Originalveröffentlichung in: Korpora Deutsch als Fremdsprache Jg. 3 (2023) Nr. 1, S. 166-169.
DOI: *https://doi.org/10.48694/kordaf.3737*

Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 1 (Juli 2023)
ISSN 2749-4411

FOLK
Das *Forschungs- und Lehrkorpus für Gesprochenes Deutsch*

Thomas Schmidt
linguisticbits.de

Abstract

Das *Forschungs- und Lehrkorpus für Gesprochenes Deutsch* (*FOLK*) ist ein Korpus des gesprochenen Deutsch in natürlichen sozialen Interaktionen, das seit 2008 in der Abteilung Pragmatik am Leibniz-Institut für Deutsche Sprache in Mannheim aufgebaut wird. *FOLK* besteht aus Audio- und Videoaufzeichnungen natürlicher Gespräche aus verschiedensten gesellschaftlichen Bereichen (private, institutionelle und öffentliche Interaktionsdomäne), die durch Transkription, weitere Annotationen und Metadaten-Dokumentation für korpusgestützte Analysen erschlossen und zur wissenschaftlichen Nutzung bereitgestellt werden. *FOLK* wird auf vielfältige Weise für Untersuchungen zum gesprochenen Deutsch genutzt, insbesondere in der Gesprächsforschung, der Korpuslinguistik und anwendungsorientierten Zweigen der Linguistik.

Keywords: gesprochene Sprache; mündliche Interaktion; Gesprächskorpora; Korpora in DaF/DaZ

Abstract

The *Research and Teaching Corpus of Spoken German* (*FOLK*) is a corpus of spoken German in natural social interactions. It has been built up since 2008 in the Pragmatics department of the Leibniz Institute for the German Language in Mannheim. *FOLK* consists of audio and video recordings of natural conversation in diverse areas of society (private, institutional and public interaction domains) that are made accessible for corpus-based analyses through transcription, annotation on several levels and metadata documentation. *FOLK* is used in various ways for research in conversation analysis, corpus linguistics and application-oriented branches of linguistics.

Keywords: spoken language; oral interaction; conversation corpora; corpora in DaF/DaZ

1. Das *FOLK*-Korpus: Primär- und Metadaten

Das *Forschungs- und Lehrkorpus* für Gesprochenes Deutsch (*FOLK*) ist ein Korpus des gesprochenen Deutsch in natürlichen sozialen Interaktionen, das seit 2008 in der Abteilung Pragmatik am Leibniz-Institut für Deutsche Sprache in Mannheim aufgebaut wird. Initiatoren des *FOLK*-Projekts sind Arnulf Deppermann und Martin Hartung (vgl. Deppermann / Hartung 2011). Das Projekt wurde von 2012 bis 2019 von Thomas Schmidt geleitet, seit 2019 leitet es Silke Reineke.

FOLK besteht aus Audio- und Videoaufzeichnungen natürlicher Gespräche aus verschiedensten gesellschaftlichen Bereichen, die durch Transkription, weitere Annotationen und Metadaten-Dokumentation für korpusgestützte Analysen erschlossen und zur wissenschaftlichen Nutzung bereitgestellt werden.

FOLK ist primär nach Interaktionsdomänen und Gesprächstypen stratifiziert (vgl. Kaiser 2018). Beispielsweise enthält *FOLK*:

1. aus der privaten Interaktionsdomäne: Tischgespräche, Telefongespräche, aktivitätsbegleitende Gespräche (z.B. Gespräch beim Kochen), Spielinteraktionen, Vorleseinteraktionen;
2. aus der institutionellen Interaktionsdomäne: Unterrichtsinteraktionen, berufliche Besprechungen (z.B. Team-Meeting, Schichtübergabe), Prüfungsgespräche, Verkaufsgespräche, Service-Interaktionen;

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-121027

3. aus der öffentlichen Interaktionsdomäne: Öffentliche Schlichtungsgespräche, Podiumsdiskussionen, Ausschuss-Sitzungen.

Bezüglich der enthaltenen Gesprächstypen strebt *FOLK* eine maximale Variationsbreite an. Soziodemographische Merkmale der SprecherInnen (Geschlecht, Alter, regionale Herkunft) gehen als sekundäre Stratifikationsparameter ins Korpusdesign ein. Auch hier wird beim Korpusausbau eine möglichst breite Abdeckung möglicher Merkmalsausprägungen (z.B. Alter, sprachlich prägende Region) angestrebt (vgl. Kaiser 2018; Reineke / Deppermann / Schmidt 2022). Typischerweise werden für *FOLK*-Gespräche auf Deutsch unter vollkompetenten MuttersprachlerInnen ausgewählt, Lernersprache und Sprachwechsel kommen aber in einigen wenigen Gesprächen als natürlicher Teil der Interaktion vor. Jedes Gespräch und alle daran beteiligten SprecherInnen werden systematisch durch Metadaten dokumentiert, die zur Datenauswahl und Analyse genutzt werden können (vgl. dazu Deppermann / Reineke i.Dr.).

2. Transkription und Aufbereitung

Aus Datenschutzgründen werden in den Audio- und Videodaten Stellen, die eine unmittelbare Identifikation der am Gespräch beteiligten Personen ermöglichen würden (wie etwa: Namensnennungen, Ortsnamen, Telefonnummern) maskiert und in den Transkripten entsprechende Pseudonyme verwendet (vgl. Reineke et al. 2017). Die Transkription der Gespräche erfolgt nach cGAT (vgl. Schmidt / Schütte / Winterscheid 2015) in literarischer Umschrift, so dass allgemeine Phänomene der Mündlichkeit („nich“ für „nicht“, „zwo“ für „zwei“, „biste“ für „bist Du“) und regionale Abweichungen von der Aussprachenorm (hessisch „runner“ für „runter“, bairisch „koans“ für „keins“) abgebildet werden. Transkripte sind in Sprecherbeiträge gegliedert, eine Unterteilung in Satzäquivalente erfolgt nicht. Die literarisch transkribierten Wörter werden nach einer Tokenisierung auf einer zweiten Ebene orthographisch normalisiert, d.h. ihrem standardorthographischen Äquivalent zugeordnet (vgl. Winterscheid et al. 2019). Auf der Grundlage dieser Normalisierung werden die Daten mit dem TreeTagger lemmatisiert und mit einem Part-Of-Speech-Tagging nach STTS 2.0 versehen (vgl. Westpfahl 2020). Für eine Auswahl an Daten liegen außerdem Annotationen von Handlungen und Themenfeldern vor (vgl. Kaiser 2023 in dieser Themenasgabe). Technisch wird der Arbeitsablauf für die Erschließung von *FOLK*-Daten mit den Tools *FOLKER* und *OrthoNormal* des *EXMARaLDA*-Systems umgesetzt (vgl. Schmidt 2016).

In der Version vom Juli 2022 (Version 2.18) umfasst *FOLK* 400 Gesprächsaufnahmen mit einer Dauer von rund 336 Stunden, von denen zu 151 Stunden auch Videoaufnahmen vorliegen. Die vollständigen Transkripte aller Aufnahmen umfassen ca. 3,2 Millionen Token (vgl. Deppermann / Reineke i.Dr.). *FOLK* ist als wachsendes Korpus konzipiert und wird kontinuierlich ausgebaut.

3. Zugriffsmöglichkeiten

FOLK ist über die Datenbank für Gesprochenes Deutsch[1] (vgl. Schmidt 2017) und über die *ZuMult*-Tools[2] (vgl. Fandrych et al. 2023 in dieser Themenausgabe) für die wissenschaftliche Öffentlichkeit (d.h. für Forschende, Lehrende und Studierende) zugänglich. Die Nutzung erfordert eine persönliche Registrierung. Aus Gründen des Datenschutzes, die sich u.a. aus den Vereinbarungen ergeben, die mit den aufgenommenen Personen getroffen wurden, darf *FOLK* nur für Zwecke wissenschaftlicher

[1] https://dgd.ids-mannheim.de (05.07.2023).
[2] https://zumult.ids-mannheim.de/ (05.07.2023).

Forschung und Lehre verwendet werden. Der DaF-/DaZ-Unterricht an Hochschulen und für Zwecke sprachdidaktischer Forschung fallen unter diese zulässigen Verwendungen, nicht aber der Einsatz von *FOLK*-Daten im Sprachunterricht außerhalb der Hochschule.

4. Anwendung

FOLK wird auf vielfältige Weise für Untersuchungen zum gesprochenen Deutsch genutzt. Zahlreiche gesprächsanalytische und interaktionslinguistische Untersuchungen basieren auf *FOLK*. In Reineke / Deppermann / Schmidt (2022) finden sich ausgewählte Beispiele solcher Untersuchungen. Schmidt (2014) bearbeitet gesprächsanalytische Fragestellungen mit korpuslinguistischen Methoden und bezieht dabei weitere mündliche Korpora aus der Datenbank für Gesprochenes Deutsch ein. Anwendungsbeispiele für den DaF-/DaZ-Bereich, die über die *ZuMult*-Tools bearbeitet wurden, finden sich in Fandrych et al. (2023 in dieser Themenausgabe). Im Projekt *LeGeDe* (vgl. Meliss et al. 2019) diente *FOLK* als Grundlage zur Erstellung einer lexikalischen Ressource des gesprochenen Deutsch[3], die auch Anwendungsperspektiven für DaF/DaZ beinhaltet (vgl. Meliss 2021)[4].

Literatur und Ressourcen

Deppermann, Arnulf / Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des ‚Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard / Müller, Marcus / Vogel, Friedemann (Hrsg.): *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen.* Berlin / New York: de Gruyter, 414-450.

Deppermann, Arnulf / Reineke, Silke (i. Dr.): Zur Verwendung von Metadaten in der interaktionsanalytischen Arbeit mit Korpora – am Beispiel einer Untersuchung anhand des Korpus FOLK. Erscheint in: *Korpusgestützte Sprachanalyse: Linguistische Grundlagen, Anwendungen und Analysen.*

Fandrych, Christian / Meißner, Cordula / Schwendemann, Matthias / Wallner, Franziska (2023): ZuMal: Zielgruppenspezifische Gesprächsauswahl aus Korpora gesprochener Sprache. In: *Korpora Deutsch als Fremdsprache* 3: 1, 13-43.

Kaiser, Julia (2018): Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 19, 515-552.

Kaiser, Julia (2023): ZuHand: Zugang zu Handlungssequenzen und handlungsbezogenen Themenausschnitten in einem qualitativ annotierten FOLK-Subkorpus. In: *Korpora Deutsch als Fremdsprache* 3: 1, 92-111.

Meliss, Meike / Möhrs, Christine / Ribeiro Silveira, Maria / Schmidt, Thomas (2019): A Corpus-Based Lexical Resource of Spoken German in Interaction. In: Kosem, Iztok / Zingano Kuhn, Tanara / Correia, Margarita / Ferreria, José Pedro / Jansen, Maarten / Pereira, Isabel / Kallas, Jelena / Jakubíček, Miloš / Krek, Simon / Tiberius, Carole (Hrsg.): *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.* 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, 783-804.

Meliss, Meike (2021): Die LeGeDe-Ressource: korpusbasierte lexikografische Einblicke und anwendungsorientierte Ausblicke. In: *Deutsch als Fremdsprache* 58: 1. Berlin: Erich Schmidt Verlag, 3-15.

[3] https://www.owid.de/legede/ (05.07.2023).
[4] Unter https://www.ids-mannheim.de/prag/muendlichekorpora/bibliographie-folk/ (05.07.2023) findet sich eine laufende Bibliographie von Arbeiten, für die *FOLK* genutzt wurde.

Reineke, Silke / Schmidt, Thomas / Schedl, Evi / Kaiser, Julia (2017): *Maskierung von Audio- und Videoaufnahmen*. Version 2.1, Überarbeitung und Ergänzung. Mannheim: Leibniz-Institut für Deutsche Sprache.

Reineke, Silke / Deppermann, Arnulf / Schmidt, Thomas (2022): Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK). Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen. In: Deppermann, Arnulf / Fandrych, Christian / Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multi-medial.* Berlin / Boston: de Gruyter, 71-102.

Schmidt, Thomas (2014): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 15, 196-233.

Schmidt, Thomas / Schütte, Wilfried / Winterscheid, Jenny (2015): *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*. Mannheim: Leibniz-Institut für Deutsche Sprache.

Schmidt, Thomas (2016): Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. In: Kupietz, Marc / Geyken, Alexander (Hrsg.): *Corpus Linguistic Software Tools, Journal for Language Technology and Computational Linguistics (JLCL* 31/1*)*, 127-154.

Schmidt, Thomas (2017): 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.

Westpfahl, Swantje (2020): *POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Studien zur deutschen Sprache (83). Tübingen: Narr.

Winterscheid, Jenny / Deppermann, Arnulf / Schmidt, Thomas / Schütte, Wilfried / Schedl, Evi / Kaiser, Julia (2019): *Normalisieren mit OrthoNormal. Konventionen und Bedienungshinweise für die orthografische Normalisierung von FOLKER-Transkripten*. Mannheim: Leibniz-Institut für Deutsche Sprache.

Biographische Notiz: Thomas Schmidt hat bis 2021 den Programmbereich Mündliche Korpora am Institut für Deutsche Sprache geleitet und war dort u.a. für die Leitung des Archivs für Gesprochenes Deutsch (AGD), den Aufbau der Datenbank für Gesprochenes Deutsch (DGD2) und (bis 2019) den Aufbau von *FOLK* verantwortlich. Thomas Schmidt arbeitet derzeit als Software-Entwickler im Bereich KI, Musik und Sprache und bietet über linguisticbits.de wissenschaftliche Dienstleistungen im Bereich von Korpora und Korpustechnologie an.

Kontaktanschrift:
Dr. Thomas Schmidt
Adam-Karrillon-Straße 13
D-55118 Mainz
Deutschland
thomas@linguisticbits.de

Lizenz: CC BY 4.0 International - Creative Commons, Namensnennung.