1. Allgemeine Angaben:

Antragsteller*in: Nicole Palliwoda (Christian-Albrechts-Universität zu Kiel)
Alexander Werth (Universität Passau)

FDZ: Archiv für Gesprochenes Deutsch (AGD) am Leibniz-Institut für Deutsche Sprache (IDS)/Thomas Schmidt

Thema des Projekts: Sprachsituation an der innerdeutschen Grenze

Berichtszeitraum: 01.07.2021-30.06.2022

Förderungszeitraum insgesamt: 01.09.2021-30.06.2022

Vorhaben und Ziele:

Das Forschungsdatenmanagement-Projekt hatte zum Ziel, Daten aus zwei Projekten zur Sprachsituation an der innerdeutschen Grenze für die Sekundärnutzung zu erschließen und bereitzustellen. Zusammengenommen handelt es sich um den größten Datenbestand zu dieser Thematik. Dieser war der wissenschaftlichen Community bislang nicht zugänglich und sollte gemäß den FAIR-Prinzipien für eine Nachnutzung aufbereitet werden. Alle Sprachdaten und dazugehörige Materialien und Metadaten sollten vom FDZ *Archiv für Gesprochenes Deutsch (AGD)*, u. a. über die *Datenbank für Gesprochenes Deutsch (DGD)*, für eine Nachnutzung in digitaler und anonymisierter Form zur Verfügung gestellt werden. Die besondere Attraktivität des Korpus liegt dabei zum einen in der sehr großen Orts- bzw. Gewährspersonendichte, sie betrifft zum anderen aber auch die Datenpluralität, indem von jeder Gewährsperson (GP) verschiedene Datentypen – freie Gespräche, Übersetzungen, Interviews und Fragebogendaten – erhoben wurden. Um die Zugänglichkeit des Korpus zu erhöhen, sollte eine Auswahl an normorthografischen Transkripten text-ton-aligniert und auf Wortebene mit Lemma- und Part-Of-Speech(POS)-Information annotiert werden. Somit werden detaillierte Korpusrecherchen nach Wörtern oder Segmentfolgen im Transkript über die DGD möglich.

Beschreibung der Daten – Projekt 1 (SPRiG):

Bei den Daten des *SPRiG-Projekts (Untersuchungen zur Sprachsituation im thüringisch-bayerischen Grenzgebiet* (2005-2011, gefördert durch die DFG)*)* handelt es sich um Daten, die von dem Kooperationsprojekt (*Erhebungen zur Dialektsituation im thüringisch-bayerischen Grenzgebiet* (1992-1994, gefördert durch die DFG)) vom *Thüringischen Wörterbuch (ThWb)* und vom *Sprachatlas von Nordostbayern (SNOB)* erhoben wurden. Das Kooperationsprojekt hatte zum Ziel, Sprachdaten unterschiedlicher Altersgruppen (AG) an der ehemaligen innerdeutschen Grenze zu erheben, um den Sprachstand dies- und jenseits der Grenze für spätere Analysen zu sichern. Die erhobenen und gesicherten Materialien wurden dann in dem SPRiG-Projekt unter der Leitung Rüdiger Harnischs (Universität Passau) systematisch analysiert. Dabei lag der Fokus auf der Analyse phonetisch-phonologischer, morphologischer und lexikalischer Merkmale. Im Zuge dessen wurden die Daten des Kooperationsprojekts dem SPRiG-Projekt überlassen. Das SPRiG-Projekt übernahm damit die Dateninhaberschaft vom Erhebungsprojekt. 2021 existieren die Ur-Einrichtungen (ThWb,

SNOB) nicht mehr, deren ehemalige Leiter sind verstorben. Die somit verwaisten Daten wurden im April 2021 von Rüdiger Harnisch an den neuen Lehrstuhlinhaber Alexander Werth übergeben.
Insgesamt lagen laut internen Angaben Sprachaufnahmen von 435 GPen aus insgesamt 21 Orten aus dem thüringisch-bayerischen Grenzgebiet vor. 20 dieser Orte liegen sich an der ehemaligen innerdeutschen Grenze geographisch gegenüber und der 21. Ort war – wie damals Berlin – zwischen Ost und West geteilt. Pro Ort wurden ca. 20 GPen befragt. Diese verteilen sich auf vier verschiedene Altersgruppen (AG) (Geburtsjahr um 1920 (AG 1), 1940 (AG 2), 1955 (AG 3), 1970 (AG 4)), um einen intergenerativen Vergleich zu ermöglichen.
Die Tonbandmitschnitte bestehen aus vier Teilen:

1. Abfrage von Wörtern/Satzgruppen zur Realisierung primärer Dialektmerkmale (insg. 415 Tonbandaufnahmen) und Übertragung von 25 standardsprachlichen Sätzen in den jeweiligen Ortsdialekt (insg. 420 Tonbandaufnahmen);
2. Fragen zur Sprachverwendung und Dialektbewertung mittels eines soziolinguistischen Fragebogens (In welchen Situationen wird Dialekt (Hochdeutsch, Umgangssprache) gesprochen? Welche Dialekte mögen Sie? Wie sprechen Sie mit Ihren Kindern? Wodurch unterscheidet sich Ihr Dialekt vom Dialekt in den Nachbardörfern?) (insg. 435 Tonbandaufnahmen od. schriftlich fixierte Interviews);
3. Abfrage von Sozialdaten der Gewährspersonen (Alter, Geburtsort, Beruf, Tätigkeit der Eltern) (insg. 435 GPen);
4. eine freie Erzählung im Dialekt (natürlich gesprochene Alltagssprache) zum Thema Grenzöffnung 1989 (insg. 435 Tonbandaufnahmen).

Alle GPen wurden vorab durch die Explorator:innen über den Zweck der Aufnahmen, d.h. deren Verwendung für wissenschaftliche Auswertung, mündlich informiert. Schriftliche Einverständniserklärungen wurden den damaligen Gepflogenheiten der Datenerhebung entsprechend nicht eingeholt. Aufgrund dessen wurde besonderer Wert auf die anonymisierte Aufbereitung der Daten zu Forschungszwecken unter dem Gebot der Datensparsamkeit gelegt, sodass kein direkter Personenbezug mehr möglich ist.
Das Audiomaterial des SPRiG-Archivs an der Universität Passau bestand aus 335 digitalisierten Tondokumenten. Jedes Tondokument entspricht einer Seite einer analog aufgenommenen Audio-Kompakt-Cassette, ist im Schnitt 30 Minuten lang (insg. ca. 10.050 min/ca. 168 h) und war somit noch nicht sprecher:innenspezifisch aufbereitet. Digitalisierte Fassungen der Tondateien wurden als WAV-Dateien und/oder MP3-Dateien auf CD-Rom und/oder DVD-Rom gesichert. Zu den Aufnahmen wurde ein Protokoll mit Informationen zur Aufnahmequalität, zur Länge und zu Sprecher:innenwechseln erstellt. Zudem existierten weitere CDs und DVDs. Die Audio-Daten wurden auf einer externen 3,5“-Festplatte (USB 2.0) gesichert. Eine Cloud-Sicherung (Leibniz-Rechenzentrum) der MP3-Dateien existierte sowie zusätzlich die Sicherung der bereits vorhandenen Transkripte (Stand der Antragsstellung: ca. 342 Word-Dokumente, die das Thema Grenzöffnung und 97 Word-Dokumente, die das Thema Sprachverwendung beinhalteten).

Beschreibung der Daten – Projekt 2 (SEIG):

Die Sprachdaten des SEIG-Projekts (*Sprechen über die ehemalige innerdeutsche Grenze*) wurden durch die Dateninhaberin Nicole Palliwoda zwischen 2013 und 2014 eigenständig erhoben. Acht Interviews konnten 2020/2021 bisher zur Grenzöffnungsthematik in einem Aufsatz analysiert werden. Hierfür wurden einfache literarische Transkripte ohne weitere Annotationen angefertigt. Insgesamt handelt es sich um Sprachaufnahmen von 29 GPen (14

männliche und 15 weibliche), von denen 16 ins AGD und die DGD überführt werden sollten. Die Personen stammen aus vier Orten an der ehemaligen innerdeutschen Grenze (Thandorf/MV (ehemalige DDR) vs. Salem (Dargow)/SH (ehemalige BRD); Struth/TH (ehemalige DDR) vs. Wanfried/HE (ehemalige BRD)), die sich auf zwei AG (AG 1: Geburtsjahr zw. 1934-1944, AG 2: Geburtsjahr zw. 1977-1984) verteilten.
Den Sprachaufnahmen zugrunde lag ein dreiteiliges Interview:

1. Ein autobiographisch narratives Interview, bei dem die GP einen Erzählstimulus erhielten: Sie sollten ihre Lebensgeschichte an der ehemaligen innerdeutschen Grenze erzählen, von ihren Erinnerungen und Erfahrungen von der Kindheit bis zum damaligen Zeitpunkt berichten.
2. Eine sog. Draw-a-Map-Task zum eigenen Sprachraum: Der GP wurde ein geografischer Kartenausschnitt mit dem Wohnort in der Mitte vorgelegt, um den ein 50/100 km Radius gezogen wurde. Auf der Karte sollte eingezeichnet und gleichzeitig kommentiert werden, wo überall gleich/anders gesprochen wird. Des Weiteren wurden Fragen zur Sprachbiographie, zur Verbundenheit und zum Status der Sprechweise behandelt.
3. Die Verortung von alltagssprachlichen Sprechproben: Die GP sollte zehn Sprechproben[1] auf einer Deutschlandkarte verorten und begründen, warum sie diese Verortung vorgenommen hat.

Zu den Sprachaufnahmen, den mentalen Karten der GPen sowie den Sozialdaten (Geschlecht, Geburtsjahr, Geburtsort, Sozialisationsorte, Wohnorte, Geburtsort der Eltern, Sozialisationsorte der Eltern, Beruf) existiert für jede GP eine unterschriebene Einverständniserklärung, die eine Weitergabe der Daten in anonymisierter Form autorisiert. Die Interviews sind im Durchschnitt 110 Minuten lang (insg. ca. 3190 min/ca. 53 h) und liegen im WAV-Format (ca. 34 GB) vor.

2. Arbeits- und Ergebnisbericht:

Die Bestände der beiden Projekte zur Sprachsituation an der innerdeutschen Grenze wurden an der Universität Passau gemeinsam bearbeitet, sie werden aus Gründen der Übersichtlichkeit hier aber getrennt voneinander beschrieben:

SPRiG-Materialien: Zu Projektbeginn lagen Sprachaufnahmen, teilweise auch Transkripte von 435 GPen vor. Es hat sich im Laufe des Projektes aber an mehreren Stellen herausgestellt, dass dieser Bestand hinsichtlich der Digitalisate und Transkriptionen stark fehlerbehaftet und auch unvollständig war, sodass eine weit über den ursprünglichen Projektplan hinausgehende Bearbeitung und auch Neudigitalisierung der Daten stattfinden musste. Hieraus ergibt sich eine Diskrepanz zwischen den Bestandszahlen bei Projektbeginn und denen bei Projektabschluss. Der dabei entstandene Mehraufwand wurde in Teilen aus dem Projekt selbst, in Teilen mit vorhandenem Personal am Lehrstuhl für Deutsche Sprachwissenschaft in Passau sowie am Leibniz-Institut für Deutsche Sprache in Mannheim abgedeckt.

[1] Die Sprechproben für die Verortungsaufgabe entstammen aus dem am Leibniz-Institut für Deutsche Sprache (IDS) durchgeführten Projekt *Gesprochenes Deutsch* (Korpus *Deutsch heute*). Diese Aufnahmen sind zudem auch über das AGD und die DGD abrufbar.

Die zu Projektbeginn vorgefundenen Digitalisate waren insofern fehlerhaft, als bei der Digitalisierung unsystematisch und deshalb für uns im Einzelnen nicht nachvollziehbar Fremdgeräusche den Weg in die Digitalisate gefunden haben (z. B. Mausklicks, Musik im Hintergrund). Auch wurden bei der Digitalisierung manche Aufnahmenummern und Orte vertauscht, mitunter fehlten auch Sprecher:inneninformationen, sodass eine Zuordnung der Aufnahme und Identifikation der Sprecherin/des Sprechers über die Aufnahme selbst vorgenommen werden musste. Die Transkripte gaben mitunter auch nicht das in den Aufnahmen tatsächlich Gesprochene wieder, sie mussten außerdem von einer Umschrift, die am gesprächsanalytischen Transkriptionssystem (GAT) orientiert war, in eine normorthografische Schreibung transformiert werden, um die Transkripte für Korpusabfragen brauchbar zu machen. Auch wurden die Transkripte aus ihrer ursprünglichen Textform in ein strukturiertes, maschinenlesbares XML-Format (EXB) überführt. In einer umfassenden Recherche zu den Datenbeständen stellte sich zudem heraus, dass bei einem Projektpartner des früheren SPRiG-Projektes, dem Thüringischen Wörterbuch, weitere Ton- und Papierdokumente gelagert waren. Diese wurden an die Universität Passau überführt und dort bearbeitet.

Da das AGD nicht mit der Digitalisierungs-Aufgabe geplant hatte und erst später (nach der IDS-Jahrestagung im März 2022) die Kapazität hatte, um die Digitalisierung vorzunehmen, waren in Passau durch die Hilfskräfte bereits die Mehrzahl der Transkripte bearbeitet und mit den ursprünglichen Audios (Erst-Digitalisaten) aligniert. Neu entstandene Digitalisate wurden direkt aligniert.
Bei den bereits auf die Original-Digitalisate alignierten Transkripten entstand eine Diskrepanz zu den Neudigitalisaten: ein Offset, der durch zwangsläufig unterschiedliche Start-Zeitpunkte entstand, und unterschiedliche und dynamische Drifts, die durch unterschiedliche Geschwindigkeiten von Tape-Recordern und die Alterung der Kassetten-Bänder erklärbar ist. Diese Probleme wurden durch technische Mittel gelöst. Der Offset wurde durch die Berechnung der Kreuzkorrelationen am Anfang der Aufnahmen ermittelt. Der Drift wurde durch Dynamische Programmierung ermittelt. Beide Parameter, Offset und Drift, wurden verwendet, um die Timeline der Transkripte an die Neudigitalisate anzupassen.

Mit Stand vom 30.06.2022 wurden im Projekt 508 Aufnahmen im Umfang von 75 Stunden digitalisiert. Zu diesen Audiodateien liegen 508 ton-text-alignierte Transkripte in einer normorthografischen Umschrift vor. Vollständig neu transkribiert und text-ton-aligniert wurden hiervon 70 Aufnahmen (10 Stunden). Bei den anderen 438 Aufnahmen und Transkripten (65 Stunden) wurden die vorhandenen Transkripte in das strukturierte EXMARaLDA-XML-Format überführt, an die neuen Konventionen angepasst und ebenfalls eine Ton-Text-Alignierung durchgeführt. Personenbezogene Daten in den Aufnahmen und Transkripten wurden vollständig anonymisiert. Die Sozialdaten der Gewährspersonen, die zu Projektbeginn in Form von handschriftlichen Sozialfragebögen vorlagen, wurden in eine Excel-Datenbank eingetragen, sind damit automatisiert durchsuchbar und werden im AGD auf das dort verwendete Metadatenmodell abgebildet.

SEIG-Materialien: Hier lagen zu Projektbeginn bereits die vollständig digitalisierten Aufnahmen (16) vor. Die auf den Sozialbögen vermerkten Sozialdaten zu den Personen (Herkunft, Geburtsjahr, aufgewachsen etc.) wurden ebenfalls in eine Excel-Datei überführt. Zudem wurden alle Papierdokumente (Einverständniserklärungen, mentale Sprachkarten des Nahbereichs der einzelnen Personen, Verortung von alltagssprachlichen Sprechproben etc.)

der 16 GPen digitalisiert, diese stehen als zusätzliches (Auswertungs-)Material im AGD und in der DGD zur Verfügung. Um den Bestand über das AGD und die DGD der Scientific-Community zugänglich zu machen, wurde der Gesamtbestand (ca. 35 Stunden Tonaufnahmen) hinsichtlich personenbezogener Daten vollständig anonymisiert. Davon wurden nur die darin enthaltenen Erzählungen zur Grenzöffnung im Umfang von 16,5 Stunden normorthografisch transkribiert und text-ton-aligniert. Von diesen lagen zu Projektbeginn bereits acht literarisch transkribierte, aber nicht annotierte Aufnahmen vor. Diese mussten noch den im Projekt vorgegebenen Konventionen angepasst und text-ton-aligniert werden. Die gezeichneten Karten der Sprechprobenverortung aus dem Interviewteil 3 wurden zudem mit den Sprachaufnahmen aus dem *Deutsch heute*-Korpus verlinkt.

Zum nächstmöglichen DGD-Release, vrsl. im Frühjahr 2023, werden die bearbeiteten Daten der Öffentlichkeit zugänglich gemacht. Die Veröffentlichung umfasst die digitalisierten Aufnahmen beider Bestände, die zugehörigen Transkripte und Sprecher:inneninformationen sowie Begleittexte zum Korpus, die dieses linguistisch verorten. Vorab wird den Mittelgeber:innen über den folgenden Link eine Einsichtnahme mit allen fertiggestellten Transkripten zur Verfügung gestellt.

https://tinyurl.com/SPIGKorpus
https://tinyurl.com/SEIGKorpus

3. Zusammenfassung

Das Forschungsdatenmanagement-Projekt *Sprachsituation an der innerdeutschen Grenze* hatte zum Ziel, Daten aus zwei Untersuchungen (DFG-Projekt: *Untersuchungen zur Sprachsituation im thüringisch-bayerischen Grenzgebiet (SPRIG)* und *Sprechen über die ehemalige innerdeutsche Grenze (SEIG)*, bisher nicht weiter veröffentlicht)) zur Sprachsituation an der innerdeutschen Grenze für die Sekundärnutzung zu erschließen und bereitzustellen. Zusammengenommen handelt es sich um den größten Datenbestand zu dieser Thematik. Dieser ist der wissenschaftlichen Community bislang nicht zugänglich gewesen und sollte für eine Nachnutzung aufbereitet werden. Mit dem Material werden Forschungsinteressen der Dialektologie, der Sprachwandelforschung und der Sprachsoziologie bedient. Hinzu kommen disziplinenübergreifende Anbindungen an Forschungsfragen der Soziologie (z. B. Sozio-Biografie), der Geschichtswissenschaft (bes. Oral History) und Politologie (z. B. Einstellungsforschung). Die besondere Attraktivität des Korpus ist dabei zum einen in der sehr großen Orts- bzw. Gewährspersonendichte zu sehen, sie betrifft zum anderen aber auch die Datenpluralität, indem von jeder Gewährsperson (GP) verschiedene Datentypen – freie Gespräche, Übersetzungen, Interviews und Fragebogendaten – erhoben wurden. Zudem ergeben sich innerhalb der DGD-Bestände Vergleichsmöglichkeiten zu thematisch ähnlichen Daten, insbesondere dem Berliner Wendekorpus und den Korpora „Deutsche Mundarten: Zwirner-Korpus“, „Deutsche Mundarten: DDR“ und „Deutsche Umgangssprache: Pfeffer-Korpus“.

Daten aus dem DFG-Projekt *Untersuchungen zur Sprachsituation im thüringisch-bayerischen Grenzgebiet (SPRIG)*:

Mit Stand vom 30.06.2022 wurden im Projekt 508 Aufnahmen im Umfang von 75 Stunden digitalisiert. Zu diesen Audiodateien liegen 508 ton-text-alignierte Transkripte in einer

normorthografischen Umschrift der Abschnitte vor, die Erzählungen zur Grenzöffnung beinhalten. Personenbezogene Daten in den Aufnahmen und Transkripten wurden vollständig anonymisiert, die Sozialdaten der Gewährspersonen, die zu Projektbeginn in Form von handschriftlichen Sozialfragebögen vorlagen, wurden in eine Excel-Datenbank eingetragen und sind damit automatisiert durchsuchbar.

Daten aus dem Projekt *Sprechen über die ehemalige innerdeutsche Grenze (SEIG)*:

Im Projekt wurden 16 Sprachaufnahmen (35 Stunden Tonaufnahmen) hinsichtlich personenbezogener Daten vollständig anonymisiert. Davon wurden die darin enthaltenen Erzählungen zur Grenzöffnung im Umfang von 16,5 Stunden normorthografisch transkribiert und text-ton-aligniert. Die auf den Sozialbögen vermerkten Sozialdaten zu den Personen wurden ebenfalls in eine Excel-Datei überführt. Zudem wurden alle Papierdokumente (Einverständniserklärungen, mentale Sprachkarten des Nahbereichs der einzelnen Personen, Verortung von alltagssprachlichen Sprechproben etc.) der 16 GPen digitalisiert. Zusätzlich wurden die durch die Proband:innen handgezeichneten Karten zur Verortung von alltagssprachlichen Sprechproben auf einer Deutschlandkarte mit den Sprachaufnahmen aus dem *Deutsch heute*-Korpus, die die Grundlage der Sprechprobenverortung bildeten und ebenfalls über die DGD abrufbar sind, verlinkt.

Bei der Erschließung und Bereitstellung der Daten war das übergeordnete Ziel eine grundständige Aufbereitung aller vorhandenen Daten sowie eines repräsentativen Samples an einheitlichen und qualitätskontrollierten Transkripten mit Fokus auf die Themen Grenzöffnung und Sprachverwendung. Dadurch wird insbesondere das interdisziplinäre Potenzial der Daten erschlossen. Eine weitere Erschließung für spezifisch variationslinguistische Zwecke – insbesondere Transkriptionen zu Wortlisten/Wortgruppen, Satzübersetzungen, Map-Tasks und Verortungen von Sprechproben – kann bei Bedarf von Nachnutzenden vorgenommen werden.

Alle Sprachdaten und dazugehörige Materialien und Metadaten wurden vom Forschungsdatenzentrum *Archiv für Gesprochenes Deutsch (AGD)* für eine Nachnutzung in digitaler und anonymisierter Form zur Verfügung gestellt und sind mit dem nächsten Release der *Datenbank für Gesprochenes Deutsch (DGD)* im Frühjahr 2023 zugänglich.