahl gewichtet?

Obwohl also beim Korpusdesign keine falschen Idealisierungen der Konzepte von Ausgewogenheit oder qualitativer Repräsentativität suggeriert werden dürfen, können und sollte diese, aufbauend auf einem theoretisch und empirisch fundierten Ansatz, nichtsdestotrotz übergeordnet angestrebt werden. Natürlichen, naheliegenden Gewichtungen wie z.B. der Omnipräsenz von Tischgesprächen vs. der bereichsspezifischen Begrenzung von universitären Prüfungsgesprächen wird bei

---

[5] Vgl. für den ersten Punkt die Hinweise in Abschnitt 4, für den zweiten Punkt: https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/Zeitbudgeterhebung/Zeitverwendung5639102139004.pdf?__blob=publicationFile.

der Erhebung bzw. Datenübernahme nach Möglichkeit ohnehin immer Rechnung getragen.

Die Kategorie "Sprache der Interaktion" als dritte der drei zentralen Klassen von Parametern, welche laut Deppermann und Hartung (2011) systematisch variiert werden sollten (im Anschluss an Merkmale des Sprechereignisses und der Sprecher, s.o.), bezieht sich vornehmlich auf die Varietät bzw. Nationalsprache. Diese Faktoren, ebenso wie weitere Aspekte etwa prominenter sprachlicher Merkmale auf den verschiedenen linguistischen Ebenen, sind in ihren Ausprägungen aber nicht im Sinne externer Steuerung apriorisch absehbar und nicht primär stratifikationsrelevant.

Auf den vorhandenen Daten operierende Annotationstests bezüglich unterschiedlicher sprachlicher Merkmale (syntaktische Einheiten, Intonationsphrasen, lexikalische Mittel, auch Sprachhandlungen etc.) für verschiedene Gesprächstypen können zukünftig aber weitere Erkenntnisse über Einfluss und Beschaffenheit der übergeordneten Merkmalskombinationen liefern (vgl. z.B. Westpfahl i.V. zum Projekt *Segmentation of Oral Corpora* (SegCor)). Weitere Schritte zur Aufbereitung des Korpus und Schaffung neuer Zugänge (für spezifische Zielgruppen, z.B. aus der Variationslinguistik oder DaF-Didaktik, vgl. das Projekt *Zugang zu multimodalen Korpora gesprochener Sprache: Vernetzung und zielgruppenspezifische Ausdifferenzierung* (ZuMult)) setzen ebenfalls an diesem Punkt an.

Praktische Prioritäten beim Aufbau, so Deppermann/Hartung (2011:445), sind die Optimierung der Kosten-Nutzen-Relationen (vgl. Schmidt 2017a), die Absicherung rechtlicher Unbedenklichkeit (vgl. die Angaben in den aktuellen Einverständniserklärungen zum Projekt, zugänglich über die Plattform Gesprächsanalytisches Informationssystem (GAIS)), die Sicherstellung eines möglichst weit reichenden Nutzerinteresses der Daten (vgl. die Ergebnisse der DGD-Nutzerstudie in Fandrych et al. 2016) und die technische Qualität der Datenerhebung (vgl. Schmidt 2016, 2017b, c).

Der folgende Abschnitt diskutiert und reflektiert zunächst die für FOLK zentrale Oberkategorie der Interaktionsdomänen und zieht andere existierende Korpusdesigns zum Vergleich heran. In Abschnitt 3 und 4 werden alle in Deppermann/Hartung (2011) vorgeschlagenen Interaktions- und im Anschluss die in FOLK dokumentierten Sprecherparameter aufgeführt und um Definitionen, FOLK-spezifische Anpassungen und Korrekturen der bisherigen Praxis ergänzt. Bei den weitgehend klar definierten und objektiv abgrenzbaren Kategorien veranschaulichen einfache Grafiken die quantitative Verteilung der Daten auf dem Stand des letzten Release (2.10, Mai 2018). Sofern nicht anders vermerkt, wird bei diesen immer die Gesprächsdauer, nicht die Token- oder Gesprächsereigniszahl, als Vergleichsmaß herangezogen. Zudem wird unterschieden und gekennzeichnet, welche Parameter künftig als stratifikationsleitend und welche als ergänzend zu behandeln sind (vgl. zu der Unterscheidung Deppermann/Hartung 2011; Love/Dembry/Hardie/Brezina/McEnery 2017). Nur an ersteren richtet sich das Korpusdesign bzw. der -ausbau aus; die meisten der anderen Kategorien sollen dennoch zur Recherche herangezogen werden können. Der Text schließt mit Erläuterungen zu künftigen Ausbaustrategien und zu den nächsten Arbeitsschritten im Projekt.

2. Interaktionsdomänen

2.1. Kategorien für FOLK

Die FOLK-Interaktionen werden bislang übergeordnet in die Kategorien "Alltag", "Institution", "Öffentlich" und "Sonstiges" eingeteilt, welche als "Interaktionsdomänen" künftig explizit die oberste Kategorienebene der Stratifikation bilden. Diese Ebene wird zusammen mit den zwei weiteren stratifikationsleitenden Parametern, Lebensbereichen und Aktivitäten (vgl. Abschnitt 3.1.1), das übergeordnete Konzept des 'Interaktionstyps' konstituieren. Die Kategorien bieten somit eine parametrisierte Aufschlüsselung des bislang eher vage oder divers gebrauchten Begriffes, welche operationalisierbar ist, da sie 'Interaktionstyp' als Merkmalsbündel begreift.

Unter die oben zuletzt genannte Domänen-Kategorie "Sonstiges" fallen alle Interviews und die experimentellen Maptask-Interaktionen, die sich weder den anderen Großkategorien noch einer weiteren eigenen Interaktionsdomäne zuordnen lassen, da es sich um unterschiedliche spezielle, elizitierte Settings handelt.[6]

Öffentliche Interaktionen sind aktuell die Schlichtungsgespräche zu Stuttgart 21 und TV-Debatten, die zudem massenmedial vermittelt sind, und Podiumsdiskussionen.

Die Arbeitsdefinition dieser Kategorie für FOLK lautet wie folgt:

> **Öffentliche Interaktionen** sind Gespräche, die im Rahmen öffentlich zugänglicher und/oder massenmedial vermittelter Anlässe stattfinden. Sie haben meist ein Publikum und behandeln z.B. politische, wissenschaftliche, andere gesellschaftlich relevante oder unterhaltende Themen.

Was die beiden Begriffe "Alltag" und "Institution" genau bedeuten bzw. umfassen, ist nicht ganz so eindeutig zu beantworten wie es zunächst scheinen mag. Laut Duden ist "Alltag" entweder als "tägliches Einerlei, gleichförmiger Ablauf im [Arbeits]leben" zu umschreiben oder der Begriff wird als Synonym zu "Werktag, Arbeitstag" erklärt.[7] Die Definition zielt also auf Handlungs- und Verhaltensroutinen bzw. gewohnheitsmäßige Muster, die einerseits auch Kommunikation am Arbeitsplatz, Arzt- und Behördenbesuche und Interaktionen in Geschäften etc. einschließen würden, andererseits aber Urlaube, Feste und Feiertage ausschließen, da "Alltag" und "Festtag" in der semantischen Definition einander gegenübergestellt werden. Bei FOLK werden dagegen auch Gespräche auf Urlaubsreisen oder bei Festen zur Alltagsinteraktion gezählt, dagegen Arzt- oder Behördenbesuche, Dienstleistungsinteraktionen in Geschäften und Interaktionen am Arbeitsplatz von diesem Bereich abgegrenzt bzw. ihm gegenübergestellt und zu institutioneller Kommunikation gerechnet (vgl. zur Abgrenzungsproblematik ausführlicher auch Schütte 2001; zum "homileïschen Diskurs" in Abgrenzung von institutioneller, aufgabenorientierter Kommunikation auch Ehlich/Rehbein [1980] (2011)).

Im Sinne der Erfassung privater, d.h. nicht-öffentlicher und nicht-institutioneller Situationen werden diese Interaktionen folgendermaßen definiert, abgegrenzt und zukünftig auch von "Alltag" zu "Privat" umbenannt:

---

[6] "Elizitiertheit" wird im XML-Schema für die Metadaten als eigener Parameter dokumentiert. Der Wert "elizitiert" gilt demnach für alle Interaktionen der Kategorie "Sonstiges"; für die Interaktionen der anderen Domänen gilt der Wert "spontan".

[7] https://www.duden.de/rechtschreibung/Alltag

> **Private Interaktionen** sind informelle Gespräche mit Familie und/oder Freunden und Bekannten, inklusive Urlaub, Festen etc., aktivitätsgeleitet oder frei, aber nicht (formelle oder auch informelle) Interaktionen in institutionellen Umfeldern (Arzt, Behörden etc.) oder in öffentlichen Kontexten.

Die Duden-Definition zum Begriff "institutionell" ist ebenfalls nicht per se ausreichend präzise – dort lauten die Bedeutungsangaben: "eine Institution betreffend, zu ihr gehörend; durch eine Institution gesichert; mithilfe einer Institution"; "als Institution geltend, wirksam".[8] Der Begriff der Institution wird wiederum definiert als "einem bestimmten Bereich zugeordnete gesellschaftliche, staatliche, kirchliche Einrichtung, die dem Wohl oder Nutzen des Einzelnen oder der Allgemeinheit dient" oder mit Hinweis auf die Soziologie als "bestimmten stabilen Mustern folgende Form menschlichen Zusammenlebens",[9] was aber als Beschreibung wiederum beispielsweise auch auf die Familie zuträfe. In FOLK wird die Kategorie konstant sowohl über den Ort als auch zugleich über institutionell oder allgemeiner professionell vor-definierte Rollen und Handlungen (zumindest einer Partei) der Gesprächsteilnehmer definiert (vgl. auch Heritage/Clayman 2010).

Eine praktikable Definition für die Einteilung in FOLK lautet folgendermaßen:

> **Institutionelle Interaktionen** sind Gespräche, die im Rahmen institutioneller Räumlichkeiten bzw. Handlungen mit Personen in der Rolle institutioneller bzw. professioneller Vertreter und mit den entsprechenden konstitutiven Aktivitäten stattfinden, also z.B. jegliche Interaktionen am Arbeitsplatz, in Ausbildungsstätten, in Behörden, in medizinischen, aber auch Dienstleistungs- bzw. Verkaufskontexten ebenso wie im organisierten Vereinsleben oder in Bereichen von Religion, Kunst, Unterhaltung und Sport.

Somit werden also auch alle Interaktionstypen im Dienstleistungs- bzw. Verkaufssektor, die in anderen Systematiken teils gesondert laufen (vgl. weiter unten in diesem Abschnitt), explizit unter institutionelle Kommunikation gefasst.[10]

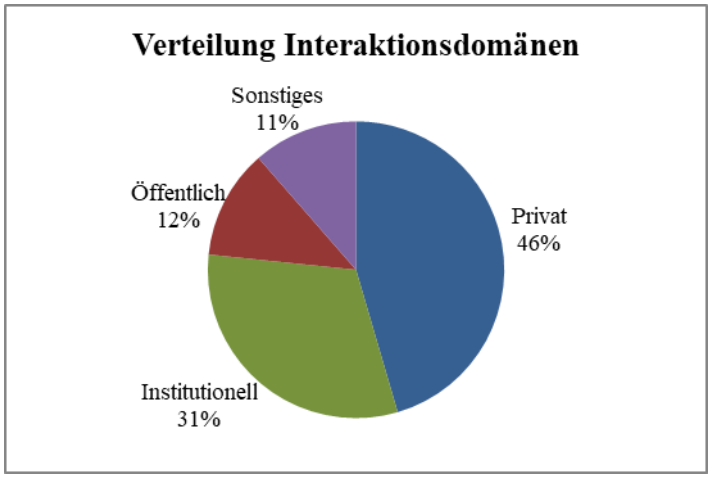
Aktuell sieht die Verteilung der FOLK-Daten auf die angesetzten Interaktionsdomänen im Überblick folgendermaßen aus:

---

8 https://www.duden.de/rechtschreibung/institutionell

9 https://www.duden.de/rechtschreibung/Institution

10 Auch mehr oder weniger private Pausengespräche unter Kollegen am Arbeitsplatz wurden in FOLK bislang der institutionellen Domäne zugeordnet. Dies ist hinsichtlich der oben genannten Definition problematisch, da es sich nicht um aufgabenorientierte und rollengebundene Kommunikation handelt. Dennoch wird die Zuteilung bei diesen Beispielen vorerst beibehalten, da die rollengebundenen Identitäten etwa von Chef und Mitarbeitern sowie die institutionelle Verortung über die Räumlichkeiten und berufsbezogene Themen bei den bislang vorliegenden Daten eine Abgrenzung von rein privaten Gesprächen erlauben.



Grafik 1: Interaktionsdomänen

2.2. Parametrisierung bei Biber

Die Umbenennung von "Alltag" in "Privat" entspricht auch der Einteilung des Parameters "setting" in "private-personal", "public" und "institutional" bei Biber (1993), dessen Arbeiten unter anderem zum Konzept des Registers für die Korpuslinguistik als grundlegend gelten können.

Biber (1993:245) führt dazu aus:

> Work on the parameters of register variation has been carried out by anthropological linguists such as Hymes and Duranti, and by functional linguists such as Halliday (see Hymes, 1974; Brown and Fraser, 1979; Duranti, 1985; Halliday and Hasan, 1989). In Biber (1993a), I attempt to develop a relatively complete framework, arguing that 'register' should be specified as a continuous (rather than discrete) notion, and distinguishing among the range of situational differences that have been considered in register studies. This framework is overspecified for corpus design work – values on some parameters are entailed by values on other parameters, and some parameters are specific to restricted kinds of texts. Attempting to sample at this level of specificity would thus be extremely difficult. For this reason I propose in Table 1 a reduced set of sampling strata, balancing operational feasibility with the desire to define the target population as completely as possible.

Bibers Tabelle für "situational parameters listed as hierarchical sampling strata" enthält (leicht gekürzt und zusammengefasst, JK) folgende Punkte:

1. Primary channel.
Written/spoken/scripted speech

2. Format.
Published/not published (+ various formats within 'published')

3. Setting.
Institutional/other public/private-personal

4. Addressee.
(a) Plurality. Unenumerated/plural/individual/self
(b) Presence (place and time). Present/absent
(c) Interactiveness. None/little/extensive
(d) Shared knowledge. General/specialized/personal

5. Addressor.
(a) Demographic variation. Sex, age, occupation, etc.
(b) Acknowledgement. Acknowledged individual/institution

6. Factuality.
Factual-informational/intermediate or indeterminate/imaginative

7. Purposes.
Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship.

8. Topics [...]

Während Punkt 1 die mediale Realisierung adressiert, wovon für FOLK natürlich nur der Wert "spoken" relevant ist, ist Punkt 2 für gesprochensprachliche Daten eher nur hinsichtlich massenmedialer Vermittlung und Öffentlichkeitsgrad interpretierbar. Punkt 3 fokussiert auf die oben angeführten Interaktionsdomänen. Punkt 4 bezieht sich auf Teilnehmerzahl und Medium hinsichtlich der Unterscheidung in *face-to-face* oder "vermittelt", Sprecherwechsel und Vertrautheitsgrad (siehe zu diesen Begriffen auch weiter unten bei den entsprechenden Parametern). Punkt 5 wird bei der Stratifikation von FOLK zunächst unabhängig von den Aspekten zur Interaktion selbst unter sekundäre bzw. Sprecherparameter gefasst (Geschlecht, Alter, Beruf, Bildung, soziale Rolle). Punkt 6 adressiert so etwas wie die Modalität, welche für die FOLK-Stratifikation etwa sprachlichen Modalitäten der Interaktion entsprechen könnte, bei Biber wohl im schriftsprachlichen Bereich, aber auch auf die Unterscheidung *fiction* vs. *non-fiction* abzielt, welche für FOLK keine große Rolle spielt. Punkt 7 verweist auf Gesprächshandlungen bzw. -zwecke und -ziele, eine Kategorie, die für FOLK hinsichtlich der einleitend genannten kommunikativen Gattungen bzw. des Interaktionsparameters "Aktivität" (vgl. Abschnitt 3.1.1) interpretierbar ist. Der letzte Parameter zu (Gesprächs-) Themen muss schließlich eine offene und erweiterbare Liste enthalten und wird als solche für die FOLK-Interaktionen auch geführt.

Während Biber gesprochensprachliche Faktoren durchaus berücksichtigt, bleiben seine Überlegungen insgesamt überwiegend schriftsprachlich orientiert, sodass einige für mündliche Interaktionen relevante Distinktionen in seiner Darstellung fehlen.

2.3. Andere Korpora

Auch bei anderen Korpusprojekten zu gesprochener Sprache finden sich Ansätze zu (mehr oder weniger expliziten) Kategorien und Parametern, welche für die Datensystematisierung und hinsichtlich angestrebter Ausgewogenheitskriterien angesetzt werden (vgl. die Zusammenstellung und Übersicht bei Merkel/Schmidt 2009 und Schmidt 2018). Im Folgenden werden einzelne weitgehend aktuelle und für den Vergleich mit FOLK interessante Projekte noch einmal selektiv aufgeführt:

1. Das Slowenische GOS-Korpus

Das Slowenische GOS-Korpus (GOvorjene Slovenščine; Verdonik et al. 2013) mit Aufnahmen von 2004 bis 2010 enthält 120 Stunden gesprochene Sprache, die eingeteilt werden nach den Kategorien "public" und "non-public" mit Subspezifikationen: "public: information (television, radio), educational (personal contact [in primary/secondary school]), entertainment (television, radio)" und "non-public: non-private (telephone, personal contact), private (telephone, personal contact)". Die Daten wurden offenbar jeweils regional (und teils auch demographisch) ausgewogen gestreut (nach den sekundären Parametern weiblich/männlich, unter/über 35, niedrige/hohe Bildung).[11]

Während institutionelle Kommunikation sozusagen ex negativo als "non-public non-private" definiert wird, werden die im FOLK-Korpus als institutionell definierten schulischen Interaktionen im GOS-Korpus aus nicht ersichtlichen Gründen als öffentlich gelabelt und nach zwei Schultypen unterschieden. In der Kategorie "public" fehlen dann auch die Telefongespräche in Abgrenzung zum persönlichen Kontakt (*face-to-face*), was bei den nicht-öffentlichen Interaktionstypen aber jeweils beides abgedeckt wird, sodass berufliche Telefongespräche separat und als institutionell kategorisiert werden.[12]

2. Corpus Gesproken Nederlands (CGN)

Auch das CGN (Corpus Gesproken Nederlands; Oostdijk 2002) macht die Grobunterscheidung in "Private" vs. "Public", jeweils mit den Subkategorien 'dialogisch'/'multilogisch' oder 'monologisch'. Bei 'Public' wird zusätzlich auch nach 'übertragen' oder 'nicht übertragen' (also massenmediale Vermittlung) unterschieden, und sowohl bei 'Public' als auch bei 'Private' nach 'spontan' vs. 'mehr oder weniger vorbereitet'. Bei 'Private' und 'spontan' findet sich zusätzlich noch eine Angabe zu 'direkt' oder 'indirekt' (= medial vermittelt, Telefon). Das "socio-situational setting" fungiert somit auch hier als Oberkategorie, hinzu kommen die Parameter Gesprächsziel, Medium, Anzahl der TN, Sprecher-Hörer-Beziehung. Eine systematische Variation von Sprecherparametern ist nicht dokumentiert und der Großteil der – wenn auch innerhalb dieser Kategorie sehr breit variierten – Daten wird von Interaktionen mit den Labels "spontaneous face-to-face" und "spontaneous telephone conversations" bestritten.

3. British National Corpus (BNC)

Im BNC (British National Corpus; Crowdy 1993)[13] sind ursprünglich 10% der 100 Mio. Wörter Gesamtbestand gesprochensprachlich, wovon wiederum 50% nach Gesprächstypen zusammengestellt wurden ("context-governed part", taxonomisch, höchste Kategorien: educational, business, public/institutional, leisure) und 50% nach soziodemographischen Merkmalen (Alter, Geschlecht, soziale Schicht, Region in GB). Von dem "context-governed part" der "spoken component" (der auch nach Sprecherparametern balanciert werden soll) sind 40% monologisch, 60% dialogisch. Die Kategorie "Public/Institutional" enthält folgende Interaktionstypen: "political speech, sermons, public/government talk, religious

[11] vgl. http://www.korpus-gos.net/Support/About

[12] Diese fehlen in FOLK wiederum bislang gänzlich.

[13] Vgl. auch http://www.natcorp.ox.ac.uk/corpus/creating.xml; http://www.natcorp.ox.ac.uk/archive/worldURG/index.xml.

meetings, parliamentary and legal proceedings". Der Typ "Educational/Informative" enthält "lectures/talks/demonstrations/news commentary, classroom interaction", der Typ "Business" enthält "company talks and interviews, trade union talks, sales demonstrations, consultations" und der Typ "Leisure": "speeches, broadcast sports commentaries, talks to clubs, phone-ins, broadcast chat shows, club meetings" (also auch (semi-)öffentliche Settings).

In den Neuerhebungen zum Spoken BNC2014 (vgl. Love/Dembry/Hardie/Brezina/McEnery 2017) sind nur noch informelle, spontane Alltagssituationen enthalten, also kein Bereich für "context-governed" oder "task-oriented" mehr. Es handelt sich um eine opportunistische Erhebung mit geplanten darauffolgenden Nacherhebungen in unterrepräsentierten Bereichen. Das Korpus postuliert seine Eignung für individuell definierte Subkorpora, was auch für FOLK ein anzustrebender Aspekt ist.

4. Göteborg Spoken Language Corpus (GSLC)

Das GSLC (Göteborg Spoken Language Corpus)[14] ist ein opportunistisch wachsendes, auf eine möglichst große Varianz von Interaktionstypen ausgerichtetes Korpus, das verschiedene Aktivitätstypen enthält. Diese werden als unterschiedliche Aktivitäts- bzw. Interaktionstyp- und Settingparameter undifferenziert aufgelistet, aber nicht ersichtlich systematisiert und variiert:

Discussion; Retelling of Article; Interview; Task-Oriented Dialogue; Informal Conversation; Role Play; Trade Fair; Arranged Discussions; Formal Meeting; Consultation; Shop; Dinner; Market; Auction; Factory Conversation; Party; Games & Play; Phone; Travel Agency; Court; Church; Lecture; Hotel; Therapy; Bus Driver-Passenger. Zu diesen Interaktionsbezeichnungen finden sich jeweils noch Subtypen.

5. Michigan Corpus of Academic Spoken English (MICASE)

Das thematisch spezifischere MICASE (Michigan Corpus of Academic Spoken English)[15] listet eine breite Variation von "speech event types" aus dem akademischen Bereich auf und gibt dazu die genauen Anteile nach Tokens an, jeweils nach Fachbereich und Studenten, außerdem genauer nach Geschlecht, akademischem Stand und Sprachstand der Sprecher, und strebt eine relative Ausgewogenheit der Parameter an. Die quantitativen Verhältnisse, inhaltlichen Erhebungsrichtlinien (bezüglich vollständiger Gespräche, Varianz hinsichtlich mehrerer Interaktions- und Sprecher-Parameter) und die Systematik und Darstellung der Metadaten werden ausführlich und transparent erläutert.

6. Corpus of informal spoken Czech

Das "Corpus of informal spoken Czech" (Projekt ORTOFON)[16] enthält informelles, spontanes gesprochensprachliches Tschechisch (2012-2017) zwischen einander vertrauten Sprechern (meist Familie und/oder Freunde, zwei oder mehr Personen) aus der gesamten Tschechischen Republik. Laut den Angaben online ist es "fully balanced regarding all the basic sociolinguistic speaker categories (gender, age group, level of education and region of childhood residence)":

---

[14] vgl. http://www.qualitative-research.net/index.php/fqs/article/view/1026/2215

[15] http://quod.lib.umich.edu/m/micase

[16] http://wiki.korpus.cz/doku.php/en:cnk:ortofon

> [..] The first three categories, i.e. gender, age, education, were assigned binary values (see picture), while the fourth category was divided into ten groups i.e. ten dialectal regions. The following picture displays the distribution of the binary categories within one dialectal region. Each region should, therefore, contain the same number of words from men and women, from speakers of ages 18-34 years and those over 35 years, and from speakers with a high school education and those with a university education.

Grafik 2: ORTOFON

> […] Taking into account the target corpus size (1.000.000 words), the target for every category presented by the combination of four variables - gender(2) × age(2) × education (2) × dialectal region of residence up to the age of 15 years (10) - was set at 12 500 words. In the effort to achieve the highest possible speaker variability within the scope of each category, a minimum of five different speakers was set. The aim of this provision to limit the influence of idiolect (sic.!).

Zumindest für definierte Subkorpora wie Tisch- und vor allem auch Telefongespräche wird auch für FOLK eine entsprechende systematische Variation über sekundäre Parameter für den künftigen Ausbau angestrebt (vgl. die Ausführungen in Abschnitt 5). Bezüglich der Dimension des Interaktionstyps wird bei ORTOFON dagegen nicht weiter unterschieden (obwohl laut der Angaben auf der Seite genauere Informationen hierzu bei den Metadaten vorhanden sind) und die aufgenommenen Interaktionen fallen alle in die Kategorie "private Interaktion".

Es zeigt sich, dass die genannten Korpusprojekte ihre Daten nach Interaktions- und/oder demographischen Sprecherparametern unterscheiden und kategorisieren. Besonders das Spoken BNC und MICASE, bezüglich demographischer Parameter auch das ORTOFON-Korpus haben relativ explizite und elaborierte Stratifikationskonzepte. Sie können somit für die Reflexion über die für FOLK anzusetzenden Parameter, die Wertbereiche und ihre Variation einige Anhalts- und Vergleichspunkte geben. Gleichzeitig bleibt die Einteilung in Interaktionstypen, wenn sie überhaupt vorhanden ist, außer bei MICASE aber überall relativ grob oder unsystematisch, weist teils starke Ungleichgewichte oder konzeptionelle und inhaltliche Lücken auf oder ist nicht transparent.

Für die weitere Ausarbeitung der FOLK-Stratifikation wird dagegen eine explizite, möglichst transparente und umfassende Systematik für die Dokumentation, die Repräsentation und den weiteren Ausbau des Korpus vorgestellt, die ab dem Release 2.11 (Frühjahr 2019) schrittweise integriert werden soll. Die im Folgenden aufgeführten Kategorien und Wertesets werden bewusst als konzeptuelle und auch methodologisch begründete Entscheidungen reflektiert, welche – trotz der möglichst induktiven Entwicklung aus den Daten – auch anders hätten getrof-

fen werden können. Sämtliche Parameter wurden aber projektintern in mehrfachen Kodierungsdurchläufen mit Mitarbeitern und Hilfskräften getestet, diskutiert und überarbeitet.

3. Weitere stratifikationsleitende und -ergänzende Interaktionsparameter

Nachdem mit den Interaktionsdomänen der erste stratifikationsleitende Parameter im vorangegangenen Abschnitt bereits diskutiert wurde, werden im Folgenden weitere Kategorien vorgestellt, definiert und diskutiert – zunächst diejenigen, die als stratifikationsleitend behandelt werden und gemeinsam den Interaktionstyp konstituieren (vgl. einleitend Abschnitt 2.1), anschließend ergänzende Parameter.

3.1. Gesellschaftlicher Lebensbereich und Aktivitäten

Bei Deppermann/Hartung (2011) bildet ein zentraler Parameter der "gesellschaftliche Sektor" bzw. der "gesellschaftliche Handlungsbereich". Der Begriff "Lebensbereich" wurde projektinternen als der verständlichere bewertet und daher ausgewählt. Dieser Parameter konstituiert die zentrale Subkategorie der ersten Unterscheidung in Interaktionsdomänen und bildet somit den zweiten stratifikationsleitenden Parameter für FOLK. Auch Schmidt (2018) setzt diesen Parameter, mit einem aus einem Praxistest entwickelten, etwas abweichenden Werteset, an. Aus eigenen Anwendungstests auf die vorhandenen Daten und mehreren projektinternen Kodierprozessen entwickelte sich das folgende Kategorienset mit weiteren Ergänzungen und Überarbeitungen der ursprünglichen Vorschläge:

- Schule und Unterricht werden zusammen mit Nachhilfe, Prüfungen in der Hochschule etc. unter dem Label "Bildung" zusammengefasst – auch praktischere bzw. berufliche Ausbildungsinteraktionen wie die Trainings in einer Hilfsorganisation fallen dann darunter.

  > Die Kategorie "**Bildung**" umfasst institutionalisierte Lehr-Lern-Interaktionen, die (mit unterschiedlichen Formen und Methoden) der Wissensvermittlung dienen.

- Die Kategorie "Behörden" umfasst Interaktionen im Bereich von Verwaltung und Recht (z.B. die in FOLK vorhandenen Gespräche im Polizeirevier und im Arbeitsamt, aber auch formale Studienberatungen).

  > Die Kategorie "**Behörden**" umfasst Interaktionen in institutionellen Einrichtungen, die mit (öffentlicher) Verwaltung und Recht befasst sind.

- Unter die Kategorie "Interprofessionelle Kommunikation" fällt Kommunikation in Unternehmen (Wirtschaft), Industrie, sozialen Einrichtungen etc., bei der professionelle Akteure untereinander agieren, wohingegen Interaktionen mit Kunden, Patienten etc. in die Kategorien Dienstleistung, Medizin, Bildung, Kunst etc. fallen.

  > Die Kategorie "**Interprofessionelle Kommunikation**" umfasst Interaktionen in allen beruflichen Feldern (Wirtschaft, Industrie, soziale oder medizinische Einrichtungen etc.) zwischen professionellen Akteuren. Handelt es sich da-

gegen um Interaktionen zwischen professionellen und nicht-professionellen Akteuren (Kunden, Patienten etc.), wird spezifiziert nach den Kategorien Medizin/Gesundheitswesen (z.B. Arzt-/Pfleger-Patienten-Kommunikation), Dienstleistung (z.B. Verkaufsgespräche), Kunst/Unterhaltung/Sport (z.B. Veranstaltungen mit Publikum), Bildung (Schüler-Lehrer-Interaktionen) usw.

- Weitere Kategorien werden wie folgt gefasst:

  Die Kategorie "**Vereinsleben und Selbstverwaltung**" umfasst Interaktionen im Bereich organisierter ehrenamtlicher Tätigkeiten zu gemeinnützigen Zwecken und auch Gremienarbeit in verschiedenen institutionellen Bereichen der Selbstverwaltung (z.B. Gemeinderats- oder Studienratssitzungen etc.).

  Die Kategorie "**Religion/Kirche**" umfasst Interaktionen in institutionalisierten religiösen Bereichen, wie Gottesdienste, Konfirmandenunterricht, Beichte etc.

  Die Kategorie "**Kunst/Unterhaltung/Sport**" umfasst Interaktionen aus institutionalisierten Bereichen der Kultur – musikalische, theatralische, künstlerische Produktion und Rezeption, sportliche Wettkämpfe etc.

  Die Kategorie "**Dienstleistung**" umfasst Verkaufsgespräche (inklusive Beratung) mit Kunden in allen Arten von geschäftlichen Einrichtungen (Supermarkt, Gartencenter, Kiosk, Apotheke etc.).

  Die Kategorie "**Medizin/Gesundheitswesen**" umfasst Gespräche in den medizinisch-versorgenden, pflegerischen Bereichen, also Interaktionen von Patienten mit Ärzten, Pflegern, Krankenschwestern etc.

- Öffentliche Gespräche werden für die Kategorie Lebensbereich untereilt in: "Politik", "Unterhaltung", "Wissenschaft", "Wirtschaft". Die Einteilung entspricht hier eher dem jeweiligen thematischen Fokus der öffentlichen, gegebenenfalls massenmedial vermittelten Interaktion. Die Themenfelder selbst reichen selbstverständlich auch in andere, institutionelle Lebensbereiche hinein und können intern noch differenzierter sein.

Divergierende Beteiligungsspezifika bleiben hier wie in anderen Bereichen neben Überlappungsbereichen und Mehrfachkodierungen ein Problem, können aber im Falle systematischer Doppelbelegungen auch zugelassen werden (vgl. z.B. die Doppelkategorisierungen "Bildung/Interprofessionelle Kommunikation" bei Fortbildungen, "Bildung/Dienstleistung" bei Fahrschul- oder Nachhilfestunden).

Für die Interaktionsdomäne "Privat" wurde bislang nur die Unterscheidung in "thematisch freie" Interaktionen vs. Interaktionen mit "begleitenden Aktivitäten" vorgeschlagen. Für das Stratifikationskonzept wird diese Unterscheidung aber erstens begrifflich noch etwas anders gefasst und definiert und zweitens auf eine untergeordnete Kategorisierungsebene verlagert.

Als dritter stratifikationsleitender Parameter wird "Aktivität" somit künftig bei privaten Gesprächen nach den Werten "nicht aktivitätsgeleitet" (bei thematisch freien und nicht-empraktischen Tischgesprächen, Kneipengesprächen etc.) vs. "aktivitätsgeleitet" (sowohl bei themen(bereichs)fixierten Planungsgesprächen als auch bei empraktischen Interaktionen wie Umräumen) unterschieden. Im zweiten Fall werden die konkreten interaktionsprägenden Aktivitäten oder bestimmenden Themen (wie "Kochen", "Essen", "Streichen", "Aufräumen", "Vorlesen" etc.) als offene Werteliste aufgeführt.

> Ein privates Gespräch (Interaktionsdomäne: privat / Lebensbereich: privat) ist **aktivitätsgeleitet**, wenn das Gespräch aus Anlass einer geplanten (möglichst klar benennbaren) Aktivität stattfindet und (zu erwarten ist, dass) diese Aktivität den Gesprächsinhalt wesentlich prägt.

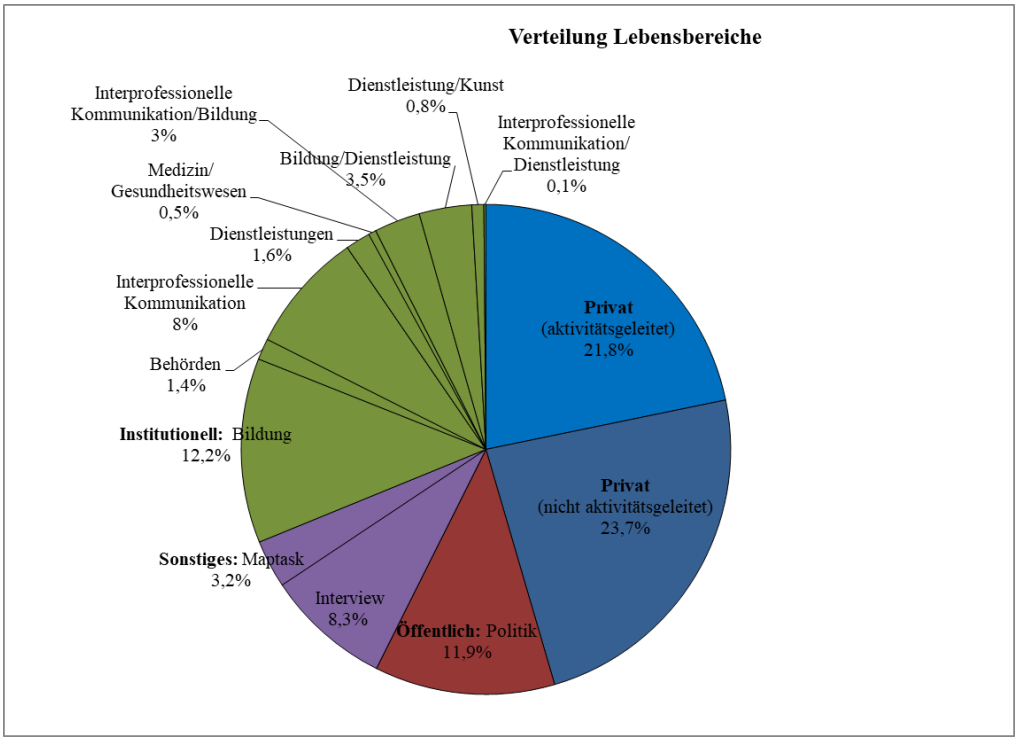
Institutionelle und öffentliche Interaktionen, die auf der Ebene der Lebensbereiche bereits durch einen Wert spezifiziert werden, gelten als per se aktivitätsgeleitet, da sie wesentlich durch bestimmte Aufgabentypen und Themen geprägt sind. Für sie wird auf der Ebene dieser Kategorie ein Wert vergeben, der die Hauptaktivität der Interaktion charakterisiert, z.B. "Meeting" bei einem institutionellen Gespräch aus dem Lebensbereich "Interprofessionelle Kommunikation". Auch Gespräche der "Sonstiges"-Domäne, die in der Kategorie "Lebensbereich" ebenfalls nicht weiter spezifiziert werden, gelten als aktivitätsgeleitet und werden als "Maptask" oder als "Interview" spezifiziert, bei den vorhandenen Interviews zusätzlich mit der genaueren Information "sprachbiographisch", "ethnographisch" oder "biographisch".

Folgende drei Aspekte sollen zukünftig eine größtmögliche Konsistenz und Stringenz bei der Kodierung neuer Gesprächsdaten sichern:

- die vereindeutigenden und abgrenzenden Definitionen,
- das aus den Daten entwickelte Werteset,
- die Kodierungen der vorhandenen Daten als Orientierung für weitere neue Gesprächstypen (vgl. auch die projektinterne Metadaten-Dokumentation als Leitfaden).

Konsistente Wertesets wie jenes für die Lebensbereiche sind für die Systematisierung (auch mit Blick auf eine mögliche Gattungssystematik) zentral. Die bisher geführten, im Datenübernahmeprozess als Kurzbezeichnung entstandenen "Ad hoc"-Benennungen für die Gesprächstypen (z.B. "Meeting in einer sozialen Einrichtung", "Spieleinteraktion mit Kindern" etc.) bestehen dagegen aus sehr inkonsistenten Angaben, die variierend auf Teilnehmer(rollen), übergeordneten Gesprächszweck, Ort/Institution, Hauptaktivität, Medium etc. referieren und auch Inkonsistenzen beim Wechsel zwischen den Bezeichnungen "Interaktion", "Gespräch" und "Kommunikation" aufweisen. Die systematisch erfassten Informationen zur Interaktionsdomäne, Lebensbereich und Aktivitäten als operationalisierbare Parameter für das Konzept des Interaktionstyps liefern die gleichen bzw. sogar umfassendere und vor allem systematischere Informationen zu den Gesprächen. Trotz dieser starken Argumente werden die Kurzbezeichnungen (unter eben diesem Begriff) aber sowohl im Metadatenschema als auch der Dokumentation parallel zu den systematisierten Kategorien weitergeführt, um den gewohnten raschen, intuitiven Zugang zu gewährleisten. Vor allem bei nicht-aktivitätsgeleiteten privaten Interaktionen oder auch bei durch das Medium charakterisierten Telefongesprächen kann so das jeweils prägende Merkmal direkt erfasst werden.

Gemessen nach der Gesprächsdauer sieht die Verteilung der Interaktionen auf die Lebensbereiche innerhalb der jeweiligen Interaktionsdomänen auf der aktuellen Datenbasis und mit der aktuellen Kodierung (ohne Spezifikation der Aktivitäten) folgendermaßen aus:



Grafik 3: Lebensbereiche

Noch stark unterrepräsentiert bzw. nicht vorhanden sind also die Bereiche Vereinsleben und Selbstverwaltung, Religion und Kirche, reine Dienstleistungen, Medizin/Gesundheitswesen (in Interaktionen mit Patienten, also nicht zwischen professionellen Vertretern). Aber auch die Kategorien Interprofessionelle Kommunikation und Bildung sollten intern noch variabler besetzt werden, ebenso wie die privaten und öffentlichen Lebens- und Aktivitätsbereiche insgesamt.

Die weiteren, im Folgenden aufgeführten Parameter sind nicht primär stratifikationsleitend, sondern als ergänzend und für Recherchen zusätzlich nutzbar zu verstehen.

3.2. Aufnahmeort

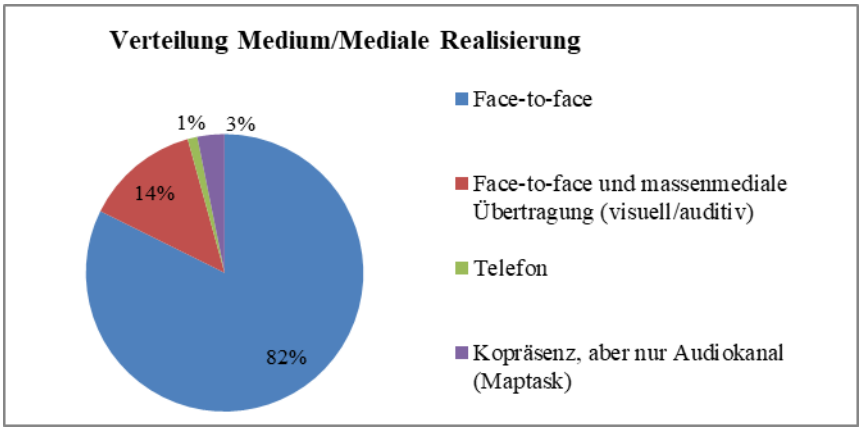
Der Ort (Land und Stadt bzw. Region) der Aufnahme wird – neben den Räumlichkeiten, gegebenenfalls auch der betreffenden Institution – als offene Angabe dokumentiert. Grundsätzlich soll die räumliche Streuung der Aufnahmen für FOLK zwar über die Herkunftsregionen der einzelnen Sprecher geführt und somit systematisch als Sprechermerkmal konzeptualisiert werden, vgl. daher in Abschnitt 4 die Ausführungen zur Präzisierung des Merkmals "Aufenthaltsregion" und zur Zuteilung der geographischen Ortsangaben zu zwei Kategoriensets. Dort wird allerding auch erläutert, dass und warum die Orientierung an Herkunftsregionen aktuell noch sehr problematisch ist. Daher wird der Parameter Aufnahmeort für die räumliche Systematisierung – vorläufig – vorgezogen.

3.3. Medium / Mediale Realisierung

Der Parameter Medium bzw. Mediale Realisierung (im Metadaten-Schema und -Formular) bezeichnet die technische Übertragungsmedialität und kann die Werte "*face-to-face*", "Telefon", "Bildtelefon" und "massenmedial vermittelt" (z.B. TV-Nachrichten, Rundfunk, Internet) erhalten.[17]

Das Problem divergierender Beteiligungsperspektiven besteht hier bezüglich der basalen Perspektiven von Produktion vs. Rezeption. Massenmedial übertragene Interaktionen (mit interagierenden Teilnehmern und/oder Studiopublikum) spielen in FOLK bislang allerdings eine untergeordnete Rolle. Der Aspekt der massenmedialen Übertragung findet sich aber auch bei den vorhandenen Aufnahmen der Produktion einer Radiosendung. Die wichtigste Unterscheidung ist *face-to-face* vs. Telefon, wobei die *face-to-face*-Interaktionen den weit überwiegenden Anteil ausmachen.

Für diesen Parameter ergibt sich in FOLK aktuell folgende Verteilung:



Grafik 4: Mediale Realisierung

3.4. Teilnehmerzahl und -konstellation

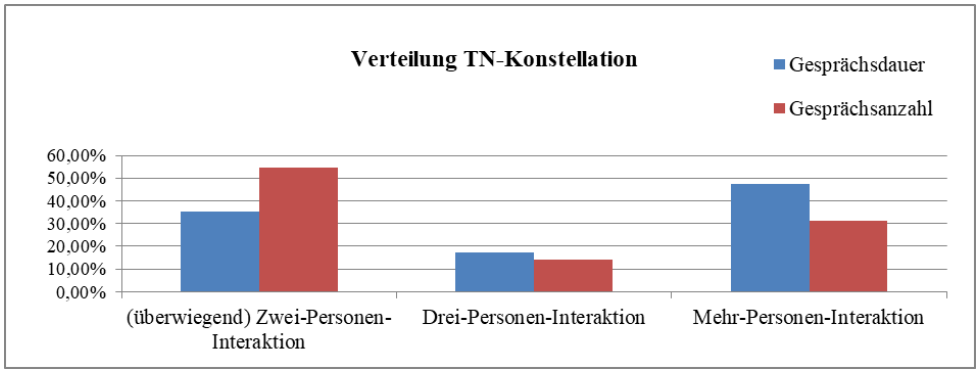
Die Zahl der Teilnehmer bildet einen wichtigen Aspekt der Gesprächscharakterisierung und wird zunächst numerisch (für sämtliche anwesenden Personen, ohne Unterschied bezüglich verbaler Beteiligung, Forscherrolle etc.) erfasst. Zusätzlich werden bei diesem Parameter die Wertbereiche "Zwei-Personen-Interaktion", "Drei-Personen-Interaktion" und "Mehr-Personen-Interaktion" (für alle Interaktionen ab vier beteiligten Personen) und zu jedem dieser drei Bereiche noch eine Kategorie mit dem Adjektiv "überwiegend" vergeben. So soll die tatsächliche Gesprächskonstellation der verbal beteiligten Teilnehmer genauer charakterisiert

[17] Einen – nicht systematisch berücksichtigten – Sonderfall bilden hier die experimentellen Maptask-Interaktionen: Die TeilnehmerInnen sind zwar ko-präsent, können sich aber nicht sehen und dementsprechend nur über den auditiven Kanal miteinander kommunizieren. In der Grafik werden sie als vierte, aktuell sogar häufiger als die Telefongespräche vorkommende Variante mitaufgeführt. Interaktionen mit dem Wert "Bildtelefon", also Videokonferenzen oder Skype-Gespräche, fehlen in FOLK bisher, nur bei einem WG-Casting wird partiell ein Teilnehmer per Skype zugeschaltet.

werden. Der Hintergrund für die Kategorien mit dem potenziellen Zusatz "überwiegend" ist der, dass eine eindeutige Zuordnung zu einem der drei Basiswerte bei einigen Interaktionen schwierig ist, unter anderem z.B. bei Prüfungen mit verbal nicht oder fast nicht beteiligten Beisitzern, Maptasks mit Experimentleitern, die nur zu Anfang kurz etwas sagen, Interviews mit sporadisch verbal beteiligtem zweitem Interviewer oder Angehörigem der/s Interviewten, Verkaufsgesprächen im öffentlichen Raum mit potenzieller Erweiterung des Kreises verbal Beteiligter etc. Hier wird die Konstellation ausgewählt, die bezüglich der konkreten Beteiligungsstruktur im Gespräch dominiert.

Hierbei ist zu beachten, dass sich die angesetzten Werte nicht unmittelbar auf die (meist theoretisch angereicherten) Konzepte der dyadischen und triadischen Konstellationen übertragen lassen, da sie sich nicht auf Gesprächsparteien, sondern immer auf konkrete einzelne Teilnehmer beziehen (vgl. z.B. eine triadische Konstellation im pädiatrischen Arzt-Patienten-Gespräch mit Arzt, Kind und zwei Elternteilen, die in FOLK als "überwiegend Drei-Personen-Interaktion" oder "Mehr-Personen-Interaktion" gelabelt würden, je nach verbaler Adressierung und Beteiligung des Kindes und der beiden Eltern). Wie bei anderen Parametern verstehen sich die Angaben primär als Such- und Strukturierungshilfe für eingehendere individuelle Analysen.

Die Verteilung in FOLK sieht aktuell, diesmal differenziert nach Gesprächsdauer und Gesprächsanzahl, folgendermaßen aus:



Grafik 5: Teilnehmer-Konstellation

Mit diesem Punkt verbunden ist auch die Frage, ob eine Forscherbeteiligung vorliegt, angegeben als "nicht beteiligt" (="nicht vorhanden"), "nicht verbal beteiligt" oder "verbal beteiligt". Diese Angabe ist zwar im Metadatenschema erfasst, wurde aber nicht immer (korrekt) dokumentiert bzw. berücksichtigt, teils ist die Angabe unklar (eine entsprechende Modifikation wurde daher auch im Formular vorgenommen). Zusätzlich wird auch die Rolle des Forschers als entweder "beobachtend" (und eventuell trotzdem teilweise verbal beteiligt) oder "teilnehmend" als authentischer, verbal beteiligter Gesprächsteilnehmer charakterisiert. Die gleichen Prinzipien gelten für eventuell teilnehmende Techniker.

3.5. Publikum

Dieser Parameter kann prinzipiell binär mit "ja" oder "nein" kodiert werden und tauchte bislang noch nicht im Metadatenschema und -formular auf, ist aber für das Interaktionssetting von Relevanz und muss somit ergänzt werden. Während ein Publikum insgesamt passiv-rezeptive Teilnehmer umfasst, können diese zusätzlich mit (gesteuertem, temporärem) Rederecht ausgestattet sein (z.B. Fragerunden bei Podiumsdiskussionen). Zudem kann es Mehr-Personen-Interaktionen mit gestuftem Publikum geben, z.B. Talkgäste vs. Studiogäste vs. TV-Zuschauer. Diese Angaben können jeweils als offene Anmerkungen ergänzt werden, lassen sich häufig aber auch aus den weiteren Angaben zu medialer Realisierung und Teilnehmerrollen erschließen. In FOLK machen Interaktionen mit Publikum bislang nur einen sehr geringen Anteil aus.

3.6. Vertrautheit

Der Grad der Vertrautheit der Teilnehmer wird mit den Werten "unbekannt" (Erstkontakt), "bekannt", "vertraut" (Freundschaft, Familienmitgliedschaft) oder "divers/gemischt" kategorisiert. Häufig wird bei den privaten Gesprächen der Wert "vertraut" vergeben werden, bei den institutionellen eher "bekannt" oder "unbekannt" und letztgenannter Wert häufiger auch bei den öffentlichen Interaktionen. Problematisch ist eine eindeutige Zuordnung, wenn komplexe Konstellationen innerhalb der Gespräche vorhanden sind (Auflistung mehrerer Werte) und/oder der Vertrautheitsgrad nicht präzise dokumentiert oder klar erkennbar ist bzw. graduell irgendwo zwischen "bekannt" und "vertraut" liegt. Für eine nachträgliche Zuordnung können die Werte des Parameters "Soziale Rollen und Beziehungen" (13) zusätzliche Präzision bringen. Dieser Parameter steht also mit den Sprecherinformationen (vgl. Abschnitt 4) in Verbindung, auch wenn er bei der Kategorisierung global für das jeweilige Gesprächsereignis erfasst wird.

3.7. Soziale Rollen und Beziehungen

Soziale Rollen sind nach Deppermann/Hartung (2011:425) zu definieren als "die Beteiligungsrechte und -pflichten der Teilnehmer gemäß ihrer offiziellen Identitäten, die konstitutiv für ihre Zulassung zu einem privaten Kontext sind bzw. aufgrund derer sie an einer institutionellen Interaktion teilnehmen; nicht gemeint sind Beteiligungsrollen, die erst durch Gesprächsaktivitäten hergestellt werden, wie z.B. Klagender-Tröster, Erzähler-Zuhörer, Freund-Freund in einem Arzt-Patient-Gespräch". Als Beispiele für soziale Rollen nennen die Autoren: "Mutter, Kind in Familientischgesprächen; Cliquenmitglied in Jugendkommunikation; Richter-Angeklagter-Zeuge-Protokollant-Rechtsanwalt in einer Gerichtsverhandlung".

Bisher wurde dieser Parameter mit einer offenen, sehr heterogenen Werteliste geführt, deren Werte sich zwischen eher interaktionsunabhängigen (institutionell oder privat geprägten) Rollen wie z.B. "Freund" oder "Polizeibeamter" und sehr interaktionsspezifischen Gesprächsrollen wie z.B. "Experte auf der Kritikerseite" (in den Schlichtungsinteraktionen) bewegen. Für die zukünftige Dokumentation und rückwirkende Vereinheitlichung werden ereignis(typ)bezogene, auf einer an-

gemessenen Abstraktionsebene gewählte Gesprächsrollenbezeichnungen angestrebt, wobei Mehrfachbezeichnungen in unklaren Fällen auch zugelassen werden. Bei privaten Gesprächen wird im Projekt an dieser Stelle in der Regel die unspezifische Bezeichnung "Gesprächsteilnehmer(in)" bevorzugt; Informationen zu familiären und/oder freundschaftlichen Beziehungen wie etwa "Mutter", "Partner(in)" etc. sollen auf die Angaben zu den sozialen Beziehungen der Teilnehmer untereinander beschränkt werden (siehe unten). Bei den institutionellen Gesprächen ergeben sich einige Überschneidungen zwischen Rollenbezeichnung und Berufen der Teilnehmer – diese sind auch zulässig; zusätzlich soll aber bei der Rolle der spezifische Gesprächstyp berücksichtigt werden (z.B. "Mitarbeiter" oder "Teilnehmer" bei einem Arbeitsmeeting, "Angestellter" dagegen als Berufsbezeichnung).

Die (sozialen) Beziehungen der Gesprächsteilnehmer untereinander werden in der bisherigen Praxis in einem weiteren Punkt – als rein sprecherbezogener, interaktionsunabhängiger Parameter, also eigentlich in Abschnitt 4 zu verorten – erfasst, wenn auch sehr heterogen und angesichts der Komplexität des Parameters bislang nicht ausreichend systematisiert: Pro Ereignis werden die Beziehungen der anderen Gesprächsbeteiligen zum jeweiligen Sprecher aufgeführt und dadurch indirekt dessen eigene Rollen konstituiert. "Mutter", "Bruder" etc. verweist z.B. auf die aktuelle Sprecherin selbst als "Tochter", "Schwester" etc. Inkonsistenzen, Unvollständigkeit und Idiosynkrasien müssen an dieser Stelle wohl in Kauf genommen werden. Wie oben zu den Rollen aufgeführt sind präzisere Angaben zu den Beziehungen besonders bei den privaten Interaktionen dennoch wichtig und werden bei der Dokumentation neuer Daten sowie der Vereinheitlichung bisher vorhandener Daten berücksichtigt.

3.8. Empraktischer Bezug

Bei empraktischen Gesprächen steht, so die Definition bei Deppermann/Hartung (2011:425), das Sprechen entweder "nicht im Fokus der Aktivität […], sondern [spielt] nur eine ergänzende, organisierende oder komplementäre Rolle […], oder […] verbale und nichtverbale, gegenständliche Handlungen [sind] eng miteinander verwoben […]". Der Parameter kann binär mit "ja" oder "nein" kodiert werden, allerdings erweist sich der Grad des empraktischen Bezugs bei einigen Gesprächen als graduell und nicht klar entscheidbar. Die Probleme, die sich hier zeigen, sind folgende: Viele Gespräche schließen in irgendeiner Form auch mehr oder weniger fokale gegenständliche Handlungen ein oder das Sprechen kann wesentlich für die Koordination und Ausführung von Tätigkeiten sein, die dennoch nicht unbedingt empraktisch im prototypischen Sinn sind. Es gilt also, den Parameter und seine Werte in sinnvoller Weise so eng und präzise zu definieren, dass die Interaktionen möglichst eindeutig zugeordnet werden können. Eine zusätzliche Differenzierung bietet zumindest bei privaten Interaktionen der Parameter "Aktivität"; auch die Angaben zu Themen und Verlauf bieten eine weitere Orientierung. Somit ist es einfacher, nur solche Interaktionen als empraktisch zu kodieren, bei denen das Sprechen konkret auf physische Bewegung, Koordination und/oder gegenständliche Handlungen (im Sinne von Objektmanipulation u.Ä.) ausgerichtet bzw. mit diesen verwoben ist. Eine Zwischenkategorie ("divers/unklar/gemischt") kann alternativ dennoch beibehalten werden. Nach der obigen Defini-

tion machen nicht-empraktische Interaktionen in FOLK aktuell den überwiegenden Teil (knapp 74%) aus.

3.9. Sprachen

Angaben zu den in der Interaktion konkret verwendeten Sprachen erlauben Aussagen zur Rolle von Mehrsprachigkeit in den Gesprächen. Die einzelnen Sprachen werden als Werte in einer offenen Liste aufgeführt. Da FOLK sich per definitionem vorrangig auf das gesprochene Deutsch fokussiert, wird als erster (oder auch einziger) Wert immer "Deutsch" angegeben. Mehrsprachigkeit spielt in den Interaktionen (noch) eine sehr geringe Rolle. Für bestimmte Kontexte, wie Lehr-Lern-Interaktionen, z.B. auch im Rahmen von Sprachtandems, kann dieser Parameter allerdings zunehmend interessant und wichtig werden. Weitere Sprachen werden generell nur dann aufgeführt, wenn zumindest ein kurzer Austausch in dieser Sprache stattfindet (z.B. passagenweise Code-Switches auf Türkisch im Interview mit türkischstämmigen Auswanderern, Essensbestellung im Restaurant auf Englisch etc.). Einzelne Wortnennungen oder die Verwendung von Anglizismen fallen nicht darunter.

4. Stratifikationsleitende Sprecherparameter

Die stratifikationsleitenden demographischen, in der Systematik sekundären Parameter sind Alter (sprechereignisbezogen) bzw. Geburtsjahr (rein personenbezogen), Geschlecht, Bildungsabschluss (bzw. -grad) und Aufenthaltsregionen (bzw. in Bezug auf die Interaktionen: Aufnahmeregionen). Ergänzende Parameter sind Berufe, Gesprächsrollen (sprechereignisbezogen, vgl. Abschnitt 3), soziale Beziehungen (sprecherbezogen, vgl. ebenfalls Abschnitt 3) und Sprachkenntnisse. Wie einleitend gesagt soll beim Ausbau vorrangig versucht werden, zu allen Ausprägungen der Variablen überhaupt Daten ins Korpus aufzunehmen. Repräsentativitätsbestrebungen im Sinne einer Annäherung an die Verteilung dieser Werte in der deutschsprachigen Bevölkerung, wie sie im Zensus dargestellt werden, sind praktisch nicht umsetzbar, vgl. zu Überlegungen dazu aber das Arbeitspapier von Winterscheid (2016) und die Ausführungen zur relativen regionalen Verteilung auf Süd – Nord – West – Ost in Abschnitt 4.4.

Wie bei den Interaktionsparametern folgen auch hier jeweils einige Ausführungen zu den einzelnen Parametern.

4.1. Geschlecht

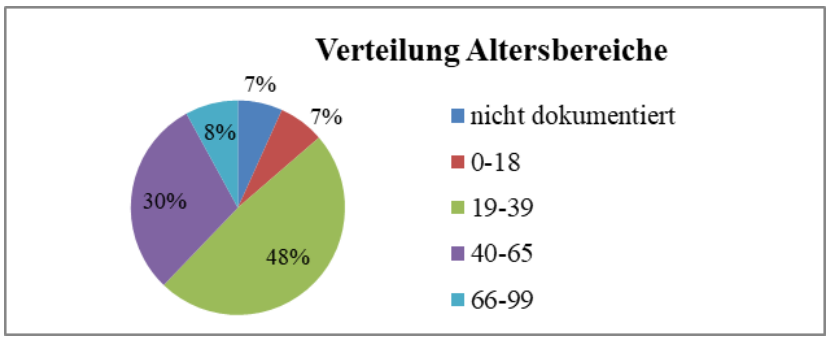
Männliche und weibliche Sprecher sind in FOLK insgesamt zu annähernd gleichen Teilen (48% zu 51%, 1% anderes oder nicht dokumentiert) vertreten. Bei zukünftigen Erhebungen und Übernahmen muss das Verhältnis innerhalb der Oberkategorien primärer Parameter jeweils überprüft werden.

4.2. Alter

Das Alter zum Aufnahmezeitpunkt als ereignisbezogener Sprecherparameter wird in der Regel über das Metadatenformular erfasst. Das Geburtsjahr als ereignisunabhängiger Parameter wurde bislang aus der Altersangabe und der Angabe zum Aufnahmejahr des Gesprächsereignisses errechnet. Zukünftig wird im Dienste größerer Genauigkeit und verringerten Aufwands aber auch dieses direkt über das Metadatenformular abgefragt.

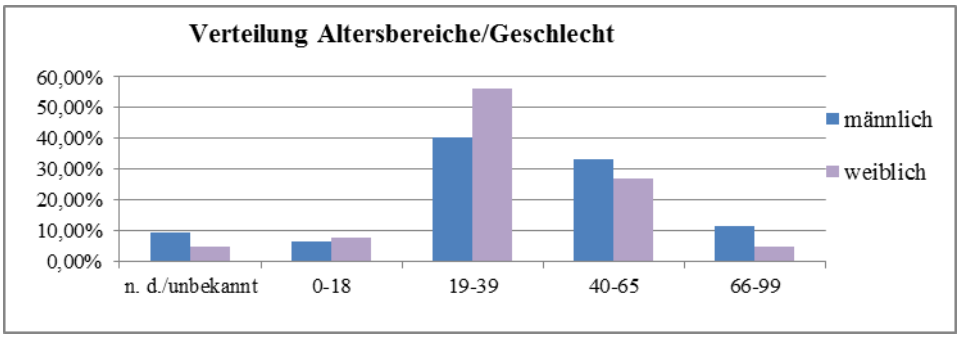
Für die Stratifikation ist zusätzlich zur Dokumentation der präzisen Angabe eine Kategorisierung nach bestimmten Altersstufen wichtig. Möglich wäre eine noch relativ genaue Einteilung in Zehnerschritte, mit einer leicht abweichenden Zäsur bei Volljährigkeit: 0-10; 11-18; 19-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90-99.

Bei dieser differenzierten Einteilung werden die bestehenden Ungleichgewichte, vor allem das aktuelle Übergewicht der 19-29-Jährigen in FOLK, sehr deutlich. Eine gröbere Einteilung und damit Reduktion der Werte bei den Parametern ist für die demographische Stratifikation aber der praktikablere und somit favorisierte Ansatz. Daher wird eine alternative Einteilung in vier Altersgruppen mit folgenden Wertbereichen angesetzt: 0-18; 19-39; 40-65; 66-99. Die entsprechende Verteilung der FOLK-Daten sieht für den Stand 2018 folgendermaßen aus:



Grafik 6: Altersbereiche

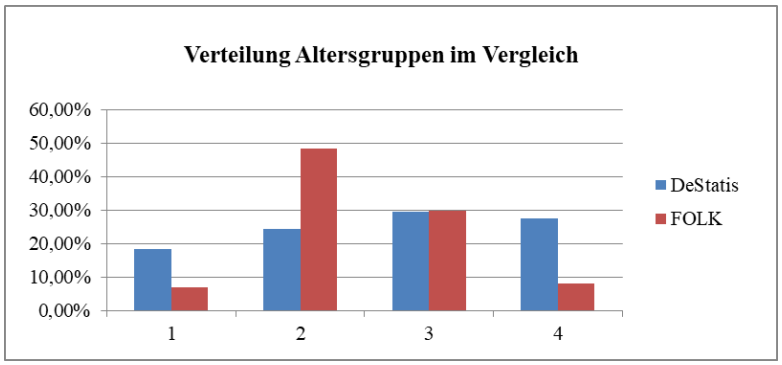
Auch Verteilungsungleichgewichte zwischen männlichen und weiblichen Sprechern zeigen sich hier:



Grafik 7: Altersbereiche/Geschlecht

Die Viereinteilung korreliert recht gut mit gesellschaftlich etablierten Lebensphasen: Kinder/Heranwachsende – junge Erwachsene – Erwachsene – Senioren. Mit dieser Struktur könnte langfristig auch eine Annäherung der Säulenhöhen erreicht werden, was bei feineren Einteilungen utopisch ist. Eine vollständige Ausgewogenheit der Altersstufen ist allerdings nicht nötig und auch nicht gewünscht: Da FOLK Aufnahmen von vollkompetenten Sprechern anstrebt, sind Kinder von 0-6 eine Sprechergruppe, die nur marginal berücksichtigt wird.

Auch bei Auswertungen des statistischen Bundesamtes[18] findet sich eine ähnliche Einteilung und Darstellung der Altersgruppen der Bundesbevölkerung: unter 20, 20-40, 40-60, 60-80, über 80. Während die ersten drei Gruppen fast genau unserer eigenen Einteilung entsprechen (bis auf die bei FOLK nicht unwichtige, abweichende Zuordnung von 18-20-Jährigen), lassen sich die letzten beiden zu unserer vierten Gruppe zusammenfassen. Im Jahr 2016 war die Verteilung im deutschen Bundesgebiet wie folgt: Die erste Gruppe machte 18,4% aus, die zweite 24,5%, die dritte 29,4% und die vierte insgesamt 27,6%. Den FOLK-Daten an die Seite gestellt (mit angepassten Wertbereichen) zeigen sich die bekannten Ungleichgewichte nochmals deutlich:



Grafik 8: Altersgruppenvergleich

4.3. Bildungsabschlüsse

Die bislang für Bildungsabschlüsse erfassten Werte in den FOLK-Sprecherdokumenten bilden eine sehr heterogene Liste. Der Anteil nicht dokumentierter Abschlüsse von Sprechern bei den bisherigen Daten ist außerdem relativ hoch. Mögliche Vereinheitlichungen und eine stärkere Kontrolle der Werte wären etwa durch die Verwendung einer Matrix aus soziologischen Studien (mit *multiple choice*-Auswahl) bei der Metadatenabfrage zu erreichen. Während dort verwendete Schemata für FOLK aber letztlich deutlich zu detailliert und zu aufwändig für die Erhebung und Dokumentation sind, stellt die in solchen Studien häufiger aufgeführte Kategorie "aktuell angestrebter Abschluss" eine sinnvolle künftige Erweiterung des Metadatenschemas dar, welche aktuell implementiert wird.

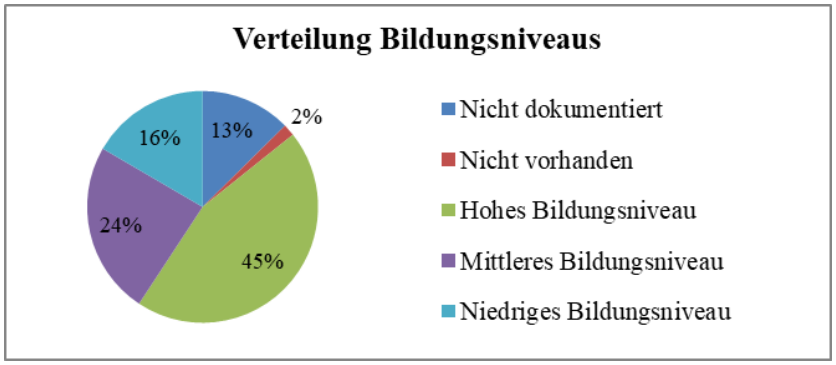
Darüber hinaus werden sowohl Bildungsabschlüsse als auch Berufe (erlernte und aktuell ausgeübte) weiterhin als offene Angaben erfasst. Wichtig für die Stra-

[18] https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Bevoelkerungsstand/Tabellen_/lrbev01.html.

tifikation ist aber, den schulischen und/oder berufsbildenden Abschluss einer klar definierten (wenn auch dadurch stark zusammenfassenden) Bildungsstufe zuordnen zu können. Hierfür bietet sich beim deutschen Ausbildungssystem eine Dreiereinteilung an:[19]

1. Hohes Bildungsniveau / "Tertiärbereich":
2. Mittleres Bildungsniveau / "Sekundarstufe II":
3. Niedriges Bildungsniveau / "Sekundarstufe I" und "Primarstufe":

Im Überblick sieht die Verteilung für die drei Bereiche in FOLK aktuell wie folgt aus (vgl. für die entsprechenden, bisher in FOLK dokumentierten Bildungsabschlüsse die Auflistung im Anhang):



Grafik 9: Bildungsniveaus

Auch die Werte der Aufstellungen des Statistischen Bundesamtes zur Bildung lassen sich, wenn auch nicht explizit so angegeben, weitgehend auf das dreiteilige System abbilden. In der aktuellsten Erhebung[20] werden die allgemeine schulische und die berufliche Bildung allerdings getrennt dargestellt und jeweils nur für Personen ab 15 Jahren gezählt. Zum Tertiärbereich zu zählen sind hier die Angaben Fachhochschul- oder Hochschulreife, Fachschulabschluss, Bachelor, Master, Diplom, Promotion. Zur Sekundarstufe II zählen Lehre, Berufsausbildung, Abschluss in polytechnischer Oberschule. Zum Bereich Sekundarstufe I und Primarstufe zählen Haupt- oder Volksschulabschluss, Realschul- oder gleichwertiger Abschluss, kein allgemeiner Schulabschluss bzw. noch in schulischer/beruflicher Ausbildung.

Sollte jemand (noch) keinen Bildungsabschluss erworben haben (z.B. Kindergartenkinder), so wird in FOLK der Wert "Nicht vorhanden" angegeben. Vor al-

[19] Vgl. dazu folgende Online-Quellen:
http://www.schulsystem.info/schulabschluesse.html
https://www.bildungsserver.de/Gesamtueberblick-zum-deutschen-Bildungssystem-506-de.html
https://www.kmk.org/fileadmin/Dateien/pdf/Dokumentation/dt-2015.pdf
https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Bildungsstand/Tabellen/Bildungsabschluss.html.

[20] https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002167014.pdf?__blob=publicationFile.

lem der Bereich der niedrigen Bildungsstufen zeigt sich im Vergleich in FOLK als deutlich unterrepräsentiert.[21]

4.4. Aufenthaltsregionen

Die Aufenthaltsregionen der Sprecher wurden bislang ebenso wie die Aufnahmeregionen der Sprechereignisse in Anlehnung an die Einteilung nach Wiesinger [1983] (2008) in 15 Sprachregionen aufgeteilt; zusätzlich wird der Wert "außerhalb deutschsprachigen Kerngebietes" aufgenommen.[22]

Abbildung 1: Karte 47.4, Die Gliederung der deutschen Dialekte (in den ersten Jahrzenten des 20. Jhs.)

Wie bei den anderen demographischen Angaben sind auch die Werte zu Aufenthaltsregionen in FOLK relativ häufig "nicht dokumentiert". Zudem werden in der Regel sämtliche Aufenthaltsregionen der Person aufgeführt, meist zwar mit Angaben zur Dauer, aber ohne weitere klare Kennzeichnung zur sprachlichen Prägung. Um zu diesem Parameter aussagekräftigere und statistisch überhaupt verwertbare Informationen zu bekommen, wird im Metadatenformular zukünftig (nach der Angabe zum Land, in der Regel Deutschland) die sprachlich prägendste

[21] Dies ergibt sich aus der ursprünglichen, primären Erhebungsstrategie im Projekt, über TeilnehmerInnen an universitären Seminaren zu Gesprächsaufnahmen zu kommen, welche folglich im studentischen Milieu verortet sind.

[22] Aufnahmen aus Österreich (Bairisch), der Schweiz (Alemannisch) und Belgien (Moselfränkisch/Ripuarisch) können in das System integriert werden.

Region als zuerst anzugebende abgefragt.[23] Im Anschluss können weitere Regionen (wie bisher mit zusätzlichen, optionalen Angaben zur Dauer) aufgeführt werden. Für die bisherigen Daten werden diese hierarchisierten Informationen so weit wie möglich aus den Angaben zur Aufenthaltsdauer rekonstruiert und nachgetragen. Einstweilen ist die Aufnahmeregion als Ereignisparameter, sozusagen als Hilfskategorie, noch der bessere Indikator[24] für die regionale Verteilung, obwohl Aufnahmeregion und prägendste Aufenthaltsregion der Sprecher selbstverständlich nicht immer, aber bislang doch noch häufig, zusammenfallen.

Zur Verschlankung der Wert-Attribut-Kombinationen für die Ausbau-Planung trägt eine zusätzliche Abbildung der bisherigen 15 Sprachregionen auf nur sechs Großregionen bei, welche (unter Beibehaltung beider Systematiken im Metadatenschema) sowohl für die Aufnahme- als auch die Aufenthaltsregionen angewendet werden kann: Nordwest – Nordost – Mittelwest – Mittelost – Südwest – Südost (vgl. den Vorschlag bei Winterscheid 2016 mit Verweis auf die Darstellung bei Ammon et al. 2004, XLIII).

Die Visualisierung bei Lameli (2008), der sich ebenfalls an Wiesinger orientiert, erlaubt eine klare Zuteilung:

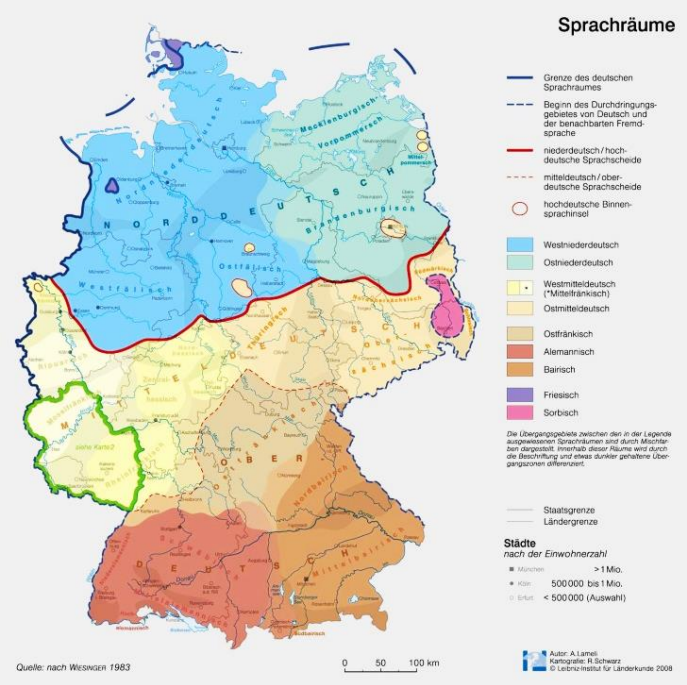


Abbildung 2: Lameli (2008)

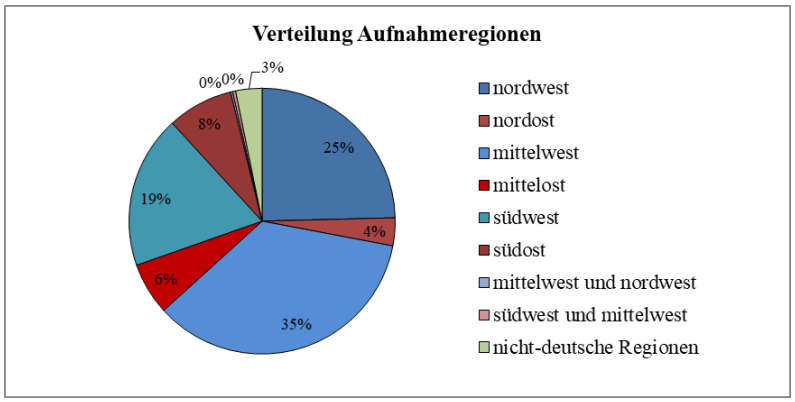
Es ergibt sich folgende, projektintern definierte Zuordnung:

- Nordwest: Nordniederdeutsch, Westfälisch, Ostfälisch;
- Nordost: Brandenburgisch, Mecklenburg-Vorpommerisch;
- Mittelwest: Rheinfränkisch, Moselfränkisch, Ripuarisch, Hessisch;

[23] Vgl. z.B. auch das in Abschnitt 2 angesprochene tschechische ORTOFON-Korpus, in dem die Hauptaufenthaltsregion bis zum 15. Lebensjahr erhoben wurde.

[24] Pro Interaktion steht hier i.d.R. nur ein Wert, nur bei den Telefongesprächen sind auch zwei Angaben möglich.

- Mittelost: (Ober-)Sächsisch, Thüringisch;
- Südwest: Alemannisch, Schwäbisch;
- Südost: Ostfränkisch; Bairisch.



Grafik 10: Aufnahmeregionen

Die Grafik zur Verteilung der Aufnahmeregionen (als Hilfsannäherung an die Herkunftsregionen der Sprecher, siehe oben) zeigt, dass Gesprächsdaten aus den östlichen Gebieten Deutschlands (in Rottönen) vorerst am dringendsten gebraucht werden, um sich ausgewogeneren Verhältnissen anzunähern. Allerdings kann hier ein zumindest grober Bezug zu den Einwohnerzahlen im Vergleich West – Ost – Nord – Süd gesetzt werden, da erstens der Westen größer ist als der Osten und zweitens der Süden dichter besiedelt als der Norden.[25]

Das Verhältnis von Einwohnern in West- zu Ostdeutschland ist gemäß dieser Einteilung und den entsprechenden Zahlen (vgl. im Anhang) insgesamt fast zwei Drittel zu einem Drittel (64,75% zu 35,25%). Auch wenn der Anspruch an Repräsentativität insgesamt und bezüglich der anderen demographischen Parameter (Alter und Bildung), wie einleitend dargelegt, keine Priorität beim Korpusaufbau darstellt, kann dieses Ungleichgewicht für die FOLK-Daten zumindest berücksichtigt werden – hier machen die westdeutschen Sprecher aktuell insgesamt 66% aus, die ostdeutschen aber nur 17% (dazu kommen noch die nicht dokumentierten Regionen).

4.5. Sprachkenntnisse

Bislang wurden alle Sprachen der Sprecher erfasst und zusätzlich der Kenntnisgrad abgefragt. Das Schema sieht 5 Stufen vor, welchen die folgenden Angaben von Sprechern zugeordnet wurden:

---

[25] Vgl. die Übersichten des Statistischen Bundesamtes zu den Einwohnerzahlen in den einzelnen Bundesländern für 2016,
https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/ Administrativ/Aktuell/02Bundeslaender.html.
Vgl. für eine Übersicht über die Zuteilung der 15 bzw. 6 Sprachregionen auf die Bundesländer die entsprechende Tabelle im Anhang, ebenso für eine Auflistung der Einwohnerzahlen pro Bundesland und der jeweiligen relativen Anteile des gesamten Bundesgebietes in Bezug auf das grobe (anteilige) Mapping auf die sechs Sprachregionen.

1 – CPE (Certificate of Proficiency in English), verhandlungssicher

2 – Konversationssicher

3 – Weniger gut/Mittel/Mittelmäßig/Okay/in Ordnung

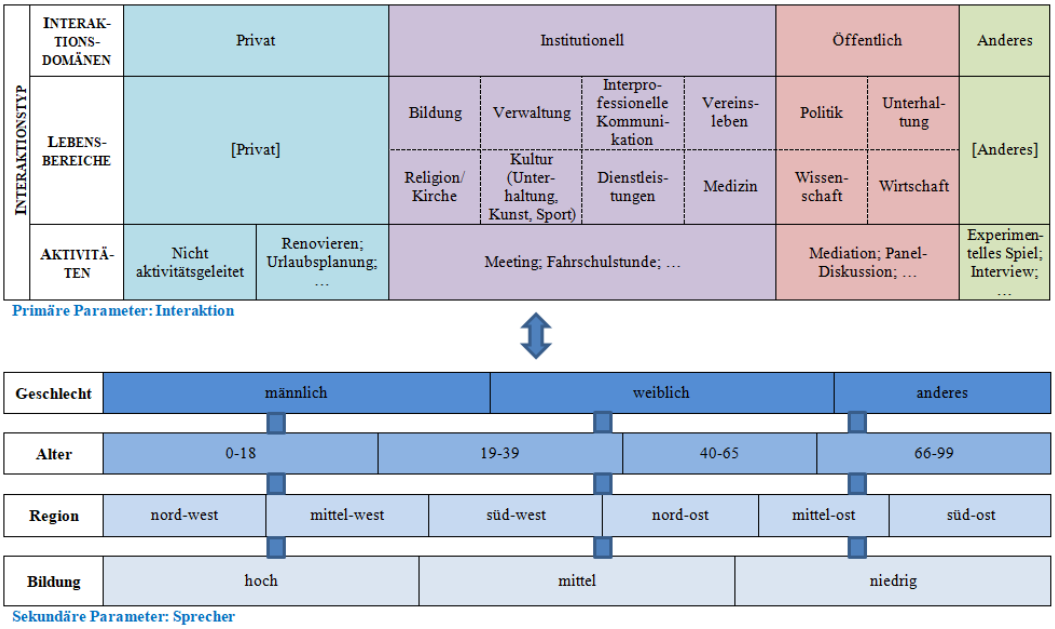
4 – Grundkenntnisse/Wenig/Schlecht/Etwas

5 – Verlernt

Die Angaben sind aber oft unvollständig oder mit sehr individuellem und vagem Vokabular ausgefüllt (z.B. "ausbaufähig", "naja", "Schulkenntnisse"). Der Kenntnisgrad wird künftig als wenig brauchbare Information daher gänzlich aus dem Schema genommen. Stattdessen wird deutlicher gekennzeichnet, was die Erstsprachen eines Sprechers sind und welches die Zweitsprachen(n), bzw. noch einfacher, ob Deutsch Erstsprache ist oder nicht.

5. Ausbauplan: Überblick, Ergänzungen, Strategien

Abschließend geben die folgenden Übersichten zunächst noch einmal einen zusammenfassenden Überblick über die in den vorangegangenen Abschnitten ausgeführte Stratifikationssystematik, die erste als schematische Darstellung, die zweite als vollständige tabellarische Auflistung aller primären und sekundären Parameter und ihrer möglichen Werte.



Grafik 11: Schema Stratifikation

		Parameter[26]	**Werte**
PRIMÄR	**STRATIFIKATIONSLEITEND**	Interaktionsdomäne	Privat; Institutionell; Öffentlich; Sonstiges
		Lebensbereich	Privat: Privat (keine Spezifizierung); Öffentlich: Politik; Unterhaltung; Wissenschaft; Wirtschaft; Institutionell: Bildung; Behörden; Inter-professionelle Kommunikation; Vereinsleben und Selbstverwaltung; Religion/Kirche; Kunst/ Unterhaltung/Sport; Dienstleistungen; Medizin/ Gesundheitswesen; Sonstiges: Sonstiges (keine Spezifizierung)
		Aktivitäten	Privat: aktivitätsgeleitet: offene Liste; nicht aktivitätsgeleitet; Sonstiges: aktivitätsgeleitet: Maptask; Interview ((sprach-)biographisch; ethnographisch); Öffentlich: aktivitätsgeleitet: offene Liste; Institutionell: aktivitätsgeleitet: offene Liste
	ERGÄNZEND	Medium/mediale Realisierung	*face-to-face*; Telefon; massenmedial übertragen (+ Mischfälle)
		TN-Zahl + Konstellation	exakte Angabe + Einteilung Zwei-Personen-Interaktion; Drei-Personen-Interaktion; Mehr-Personen-Interaktion; überwiegende Konstellation (ggf. m. Präzision bzgl. Forscherbeteiligung)
		Publikum	ja; nein (+ offene Angabe zu gestuftem Publikum)
		Vertrautheit	unbekannt; bekannt; vertraut; divers/gemischt
		soziale Rollen und Beziehungen	offene Angaben
		Empraktischer Bezug	ja; nein; divers/unklar/gemischt
		Sprache(n) der Interaktion	offene Angabe(n)
SEKUNDÄR	**STRATIFIKATIONSLEITEND**	Geschlecht (personenbezogen)	männlich; weiblich; anderes
		Alter (interaktionsbezogen)	exakte Angabe + Vierer-Einteilung (0-18; 19-39; 40-65; 66-99)
		Aufenthalts-/ Aufnahmeregion (personen-/ interaktionsbezogen)	Offene Angabe + Zuordnung zur Einteilung nach Wiesinger (15 Regionen) + Einteilung nach Lameli (sechs Regionen)
		höchster + derzeit angestrebter Bildungsabschluss (personenbezogen)	Offene Angabe – Zuordnung zur Einteilung nach drei Bildungsstufen
	ERGÄNZEND	Beruf (personenbezogen)	offene Angabe
		Sprachkenntnisse (personenbezogen)	Erstsprache (Deutsch ja/nein); weitere Sprachen

Tabelle 1: Stratifikationsparameter

[26] Über die hier aufgeführten Parameter hinaus werden projektintern im XML-Schema für die Metadaten noch einige weitere Informationen dokumentiert, v.a. zu organisatorischen und technischen Modalitäten der Aufnahme.

Alle aufgeführten Sprecherparameter bzw. die Verteilung ihrer Werte in den FOLK-Daten müssen für ein vollständiges Bild über die quantitativen Verhältnisse schließlich mit den Interaktionsparametern, zunächst vor allem den übergeordneten Domänen und dann den Lebensbereichswerten, gekreuzt werden. Somit können für den zukünftigen Ausbau ein schrittweiser Ausgleich von Ungleichgewichten bzw. ein "Auffüllen" leerer Wert-Ausprägungen angestrebt werden.

Im Ausbauplan (2017) finden sich Empfehlungen, die hier nochmals erläutert und durch weitere Überlegungen ergänzt werden. Eine "breite", also auf eine große Abdeckung hinsichtlich sekundärer Stratifikationsparameter (vor allem Alter, Bildungsgrad und regionale Herkunft) zielende Erhebung ist aufgrund der einleitend diskutierten Ausgewogenheitsproblematik zunächst nur für einen privaten Gesprächstyp konkret geplant. Dieser ist erstens mit verhältnismäßig wenig Aufwand (bezüglich Rekrutierung, Komplexität des Settings und Zugänglichkeit für ForscherInnen und TeilnehmerInnen) zu erheben und zählt zweitens zu den basalen, hochfrequenten Kommunikationsroutinen des alltäglichen Lebens. Es handelt sich um private (auf den auditiven Kanal beschränkte) Telefongespräche[27] zwischen zwei Personen. Mit Hilfe einer am IDS neu installierten Anlage können zukünftig solche Telefonate von TeilnehmerInnen, die sich zuvor für eine Aufnahme gemeldet und ihr Einverständnis gegeben haben, zeitlich und räumlich flexibel mitgeschnitten und unmittelbar danach technisch aufbereitet werden. Der Anspruch der breiten Streuung ist allerdings selbst für diesen begrenzten Bereich recht hoch: Wenn für sechs Sprachregionen je vier Altersstufen und drei Bildungsgrade (zusätzlich auch eine annähernd ausgewogene Verteilung über beide Geschlechter) berücksichtigt werden sollen, ergeben sich allein 72 Kombinationen. Das Konzept muss daher eher als Ideal zur Orientierung betrachtet werden. Für eine weitere parallele Erhebungsinitiative können zu einem späteren Zeitpunkt familiäre Tischgespräche (als Mehr-Personen-Interaktionen) in den Blick genommen werden.

Beide Gesprächstypen bieten bei einer erfolgreichen Erhebung eine gute Möglichkeit, bezüglich der sekundären Parameter ausgewogene Subkorpora innerhalb von FOLK anzubieten, die außerdem zusammengenommen grundlegende Interaktionsparameter innerhalb der Kategorie "Privat" variieren, vor allem Medium (*face-to-face* vs. vermittelt über Telefon) und Teilnehmerzahl (Zwei- vs. Mehr-Personen-Gespräch). Während bei den Tischgesprächen zunehmend Videoaufnahmen angestrebt werden, sind die genuin auf den Audiokanal beschränkten Telefongespräche bezüglich des visuellen Kanals bzw. Multimodalität (weitgehend) indifferent und somit unproblematisch.

Zusätzlich zum Fokus auf diesen beiden Gesprächstypen sind weitere "opportunistische" oder auch geplante Erhebungen und aus Kooperationen erwachsende Datenübernahmen anderer Gesprächstypen jederzeit möglich und erwünscht. Mittels einer im Herbst 2018 initiierten Werbe-Initiative werden WissenschaftlerInnen, aber auch wissenschaftliche Laien durch Aufrufe über Mailinglisten, Facebook, Twitter, weitere Webseiten und Poster dazu angeregt, interessante Interaktionstypen für FOLK zu erheben und/oder bereits existierende Daten weiterzugeben. Bei Auswahl und Übernahme wird darauf geachtet, nicht zu große Datenmengen eines Typs aufzunehmen und keine zu großen Übergewichte der Redean-

[27] In Anlehnung an die Systematik des in Abschnitt 2 vorgestellten GOS-Korpus wäre es möglich, später auch Aufnahmen von beruflichen Telefongesprächen zu integrieren.

teile einzelner Sprecher zu erhalten. Auch die sonstigen Erkenntnisse über möglichst auszugleichende Ungleichgewichte bezüglich Alter, regionaler Herkunft und Bildung der Sprecher werden weiterhin berücksichtigt (vgl. Winterscheid 2016). Als geeignet bei der Übernahme größerer Sammlungen erweist sich generell ein Vorgehen, das bei der Übernahme des GeWiss-Korpus angewendet wurde: Das Gesamtkorpus wird in das mit der DGD verbundene Archiv für Gesprochenes Deutsch übernommen und eine nach den Stratifikationsparametern als sinnvoll bewertete Auswahl davon in FOLK integriert.

Tests bezüglich geeigneter Videotechnik und -einrichtung für komplexere Settings sollen für empraktische Interaktionstypen mit primär handlungsbegleitendem Sprechen angestrebt werden, wie z.B. beim Aufbau von IKEA-Möbeln oder weiteren Koch- und Backinteraktionen. Weiterhin wünschenswert und denkbar sind Interaktionen im Reisebüro und in weiteren Dienstleistungssektoren, behördliche und Vereins-Interaktionen, Sprechen mit Tieren, massenmedial vermittelte Interaktionen wie Talkshows.[28] Insgesamt machen die Videoaufnahmen in FOLK bislang nur 30% aus.

Hinsichtlich der Grobeinteilung in Interaktionsdomänen wurde aus Grafik 3 in Abschnitt 3 auch ersichtlich, dass es in FOLK bislang vergleichsweise wenige Daten aus der Kategorie "Öffentliche Kommunikation" gibt und dass die vorhandenen Daten aus sehr wenigen Interaktionen (vor allem Stuttgart 21) bestehen. Zeitweise wird daher voraussichtlich ein Teil der Kapazitäten für einen Ausgleich in diesem Bereich eingesetzt werden, z.B. durch Aufnahmen von mehr und unterschiedlichen Podiumsdiskussionen (in Vorbereitung).

Inwiefern letztlich die Kombination aus einer parametrisierten und einer Gattungssystematik für die Gesamtstratifikation als konsequente Taxonomie in FOLK umsetzbar ist, muss noch praktisch überprüft werden. Festzuhalten bleibt, dass die Systematik zunächst vor allem von den Interaktionsdomänen, Lebensbereichen und Aktivitätsspezifizierungen ausgehend konzipiert wird, anschließend verbunden mit weiteren basalen Parametern wie Teilnehmerkonstellation, Medium, Vertrautheitsgrad etc. Zu einem späteren Zeitpunkt kann eine Ableitung und Beschreibung von spezialisierten Gattungen als komplexe Kombinationen relevanter Merkmalsausprägungen auf den Parameter-Dimensionen erprobt werden. Regelmäßige statistische Auswertungen und entsprechende Anpassungen der Ausbaustrategie für das jeweils nächste Release werden zukünftig eingesetzt, um schrittweise eine größere Ausgewogenheit zu erreichen.

Trotz der zukünftig noch zu bewältigenden Projektaufgaben und Weiterentwicklungen reicht FOLK bereits zum aktuellen Zeitpunkt an den Status eines nationalen Gesprächskorpus heran, welches – auch im Vergleich zu anderen existierenden Datensammlungen – quantitativ wie qualitativ hohe Standards sowohl bei der technischen als auch der inhaltlichen Aufbereitung erfüllt.

[28] Für die diversen Interaktionstypen insgesamt wichtig ist, dass die soziodemographischen Variablen hier natürlich nicht unabhängig vom Gesprächstyp frei variiert und daher selbst unter idealen Erhebungsmöglichkeiten systematische Variationen der Sprechervariablen auch nicht sinnvollerweise angestrebt werden können.

6. Literatur

Ammon, Ulrich / Bickel, Hans / Ebner, Jakob / Esterhammer, Ruth / Gasser, Markus / Hofer, Lorenz / Kellermeier-Rehbein, Birte / Löffler, Heinrich / Mangott, Doris / Moser, Hans / Schläpfer, Robert / Schloßmacher, Michael / Schmidlin, Regula / Vallaster, Günter (Hg.) (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin u.a.: de Gruyter.

Bergmann, Jörg (1987): Klatsch. Zur Sozialform der diskreten Indiskretion. Berlin/New York: de Gruyter.

Biber, Douglas (1993): Representativeness in Corpus Design. In: Literary and Linguistic Computing 8, 4, 243-257.

Brown, Penelope / Fraser, Colin (1979): Speech as a marker of situation. In: Scherer, Klaus R. / Giles, Howard (eds.), Social markers in speech. Cambridge: Cambridge University Press, 33-62.

Crowdy, Steve (1993): Spoken Corpus Design. In: Literary and Linguistic Computing 8, 4, 259-265.

Deppermann, Arnulf / Hartung, Martin (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: Felder, Ekkehard / Müller, Marcus / Vogel, Friedemann (Hg.), Korpuspragmatik. Berlin: de Gruyter, 414-450.

Deppermann, Arnulf / Schmidt, Thomas (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). In: Domke, Christine / Gansel, Christa (Hg.), Korpora in der Linguistik - Perspektiven und Positionen zu Daten und Datenerhebung. V&R unipress, 4-17.

Duranti, Alessandro (1985): Sociocultural Dimensions of Discourse. In: van Dijk, T.A. (eds.), Handbook of Discourse Analysis 1, London: Academic Press, 193-230.

Ehlich, Konrad / Rehbein, Jochen [1980] (2011): Sprache in Institutionen. In: Althaus, Hans P. / Henne, Helmut / Wiegand, Herbert E. (Hg.), Lexikon der Germanistischen Linguistik. 2., vollst. neubearb. u. erw. Aufl. Tübingen: Niemeyer, 338-345.

Fandrych, Christian / Frick, Elena / Hedeland, Hanna / Iliash, Anna / Jettka, Daniel / Meißner, Cordula / Schmidt, Thomas / Wallner, Franziska / Weigert, Kathrin / Westpfahl, Swantje (2016): User, who art thou? User profiling for oral corpus platforms. In: Calzolari, Nicoletta / Choukri, Khalid / Declerck, Thierry / Goggi, Sara / Grobelnik, Marko / Maegaard, Bente / Mariani, Joseph / Mazo, Helene / Moreno, Asuncion / Odijk, Jan (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 280-287.

Günthner, Susanne (1995): Gattungen in der sozialen Praxis. Die Analyse kommunikativer Gattungen als Textsorten mündlicher Kommunikation. In: Deutsche Sprache 25/1, 193-218.

Günthner, Susanne (2000): Vorwurfsaktivitäten in der Alltagsinteraktion. Grammatische, prosodische, rhetorisch-stilistische und interaktive Verfahren bei der Konstitution kommunikativer Muster und Gattungen. Tübingen: de Gruyter.

Günthner, Susanne / Knoblauch, Hubert (1994): 'Forms are the Food of Faith' - Gattungen als Muster kommunikativen Handelns. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 46, 4, 693-723.

Halliday, Michael A. K. / Hasan, Ruqaiya (1989): Language, context and text: Aspects of language in a social-semiotic perspective (2nd ed.). Oxford: Oxford University Press.

Henne, Helmut / Helmut Rehbock (1995): Einführung in die Gesprächsanalyse, 3. durchgesehene und um einen bibliographischen Anhang erweiterte Auflage, Berlin: de Gruyter.

Heritage, John / Clayman, Steven (2010): Talk in action. Interactions, identities, and institutions. Chichester u.a.: Wiley-Blackwell.

Hymes, Dell H. (1968): The ethnography of speaking. In: Fishman, Joshua A. (eds.), Readings in the sociology of language. The Hague, Paris: Mouton, 99-138.

Hymes, Dell H. (1974): Ways of speaking. In: Bauman, Richard / Sherzer, Joel (eds.), Explorations in the ethnography of speaking. Cambridge, 433-451.

Lameli, Alfred (2008): Deutsche Sprachlandschaften. In: Nationalatlas aktuell 9 (08/2008). Leipzig: Leibniz-Institut für Länderkunde (IfL). Nochmals publiziert in: Bode, Volker / Lentz, Sebastian / Tzschaschel Sabine (Hg.) (2011), Deutschland aktuell. Kartenbeiträge zu Wirtschaft, Gesellschaft, Kultur, Politik und Umwelt. Leipzig: Leibniz-Institut für Länderkunde (IfL).

Love, Robbie / Dembry, Claire / Hardie, Andrew / Brezina, Vaclav / McEnery, Tony (2017): The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. In: International Journal of Corpus Linguistics 22, 3, 319-344.

Luckmann, Thomas (1986): Grundformen der gesellschaftlichen Vermittlung des Wissens: Kommunikative Gattungen. In: Friedhelm Neidhardt / Rainer M. Lepsius / Johannes Weiß (Hg.), Kultur und Gesellschaft. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 27. Opladen, 191-211.

Luckmann, Thomas (1988): Kommunikative Gattungen im kommunikativen "Haushalt" einer Gesellschaft. In: Gisela Smolka-Koerdt / Peter M. Spangenberg / Dagmar Tillmann-Bartylla (Hg.), Der Ursprung von Literatur. München: Fink, 279-288.

Merkel, Silke / Schmidt, Thomas (2009): Korpora gesprochener Sprache im Netz – eine Umschau. In: Gesprächsforschung 10, 70-93.

Oostdijk, Nelleke (2002): The Design of the Spoken Dutch Corpus. In: Peters, Pam / Collins, Peter / Smith, Adam (eds.): New Frontiers of Corpus Research. Amsterdam: Rodopi, 105-112.

Schmidt, Thomas (2014a): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: Gesprächsforschung 15, 196-233.

Schmidt, Thomas (2014b): The Database for Spoken German – DGD2. In: Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA), 1251-1457.

Schmidt, Thomas (2014c): The Research and Teaching Corpus of Spoken German – FOLK. In: Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA), 383-387.

Schmidt, Thomas (2016): Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. In: Compilation, transcription, markup and annotation of spoken corpora. In: Kirk, John M. / Andersen, Gisle (eds.), Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], 396-418.

Schmidt, Thoams (2017a): Memo Ausbauplan 2017.

Schmidt, Thomas (2017b): Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. In: Kupietz, Marc / Geyken, Alexander (eds.), Corpus Linguistic Software Tools. In: Journal for Language Technology and Computational Linguistics (JLCL) 31/1, 127-154.

Schmidt, Thomas (2017c): DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. In: Zeitschrift für Germanistische Linguistik 45, 3, 451-463.

Schmidt, Thomas (2018): Gesprächskorpora. Aktuelle Herausforderungen für einen besonderen Korpustyp. In: Kupietz, Marc / Schmidt, Thomas (Hg.), Korpuslinguistik. Berlin/Boston: de Gruyter, 209-230.

Schütte, Wilfried (2001): Alltagsgespräche. In: Brinker, Klaus / Antos, Gerd / Heinemann, Wolfgang / Sager, Sven F. (Hg.): Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. 2. Halbband. - Berlin/New York: de Gruyter, 1485-1492.

Steger, Hugo / Deutrich, Karl-Helge / Schank, Gerd / Schütz, Eva (1974). Redekonstellation, Redekonstellationstyp, Textexemplar, Textsorte im Rahmen eins Sprachverhaltensmodells. In: Moser, Hugo et al. (Hg.), Gesprochene Sprache. Düsseldorf: Schwann, 39-97.

Verdonik, Darinka / Kosem, Iztok / Zweitter-Vittez, Ana / Krek, Simon / Stabej, Marko (2013): Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. In: Language resources and evaluation 47, 4, 1031-1048.

Wiesinger, Peter [1983] (2008): Die Einteilung der deutschen Dialekte. In: Handbücher zur Sprach- und Kommunikationswissenschaft. Bd. 1/2 Dialektologie. Berlin u.a.: De Gruyter, 807-900.

Winterscheid (2016): Korpusstratifikation.

Online-Quellen der Veröffentlichungen des Statistischen Bundesamts DeStatis:

Aufstellungen zur Altersverteilung 1950-2017:

https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Bevoelkerungsstand/Tabellen_/lrbev01.html

Aufstellungen zu Bundesländern mit Hauptstädten nach Fläche, Bevölkerung und Bevölkerungsdichte am 31.12.2016:

https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/Administrativ/Aktuell/02Bundeslaender.html

Aufstellungen zu Bildungsstand 2008-2017:

https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Bildungsstand/Tabellen/Bildungsabschluss.html

Bildungsstand der Bevölkerung. Ergebnisse des Mikrozensus 2016. Destatis 2018:

https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002167014.pdf?__blob=publicationFile

Zeitverwendungserhebung. Aktivitäten in Stunden und Minuten für ausgewählte Personengruppen 2012/2013. Wiesbaden 2015:

https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/Zeitbudgeterhebung/Zeitverwendung5639102139004.pdf?__blob=publicationFile

7. Anhang

7.1. Dokumentierte Bildungsabschlüsse in FOLK

1. Hohes Bildungsniveau / "Tertiärbereich":

Habilitation; Promotion; Hochschulabschluss (Diplom, Master, Magister, 1./2. Staatsexamen, Bachelor); Fachhochschulabschluss (-reife, Diplom); Diplom ohne genauere Angabe; Fachschule (Diplom); Konzertreifeprüfung; Meister

2. Mittleres Bildungsniveau / "Sekundarstufe II":

Abitur; Fachabitur; Vordiplom; Kaufmännische Handelsschule; Berufskolleg (und abgeschlossene Berufsausbildung, Lehre)

3. Niedriges Bildungsniveau / "Sekundarstufe I" und "Primarstufe":

Mittlere Reife / mittlerer Bildungsabschluss / Realschulabschluss; Wirtschaftsschule; Werkrealschule; Hauptschulabschluss (/ Volksschule); Grundschule

4. Nicht vorhanden; Nicht dokumentiert

7.2. Systematik und Verteilung der Sprachregionen

6 Großregionen (nach Lameli 2008)	15 Sprachregionen (nach Wiesinger [1983] 2008)	16 Bundesländer
Nordwest	Nordniederdeutsch; Westfälisch; Ostfälisch	Schleswig-Holstein; Hamburg; Niedersachsen; Bremen; Nordrhein-Westfalen
Nordost	Mecklenburg-Vorpommerisch; (Mittelpommerisch); Brandenburgisch	Mecklenburg-Vorpommern; Berlin; Brandenburg; Teil von Sachsen-Anhalt
Mittelwest	Ripuarisch; Moselfränkisch; Hessisch; Rheinfränkisch;	Hessen; Rheinland-Pfalz; Saarland; Teil von Nordrhein-Westfalen; (Teil von Baden-Württemberg)
Mittelost	(Ober-)Sächsisch; Thüringisch	Sachsen; Thüringen; Teil von Sachsen-Anhalt
Südwest	Schwäbisch; Alemannisch	Baden-Württemberg
Südost	Bairisch; Ostfränkisch	Bayern

Tabelle 2: Sprachregionen/Bundesländer

Bundesland	**Einwohnerzahl**	**%**	**Mapping**[29] **6er-Regionen-Einteilung**	**%**
Schleswig-Holstein	2 881 926	3,49%	nordwest	nordwest ges.: 25 839 872 = 31,31%
Hamburg	1 810 438	2,19%	nordwest	
Niedersachsen	7 945 685	9,63%	nordwest	
Bremen	678 753	0,82%	nordwest	
Nordrhein-Westfalen	17 890 100	21,68%	nordwest; mittelwest (70/30 gezählt)	
Hessen	6 213 088	7,53%	mittelwest	mittelwest ges.: 16 642 822 = 20,17%
Saarland	996 651	1,21%	mittelwest	
Rheinland-Pfalz	4 066 053	4,93%	mittelwest	
Baden-Württemberg	10 951 893	13,27%	südwest (mittelwest)	südwest ges.: 10 951 893 = 13,27%
Bayern	12 930 751	15,67%	südost	südost ges.: 12 930 751 = 15,67%
Berlin	3 574 830	4,33%	nordost	nordost ges.: 8 798 278 = 10,66%
Brandenburg	2 494 648	3,02%	nordost	
Mecklenburg-Vorpommern	1 610 674	1,95%	nordost	
Sachsen	4 081 783	4,95%	mittelost	mittelost ges.: 7 358 037 = 8,92%
Sachsen-Anhalt	2 236 252	2,71%	mittelost; nordost (50/50 gezählt)	
Thüringen	2 158 128	2,62%	mittelost	
Deutschland	82 521 653	100,00%		

Tabelle 3: Bundesländer/Sprachregionen

Dr. Julia Kaiser
Institut für deutsche Sprache
R5, 6-13
68161 Mannheim

kaiser@ids-mannheim.de

Veröffentlicht am 29.1.2019

[29] Diese Verteilungsstatistik ist nur als grobe Annäherung zu verstehen. Bei Bundesländern, die eindeutig mehreren Sprachregionen zuzuordnen sind, werden die Anteile nach der jeweiligen Fläche ungefähr aufgeteilt; variierende Besiedlung wird dabei nicht berücksichtigt. Der kleine Anteil von Baden-Württemberg am Sprachgebiet Mittelwest wird bei der Zählung außer Acht gelassen.