To appear in: Boas, Hans (Ed.): Comparative language island research: data, methods, and goals. – Leiden: Brill. 31 Pp.

Data of German Speech Minorities in the Archive for Spoken German: An Overview

Jan Gorisch

Thomas Schmidt

Ulf Michael Stift

Abstract

Speech islands are historically and developmentally unique and will inevitably disappear within the next decades. We urgently need to preserve their remains and exploit what is left in order to make research on language-in-contact and historical as well as current comparative language research possible.

The Archive for Spoken German (AGD) at the Institute for German Language collects, fosters and archives data from completed research projects and makes them available to the wider research community.

Besides large variation corpora and corpora of conversational speech, the archive already contains a range of collections of data on German speech minorities. The latter will be outlined in this chapter. Some speech island data is already made available through the personal service of the AGD, or the database of spoken German (DGD), e.g. data on Australian German, Unserdeutsch, or German in North America. Some corpora are still being prepared for publication, but still important to document for potentially interested research projects. We therefore also explain the current problems and efforts related to the curation of speech island data, from the digitization of recordings and the collection of metadata, to the integration of transcriptions, annotations and other ways of accessing and sharing data.

1 Introduction

Due to its uniqueness, speech island data[1] has been used for language and linguistics research in various ways: It has been investigated how speech minorities and languages have come into place and how they developed over time in terms of language change (Costello, 1978; Salmons, 1994, Fuller, 2000), speech change and sound change (Boas, 2002a; Keiser, 2000). Speech island research enables us to investigate how languages behave in different environments. Speech islands exist in close proximity to other established languages, e.g. German in contact with English (e.g. Clyne, 1981), German in contact with Russian (e.g. Berend, 2011), or even more complex: Russian-German in contact with Spanish, e.g. Volga German in Argentina (Ladilova, 2013), etc. Most speech island research also involves the analysis of cultural developments, comparing the original culture of the speech island community with the culture of the surrounding language community and how one influenced the other.

In general, research is more valid, transparent and reproducible when it is based on empirical data and works with quantification. This is true also for research on speech islands. Therefore, we as an archive for data of spoken German have the responsibility to preserve speech island data and to lend support to projects compiling new data collections. We have to inform and give advice to research projects on decisions about recordings, collection of metadata and participants' consent, transcription, annotation, data security, data protection, etc.

Doing research in this area is always linked to data. In archives such as the AGD (Stift & Schmidt, 2014), data are sometimes accessible in a highly structured form, e.g. in databases such as the database for spoken German (DGD, Schmidt, 2014a), sometimes they are available in less structured conditions but still well documented, sometimes they are only rudimentarily documented, undigitized, untranscribed and hardly visible to the research community. Against this background, the aim of this paper is twofold: on the one hand, we want to provide an overview of the speech island data currently hosted at the Archive for Spoken German (AGD). This overview goes beyond the information "officially" displayed on the archive's website (http://agd.ids-mannheim.de/korpus_index.shtml), since it also includes resources which, for various reasons, could not yet be made available in the archive's platforms (several sections under 2.1), and because it also takes into account relevant data from corpora which were not explicitly designed as speech island corpora (section 2.2). On the other hand, we want to discuss, from an archive's perspective, two more general aspects of the work with speech island corpora, namely the question of data curation (section 3) and of enabling comparative research on the basis of curated corpora (section 4).

We would like to dedicate this article to the memory of our dear colleague and co-author Ulf-Michael Stift who passed away suddenly and unexpectedly at the turn of the year 2022. A historian by education, Ulf had been the AGD's archivist from 2008 on, specializing in the archive's historical stock. Without his diligent work on archived project files, reel-to-reel-

[1] We do not use the term "speech island (data)" in this chapter as a technical term. We are aware that its suitability for describing some German-based varieties can be contested. However, contenders such as "speech minority", "extra-territorial variety" or even "German abroad" are problematic in their own ways. There is probably no established term which would fully adequately subsume all the different types of data that we cover in this chapter.

tapes and other analogue materials, and without his historian's informed view on that data, digitizing the archive would have been an impossible task, and this article would have been impossible to write. Ulf will be dearly missed, not only in his capacity as a highly professional colleague, but also as a kind and reliable person.

2 AGD and DGD speech island data

Besides large corpora of spoken varieties within Germany and the surrounding countries (such as the "classical" German dialect corpora by Zwirner et al. 1958) and corpora of conversational speech (such as FOLK, Schmidt, 2014b), the Archive of Spoken German hosts a range of collections of data on German speech minorities, some of which are published in the database for spoken German (DGD). It is not always straightforward to draw an exact line between recordings in the variation corpora and recordings that have been made with the intention to record specifically speech island data. For example, at the time when recordings for the Zwirner corpus (Zwirner and Bethge, 1958) were made, there were migrations of people who formerly lived in speech islands, mostly in Eastern Europe, and who were then recorded outside of their speech community and categorised according to the dialectal region that their speech island dialect was most similar to. In order to give interested speech island researchers a complete overview of German speech island data in the AGD and DGD, we include in this section, besides "pure" speech island corpora (Section 2.1), also those corpora that have been identified to contain speech island data (Section 2.2).

For quicker reference, shortcuts to subsections are provided via Table 1 where we indicate for each speech island the corpora (abbreviations) in which the corresponding data can be found. The table shows that corpora such as MV (Mundarten und Varia) and OS (formerly German territories in Eastern Europe) contain data from several speech islands.

Table 1: Overview on the various German speech islands according to corpora and countries.

Speech island	**Corpus abreviation**	**Country/region**
Australian German	AD, MV	Australia
Swabian in Panambi and other colonist German in Southern America	BA, MV	Brazil, Paraguay
Swabian in the Black Sea region ("Schwarzmeerschwäbisch")	MV	Ukraine
Mennonite Low German ("Plautdietsch")	NO, MEND, MV, OS, ZWTV	USA, Mexico, Paraguay, Bolivia, Brazil, Canada, Russia/Ukraine, Poland
Saxonian in Transylvania ("Siebenbürgersächsisch")	RU, MV, OS, ZW	Romania
Volhynia German ("Wolhyniendeutsch")	NO, MV, OS, ZW	Ukraine, Belarus

Pennsylvania Dutch (“Pennsylvaniadeutsch”)	NO, MV, PEND, ZWTV	Pennsylvania, Ohio/USA
German in Russia (“Russlanddeutsch”)	RS, RUDI, OS, ZWTV	Russia
German in the Carpathians (“Karpatendeutsch”)	SL, OS, ZW	Slovakia
Ali Pidgin	OZ	Ali Island/Papua New Guinea
Unserdeutsch (“Rabaul Creole German”)	UNSD, OZ	Papua New-Guinea, Australia
German in Namibia (“Namdeutsch”)	NAMD, AF	Namibia
Palatine dialect in Canada (“Siedlungspfälzisch”)	MV	Ontario/Canada
Dane County Kölsch and other varieties in Wisconsin	MV	Wisconsin/USA
German in Puhoi (“Puhoideutsch”)	MV, NZ	New Zealand
Baltic German	OS, ZW, ZWTV	Estonia, Latvia
German in Poland	OS, ZW	Poland
Galician German	OS, ZW, ZWTV	Poland, Ukraine
Bukovina German	OS, ZW	Ukraine, Romania
Bessarabian German	OS, ZW, SV, ZWTV	Ukraine, Moldova
Hungarian German (Bavarian dialect in northern Hungary)	ZW, OS	Hungary
Swabian at the Danube (“Donauschwäbisch”): Southern Hungary, Batschka, Slavonia, Syrmia, Banat	OS, ZW, MV	Hungary, Croatia, Bosnia-Hercegovina, Serbia, Romania
Dobrudscha German	OS, ZW	Romania, Bulgaria
Dialect of Gottschee (Kočevje)	ZW, MV	Slovenia
German in northern Italy (“Mochenisch”): Val Fersina/Fersental	MV	Italy
Emigrant German in Israel	IS, ISW, ISZ	Israel

2.1 Speech island corpora

The corpora we describe in this section were designed as speech islands corpora from the outset, i.e. their primary purpose is to document one or a few specific varieties of “German Abroad”.

2.1.1 AD – Australiendeutsch

The AD corpus comprises data from the German speech islands in Australia recorded by Michael Clyne from Monash University in Melbourne in the 1960s and 1970s (cf. Table 2). As this corpus is further discussed in Section 3, we limit ourselves here to an overview of the data (see also Wagener & Bausch (1997, pp. 206-207)).

The corpus is structured in four sub corpora[2], which are published and accessible via the DGD:

1. AD(S) – Monash Corpus of South Australian German – Barossa Valley
2. AD(V) – Monash Corpus of Australian German – Western District of Victoria
3. AD(W) – Monash Corpus of Australian German – The Wimmera
4. AD(P) – Monash Corpus of German Pre-War-Speakers (Melbourne)

The recording situations were speech-biographic interviews including picture descriptions, and conversational speech. Transcripts are partially available. Metadata for all sub corpora are relatively sparse, but should be sufficient for reliable analysis of the data.

Table 2: Australian Speech island data in the AGD and DGD

Sub corp.	Year	Number of Recordings in AGD	Tr.	Avail. in DGD
AD(S)	1967 + 1976	50 of 51 (tape, CD – digitized), plus 2 rec. from Peter Paul (tape, digitized, not edited)	34	46 rec.
AD(V)	1966 + 1970	55 of 56 (tape, CD – digitized)	41	54 rec.
AD(W)	1969 + 1972 + 1973	91 (tape, CD – digitized)	81	91 rec.
AD(P)	1969	29 of 33 (CD – digitized)	29	29 rec.

2.1.2 BA – Brasiliendeutsche Mundarten

[2] Three more sub-corpora from Melbourne (recorded 1995-1998) have been exempted from curation so far, because they are made up for a larger part of language material other than German. However, we continue to archive the audio files and transcripts at the AGD. These are:

- AD(DG) – Monash Corpus of Australian German Third Generation Project, (13 recordings from 1995-1996 ; no transcripts ; 4 recordings are corrupt).
- AD(UE) – Trilingual Corpus Hungarian, German and English (17 recordings from 1996; 37 transcripts). Information received from Australia suggest that the original number of recordings was 38.
- AD(NE) – Trilingual Corpus Dutch, German and English (32 recordings from 1995-1998 ; 36 transcripts). Here too, documentation speaks of 38 recording in the original collection.

The BA corpus documents Swabian dialects as spoken by speakers in Panambi (formerly Neu-Württemberg) in Rio Grande do Sul in South Brazil. The corpus was compiled by Ute Bärnert-Fürst (then affiliated to the Universities of Mannheim and Campinas/Brazil) in 1985 in preparation of her master's thesis and consists of audio recordings of narratives by and interviews with speakers of three different generations. Besides German, the recordings, especially with the youngest generation who have reduced fluency in German, contain a substantial amount of Brazilian Portuguese.

The AGD has at its disposal altogether 18 audio recordings from this corpus, totaling 23:33 hours. The original reel-to-reel tapes were copied to DAT tapes by the AGD and returned to Ute Bärnert-Fürst. Metadata on speakers and recording situations is fragmentary, no transcripts are available, although Wagener & Bausch (1997, pp. 144) mention transcripts in literal notation made by Ute Bärnert-Fürst. Since data protection issues are waiting to be cleared, and this is complicated by the fact that larger portions of the data are in Portuguese, the corpus is currently not available from the official AGD platforms. Further information on the corpus can be found in Wagener & Bausch (1997, pp. 144-145) and Bärnert-Fürst (1994).

2.1.3 RU – Rumäniendeutsche Mundarten

The RU corpus documents German dialects spoken in Siebenbürgen (Transsylvania) in Romania ("Siebenbürgersächsisch"). The recordings were done by the Institute of Linguistics of the University of Bukarest, the "Arbeitsstelle Siebenbürgisch-Sächsisches Wörterbuch" in Hermannstadt/Sibiu and the "Institut für Folkloristik" in Klausenburg/Cluj-Napoca with Ruth Kisch, Heinrich Mantsch, Helga Stein and Hanni Kirschlager as principal investigators (Wagener & Bausch 1997: 69). The corpus was compiled by the "Heimathaus Siebenbürgen" in Gundelsheim/Neckar and contains recordings of narratives, conversations, fairy tales, elicited speech, partially also of music and chant of German speakers from Bukarest (1966-1975), Hermannstadt (1962) and Klausenburg (1960-1981).

Analogue versions of the recordings were transferred from Gundelsheim to the IDS in 1994 for digitization. One set of digital copies remained at the IDS while the original tapes and another digital copy were returned to Gundelsheim. In 2008, a part of the data (the Wenker sentences) were transferred from the IDS to Marburg for integration into the Sprachatlas. In 2009, the documentation center "Siebenbürgische Bibliothek mit Archiv" in Gundelsheim transferred digital copies of the recordings to the Ludwig-Maximilians-Universität in Munich where a project headed by Thomas Krefeld has made them available - including detailed metadata and transcriptions - via the platform of the "Audioatlas Siebenbürgisch-Sächsicher Dialekte" (ASD at http://www.asd.gwi.uni-muenchen.de/ - see also Krefeld et al. 2015). Since 2016, the corpus is also listed among the resources hosted by the Bavarian Archive for Speech Signals (BAS: http://www.bas.uni-muenchen.de/forschung/Bas/BasASDeng.html).

The AGD continues to store the 2275 audio files (around 360h) resulting from the digitization effort in the 1990s. Since, however, the resource has undergone substantial curation work at LMU Munich, we consider the ASD platform and BAS the primary contacts for researchers interested in this data and are not planning any further work on this corpus.

2.1.4 NO – Deutsch in Nordamerika

The title of "NO - Deutsch in Nordamerika" actually subsumes two disparate collections of recordings of German speaking communities (Mennonite Plautdietsch, Volga German, Volhynia German, Pennsylvania Dutch) in Kansas and Nebraska. The older collection can be traced back to the "Robert Buchheit Low German Collection" of the Mennonite Archive at Bethel College in Newton/Kansas (Buchheit, 1982) and was acquired by the IDS in the 1990s. It contains 17 recordings of narratives and different elicitation tasks (word lists, test sentences, rhymes and sayings, translations), probably made around 1978 in diverse locations in Kansas. The second collection stems from a project undertaken by Peter Wagener (then head of the archive at the IDS) in 1995 and contains 32 recordings of narratives and recitals of poems by speakers in diverse locations in Kansas and Nebraska.

All recordings were digitized at the AGD and are currently stored on DAT tapes. Metadata for the older collection are very sparse, while Wagener's collection has been thoroughly documented according to the archive's standards of the time. Conditions for distribution of these data are unclear. The AGD plans to integrate digital copies of the DAT tapes into the regular archive and then reassess quality of recordings and options for further curation and distribution of the data, possibly including a split of the "corpus" into two separate entities.

2.1.5 RS – Russlanddeutsche Mundarten

The RS corpus documents German dialects (mostly Low German and Franconian) spoken in the Altai region in Siberia. The recordings were made by Hugo Jedig in 1960 in his effort to document the language of German speech communities in the Soviet Union (Berend and Jedig, 1991; Berend, 1998; Wagener & Bausch, 1997: 123-124). They contain mostly narratives about rural work and everyday life.

Copies of the original reel-to-reel tapes were secured for the AGD by Nina Berend around 1994 and digitized around 2010, resulting in altogether 106 audio files with a total duration of roughly 7 hours. Rudimentary metadata about time and place of the recordings and the speakers' names are available. Owing in part to the deficient recording equipment, the quality of the recordings is in general rather poor. Since no explicit information about speakers' consent is available, the legal conditions for distribution of these data are somewhat unclear. For these reasons, the corpus is currently archived at the AGD, but not made officially available through the DGD or the archive's personal service. A part of the data has been reused for the database RuDiDat (cf. following section).

2.1.6 RUDI – Russlanddeutsche Dialekte

RUDI is a corpus of Russian German dialects which constitutes the basis for the Database of Russian German Dialects (RuDiDat[3]), a freely accessible search tool based on word indices containing historical and authentic audio recordings of interviews and narratives as well as private memories by speakers from the German speech islands in the former states of the Soviet Union.

[3] RuDiDat: http://prowiki.ids-mannheim.de/bin/view/Russlanddeutsch/WebHome

The corpus was compiled by Nina Berend and published via the DGD in May 2018. It contains data from 7 Russian German dialects: Swabian (recorded in Pawlodar), Bavarian, South Franconian, Low German and Palatin (recorded in the Altai region), Hessian (recorded in Omsk) and Volhynia German (recorded in Omsk and Koktschetaw). Altogether, there are 10 hours of audio recordings from 20 speakers, recorded in the years between 1959 and 1989. Detailed transcripts in modified orthography with orthographic normalisation, lemmatisation and POS tagging are available for the entirety of the corpus.

2.1.7 SL – Deutsch in der Slowakei

The SL corpus documents the "Karpatendeutsch" variety (Carpathian German - a Bavarian dialect) spoken in three speech islands: the Zips region and the regions around Kremnica and Bratislava in Slovakia. The recordings were made between 1965 and 1968 in a project headed by Helmut Protze (see Protze, 1965, 2006), then a member of the "Zentralinstitut für Sprachwissenschaften" at the Berlin Academy of Sciences in the former GDR. They consist of narratives, conversations and elicited speech including some recordings of Wenker sentences.

The AGD received the data from the Zentralinstitut in the early 1990s and now has as its disposal an estimated number of 250 recordings on 42 reel-to-reel tapes. Raw digitization of the tapes was completed in early 2018. Metadata on paper are available and will also be transferred to digital form as soon as the archive's capacities allow. It should then be possible to assess if and how the corpus can be made available.

2.1.8 OZ – Deutsch in Ozeanien

The corpus OZ assembles two disparate datasets documenting German-based contact varieties spoken in Oceania. Two recordings (1:35h total duration) made by Peter Mühlhäusler in 1972 and 1975 document seven speakers of Ali Pidgin (Mühlhäusler, 1977 and 1984). 9 recordings (5:42h total duration) made by Craig Volker between 1978 and 1980 document different speakers of Unserdeutsch (Volker, 1989; see also section 3.2 below). The recordings contain narratives and word lists in the respective varieties, with a fair amount of mixing with standard German, English and Tok Pisin. Compact cassettes of the recordings were obtained by Stefan Engelberg and Doris Stolberg in the context of the IDS project "Lexikalischer Wandel unter deutsch-kolonialer Herrschaft - Deutsch in der Südsee". The AGD has digitized these recordings and collected further information, but only little metadata was available for them. The corpus is now distributed through the archive's personal service.

2.1.9 MEND – Mennonitendeutsch

The corpus MEND consists of the data on Mennonite communities in Mexico (Cuauhtemoc/Chihuahua), Paraguay (Fernheim/Filadelfia, Loma Plata), USA (Seminole), Brasil (Colonia Nova) and Bolivia (Santa Cruz) compiled by Göz Kaufmann for his habilitation thesis (Kaufmann, 2007) in the first years of the new millennium. All recordings contain the same type of elicited data, namely translations of a set of 46 sentences (given to the speakers in Spanish, English or Portuguese) into Mennonite Low German ("Plautdietsch"). Altogether, 321 speakers were recorded, resulting in audio material with a total duration of around 40h. Since the end of 2016, Göz Kaufmann has been supporting the AGD in the curation of these data. Digitization of the tapes was completed in summer 2017 and the alignment of existing transcripts with the audio was completed in summer 2018. Metadata for the corpus is currently being transformed to the archive's standards. Given the thorough documentation of the corpus and the continuous support of the corpus compiler, we expect to be able to publish a first version of the corpus in the DGD in the course of 2018.

2.1.10 AF – Deutsch in Afrika

The corpus AF documents German spoken in Namibia. The recordings were made by Peter Wagener in 2005 in Ojiwarongo and Windhoek and contain narrative monologues. According to the available corpus documentation, the seven recordings now archived as digital audio files at the AGD (total duration: 90 minutes) were meant to be a first contribution to a larger collection. However, no traces of further recordings could be found.

Metadata on the recording situation and on the speakers are available as well as text transcripts. Speakers' consent for publication of the data is partly documented, partly unclear. In principle, conditions for further curation of the corpus and its eventual integration into the personal archive service are thus favorable.

2.1.11 NZ – Deutsch in Neuseeland

In 2016, Stefan Engelberg secured a set of recordings of the German variety of Puhoi in New Zealand, from Gerard Straka, the son of one of the speakers documented in the recordings. There is a possible overlap between these data and a New Zealand recording in the MV corpus (see below). The recordings contain elicited data and free speech and amount to a total duration of roughly three hours.

The data was submitted to the AGD as 25 audio WAV files. In the context of the IDS' work on German in the Pacific region, we are currently exploring options for integrating these data with another 11h of data on New Zealand German as stored in a collection at the University Library of Auckland (filed there as the "Droescher collection of material relating to Puhoi, 1957-1963" and obtained by the AGD in early 2018) and possibly further related data from other sources.

2.1.12 IS, ISW and ISZ – Emigrantendeutsch in Israel

The corpora IS (Emigrantendeutsch in Israel), ISW (Emigrantendeutsch in Israel: Wiener in Jerusalem) and ISZ (Zweite Generation deutscher Migranten in Israel) were all compiled between 1988 and 2012 in different research projects headed by Anne Betten (cf. Betten, 1995; Betten & Du-nour, 2000; Betten, 2013). The IS corpus documents the German of Jews some fifty or sixty years after they migrated from German speaking regions in Europe to Israel, ISW does the same with a special focus on Austrian emigrants from Vienna, and ISZ complements the other two corpora with recordings of the German of the second generation, mostly the children of speakers recorded in IS and ISW. By far the largest part of the recordings are auto-biographic interviews covering a large spectrum of cultural, political and personal topics.

Starting with IS in the 1990s, the data were given to the AGD in several steps by Anne Betten who has remained in close contact with the archive to collaborate on curation and further work on the corpora. Currently, all digitized audio recordings and a few digital videos of the corpora are made available through the DGD, and their documentation can be considered fairly complete. A substantial part of the data has also been transcribed, and transcripts which have undergone sufficient quality checks are also made available through the DGD (the others are stored as "raw" transcripts in the archive and can be made available upon request to Anne Betten via the personal service of the AGD). With altogether close to 500 hours of recordings and almost half a million transcribed tokens available in the DGD (for details, see Table 3), the IS corpora count among the largest of the archive. Curation of the data is ongoing as far as the archive's capacities allow. Most recently, we have completed time-aligned XML versions of the ISW transcripts and published them in version 2.10 of the DGD thus improving possibilities for corpus linguistic analyses.

Table 3: Overview of the corpora IS, ISW and ISZ.

Corpus	Recordings	Transcripts
IS	181, roughly 285h	22 in DGD, about 250.000 tokens 83 raw (not in DGD)
ISW	28, roughly 50h	20 in DGD, about 220.000 tokens 4 raw (not in DGD)
ISZ	86, roughly 130h	64 raw (not in DGD)

2.1.13 Current activities: Further corpora

The archive is currently negotiating the acquisition of further corpora of German abroad. Collaborations with Peter Maitz' Unserdeutsch project and Heike Wiese's and Horst Simon's Namdeutsch project have been started in 2015 and 2017, respectively (see Section 3.2), and the corpora compiled in these projects will be integrated into the AGD after completion. Further negotiations concern data of Volga German in Argentina (Ladilova, 2013), of Pennsylvania Dutch in the US (PEND, recorded by Ludwig Eichinger in 1994 and 1999, cf. Eichinger, 1997, p.175) and of speech islands in German-Bohemian settlements in the Ukraine, Romania, USA and New Zealand (recorded by Alfred Wildfeuer, Wildfeuer, 2017, and Nicole Eller-Wildfeuer, Eller-Wildfeuer, 2017).

2.2 Corpora containing speech island data

2.2.1 MV – Binnen- und auslandsdeutsche Mundarten: Varia

As the term "Varia" in the name of the MV corpus suggests, MV is not a homogeneous corpus, but rather a collection of recordings of German speech varieties from within Germany and from outside Germany. The data originate from various external sources. Besides data on dialects in Germany and neighboring countries, MV also contains data from 9 speech islands. These data are already included in the DGD and documented according to AGD standards (cf. Section 2.2.1.1). For another 13 speech islands (cf. Section 2.2.1.2), the documentation of the recordings, as well as the metadata, was only partially available to the archive and the sound quality may sometimes be quite poor. We are currently reassessing and curating these data and expect to be able to make a part of them available via the DGD in late 2018.

2.2.1.1 MV – speech island data in the DGD

The speech island data in the MV corpus that made it into the DGD were recorded at various places: the USA, Canada, Mexico, Australia, FR Germany in the years 1964-1973. The recording situations are mainly initiated narratives.

Altogether, there are 9 speech islands contained in the MV corpus available via the DGD. Some transcripts can be found in PHONAI, but no transcripts have been digitized for the DGD yet.

1. Mennonite German from South Russia / Ukraine – Canada / British Columbia. 24 recordings from 1965/66 by Wolfgang W. Moelleken. 8 transcripts were published in PHONAI Vol.10 (Moelleken, 1967; Moelleken, 1972).
2. Siedlungspfälzisch (Palatine German) – Canada / Ontario. 2 recordings from 1969 by Wolfgang W. Moelleken. 2 transcripts were published in PHONAI Vol. 18 (Karch & Moelleken, 1977)
3. Low German dialects and other North American German in Wisconsin – USA (Eichhoff, 1985; Eichhoff, 1996; Eichhoff, 2003). 64 recordings from 1968 (plus 55 recordings not in the DGD) by Jürgen Eichhoff. No transcripts.
4. Dane County Kölsch – USA / Wisconsin. 2 recordings (one of them a double recording with two speakers) from 1973 by Peter A. McGraw. 1 transcript in PHONAI Vol. 21 (McGraw, 1979)
5. Mennonite German from South Russia / Ukraine – Mexico (Moelleken, 1987). 4 recordings from 1966 by Wolfgang W. Moelleken. No transcripts.
6. Australian German from Barossa Valley, Kapunda – Australia / South Australia. 3 recordings from 1964 by Peter Paul. 3 transcripts in PHONAI Vol. 6 (Paul, 1970), cf. also Paul (1965).
7. Banat / Stefansfeld – Serbia; 9 recordings made in Chicago/USA by Ronald N. Werth in 1972. No transcripts.
8. Banat / Neubeschenowa – Romania; 1 double recording with two speakers made in Köln/Germany by Peter Wanko in 1974. No transcripts.
9. Siebenbürgen / Brenndorf near Kronstadt (Braşov) – Romania; 1 recording made in Wolfenbüttel/Germany by Hilde Heuer in 1970. 1 Transcript (not in DGD, not digitized).

2.2.1.2 MV – speech island data in the archive

Some of the speech island data in MV are not yet available via the DGD, but have recently been added to the archive’s work plan for corpus curation and should be available (with the exception of the Brazilian data) in the DGD by the end of 2018.
This concerns the following 13 speech islands:

1. Mennonite German – Poland Weichsel-Delta. Principal investigator Wolfgang W. Moelleken. 3 recordings (digitized) from 1981. Metadata available in DSAv standard.
2. Mochenisch dialect of the Fersental – Italy, Trentino (Anthony R. Rowley, Regensburg). 3

recordings (digitized) from 1976. Metadata available in DSAv standard. 3 transcripts published in PHONAI Vol. 31 (Rowley, 1986)

3. Gottscheer dialect of Kočevje – Slovenia (Günter Lipold, Phonogram Archive Vienna). 21 recordings (copies from Vienna; no digitization planned) from 1974-75. Metadata copied from Vienna. 7 transcripts published in PHONAI Vol. 26 (Lipold, 1984).
4. Volhynia German – Canada (Prof. Reinhold, Edmonton / University of Alberta / Canada). 2 recordings (digitized). The recording year is unknown. No transcripts.
5. Mennonite Low German – Canada. Principal investigator Grace E. Wiebe (Wiebe, 1983). 1 recording in 1983 of 17:50 minutes (digitized). No transcript.
6. Mennonite German from South Russia / Ukraine. 3 recordings of 8 speakers made in Canada / Mexico by Dr. Jack Thiessen (Winnipeg) in 1966 (not digitized, digitization planned). Metadata available in DSAv standard. No transcripts.
7. Pennsylvania German – USA. 27 recordings by Prof. Ralph Charles Wood (Philadelphia / Pennsylvania) in 1954-59 (digitized). Metadata is partially available. Phonetic transcripts are partially available (not digitized).
8. Standard German – South Africa (Frl. H. Stielau, Durban). 1 recording of four speakers (digitized). Few metadata available: Two speakers born in southwest Africa; two speakers born in Germany. The recording year is unknown.
9. Queensland Speech Survey – Australia, Queensland (Elwyn H. Flint, Brisbane). 9 recordings of several speaker groups in 1964 (digitized). Languages used are German and English. Metadata for speakers are partially available. No transcripts.
10. Black Sea Swabian (Schwarzmeer-Schwäbisch, Lustdorf) – Australia (Anatole Bond, University of Queensland, Brisbane). 1 recording from 1966 (digitized). Metadata and transcript in Bond (1978).
11. Barossa Valley dialectal samples Silesian – Australia (Peter Paul, Adelaide). 1 recording from 1962 (digitized). Speaker-data and transcriptions available in bachelor thesis of Peter Paul (Paul, 1962).
12. Puhoi German, Egerland dialect – New Zealand (Werner O. Droescher, Auckland). 3 recordings from 1967 (digitized). Phonetic analysis of one speaker published in PHONAI Vol.15 (Droescher, 1974). No transcripts.
13. Colonist German (Kolonistendeutsch: Hunsrückisch, Pommersch Platt, Westfälisch Platt, Mennonite German) – Brasil, Paraguay. 141 recordings (digitized) by Erich Fausel (Reutlingen) in 1956-63 (Fausel, 1959). The recordings contain Wenker sentences. Metadata on recordings and speakers are available in hand-written form.

2.2.2 OS – Deutsche Mundarten: Ehemalige deutsche Ostgebiete

In the OS corpus, there are several recordings from German speech islands in middle-eastern, eastern and south-eastern Europe (Bellmann & Göschel, 1970). We describe here those recordings of OS that stem from outside the contiguous German speaking area. Principal investigators were the Deutsches Spracharchiv, then in Münster/Braunschweig, and the Deutscher Sprachatlas in Marburg. The recordings took place in the Federal Republic of Germany and in Austria in the years 1962 to 1965. The recording situations are narratives and Wenker sentences. Transcripts are only available for speech islands in southeast Europe

(Romania, Serbia, Croatia) and for some recordings from Poland, Estonia and Latvia.

The recordings listed below comprise 349 of 981 recordings in the entire OS corpus. All have been digitized from tape and are available in the DGD.

1. Baltic States: Estonia, Latvia, Russia: 18 recordings and 2 transcripts on DGD.
2. Poland: Northwest Poland, Northeast Poland, Central Poland, Galicia: 44 recordings and 3 transcripts in the DGD.
3. Russia: Volga: 2 recordings
4. Ukraine: Volhynia, Galicia: 35 recordings.
5. Ukraine: Mennonite German in the Don and Black Sea regions: 21 recordings.
6. Ukraine/Moldova: Bessarabia: 35 recordings.
7. Slovakia/Ukraine: Zips, Bratislava, Karpaten: 4 recordings.
8. Hungary: North Hungary: 3 recordings.
9. Hungary: South Hungary / Batschka, Baranya amongst others: 3 recordings.
10. Croatia: Slavonia / Bosnia: 5 recordings and 1 transcript on DGD.
11. Serbia/Croatia: Syrmia: 8 recordings.
12. Serbia: Batschka: 30 recordings, 1 transcript on DGD.
13. Serbia: Banat: 19 recordings and 1 transcript on DGD; 1 transcript in PHONAI Vol.9 (Grubačić, 1971); 7 literal transcripts (not digitized).
14. Romania: Banat: 28 recordings and 5 transcripts on DGD and 5 literal transcripts (not digitized).
15. Romania: Siebenbürgen (Transsylvania): 67 recordings and 14 transcripts on DGD and 22 literal transcripts (not digitized).
16. Ukraine/Romania: Bukovina: 17 recordings.
17. Romania: Dobrudscha: 8 recordings.

2.2.3 ZW – Deutsche Mundarten: Zwirner-Korpus

The entire ZW corpus comprises 5796 recordings. Thereof, 378 recordings can be classified as speech island data from central-eastern, eastern and south-eastern Europe. Principal investigators vary, however the main director of the project was Eberhard Zwirner (Deutsches Spracharchiv Münster/Braunschweig). The speech island recordings took place in the Federal Republic of Germany in the years 1955 - 1959, 1961, 1964 and 1972. The recording situation was narratives, partially dialogue recordings with 2 speakers. Transcripts are only available for speech islands in south-eastern Europe (Romania, Serbia, Croatia, Hungary), and for some recordings from Poland and the Ukraine.

All recordings have been digitized from tape and are available on the DGD.

1. Baltic states: Estonia, Latvia: 12 recordings.
2. Lithuania: 3 recordings.
3. Poland: Northeast Poland, Central Poland, Galicia: 15 recordings and 3 transcripts on DGD.
4. Ukraine: Volhynia, Galicia, Bukovina: 19 recordings and 3 literal transcripts (not digitized).
5. Ukraine: 2 recordings.
6. Ukraine/Moldova: Bessarabia: 26 recordings and 9 literal transcripts (not dig.).

7. Slovakia: Zips: 10 recordings and 1 literal transcript (not dig.).
8. Slovakia: Bratislava, Kremnitz: 11 recordings.
9. Hungary: North Hungary, Northwest Hungary: 28 recordings and 5 literal transcripts (not dig.).
10. Hungary: Central Hungary: 25 recordings and 5 literal transcripts (not dig.).
11. Hungary: South Hungary/Batschka, Baranya, Banat and others: 60 recordings and 2 transcripts on DGD and 8 literal transcripts (not dig.).
12. Slovenia: Gottschee: 1 recording.
13. Croatia: Slavonia/Bosnia: 11 recordings and 2 transcripts on DGD and 1 transcript published in PHONAI Vol.19 (Popadić, 1978) and 2 literal transcripts (not dig.).
14. Serbia: Syrmia: 20 recordings and 2 transcripts on DGD and 5 literal transcripts (not dig.).
15. Serbia: Batschka: 65 recordings and 5 transcripts on DGD and 1 transcript published in PHONAI Vol.15 (Geršić, 1974) and 2 literal transcripts (not dig.).
16. Serbia: Banat: 22 recordings and 1 transcript published in PHONAI Vol.6 (Grubačić, 1970) and 2 literal transcripts (not dig.).
17. Romania: Banat: 9 recordings and 2 transcripts on DGD.
18. Romania: Siebenbürgen (Transsylvania): 10 recordings.
19. Romania: Bukovina: 16 recordings.
20. Romania: Dobrudscha: 5 recordings.
21. Serbia: Banat/Stefansfeld: 8 recordings.

2.2.4 SV – Deutsche Mundarten: Südwestdeutschland und Vorarlberg

The SV corpus (Ruoff, 1973; Fiess, 1975) contains speech island data from Bessarabia (Ukraine/Moldova). The principal investigator was Arno Ruoff ("Tübinger Arbeitsstelle Sprache in Südwestdeutschland"/Deutsches Spracharchiv, Tübingen). The recordings took place in Württemberg in the FR of Germany in 1966-1967. The entire SV corpus comprises 242 recordings, 25 of which are from speech islands. The recording situation is narratives, and speakers were selected from various generations. No transcripts are currently available in the archive or in the DGD. Some phonetic transcripts are available via the Ludwig Uhland Institut in Tübingen. The project "Sprachalltag II: Sprachatlas - Digitalisierung - Nachhaltigkeit" in Tübingen is currently creating transcripts for all recordings in SV. When the project is completed, these will also be integrated in the DGD.

2.2.5 ZWTV – Korpus Zwirner – Sprachproduktion im Tonstudio/ Varia

The ZWTV corpus contains studio recordings from speech islands, namely Baltic German, Galicia, Bessarabia[4], Pennsylvania and Mennonite German. The principal investigators were Eberhard Zwirner, Wolfgang Bethge (Deutsches Spracharchiv Münster/Braunschweig), Edeltraud Knetschke (Deutsches Spracharchiv at IDS Mannheim) and Julius Krämer (Pfälzisches Wörterbuch Kaiserslautern). The recordings were made in Braunschweig, Lüneburg, Klein

[4] Speakers from the Baltic States, Galcia and Bessarabia became German residents after 1945 and could therefore be recorded in the studio there.

Hehlen near Celle and Mannheim, all in the Federal Republic of Germany in the years 1955-1957 and 1987.

The recording situation was narratives and partially dialogues. All 17 recordings of speakers from speech islands were digitized from tape. There are 2 transcripts for 2 of the recordings (not digitized). One of them was published in the Lautbibliothek Deutscher Mundarten (LDM) Vol.9 (Laur, 1958). Neither the recordings nor the transcripts are available via the DGD or the personal service of the AGD.

1. Baltic states/Riga: 2 recordings from 1955 and 1956. 1 transcript published in LDM Vol.9 (not digitized).
2. Ukraine/Galicia: 8 recordings from 1955 and 1956. 1 transcript (not digitized).
3. Ukraine/Bessarabia: 1 recording from ca. 1956.
4. USA: Pennsylvania German: 1 recording from 1956; speaker was guest of the DSAv.
5. USA: Low German: 2 recordings with 4 emigrated speakers born in North-Lower Saxony; recording year 1956.
6. USA: Minnesota/Iowa: 1 recording from 1955; speaker was guest of the DSAv.
7. South Africa: 1 recording from 1956; speaker was guest of the DSAv.
8. Uruguay/Russia (Altai region): Mennonite German: 1 recording in 1987 with two speakers.

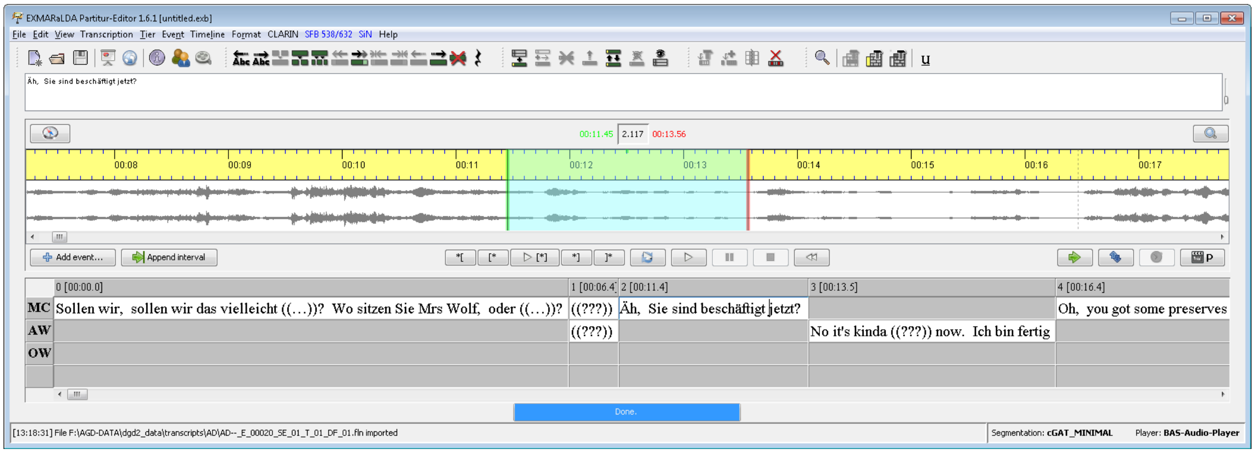
3 Data curation

The tasks involved in data curation can be manifold. They depend on the data that researchers or research projects transfer to the archive. Sometimes, projects approach the archive already at the planning stage in order to secure a place for the data to be archived and for technical support and advice on various aspects of data collection: from planning audio or video recordings, to legal aspects such as consent forms and data protection and metadata forms transcription tools and conventions, and data storage and backups. Such a workflow is exemplified in section 3.2. It has advantages for both the research project in order to run smoothly as well as for the archive in order to avoid the work expenses that are involved when data are transferred in idiosyncratic or fragmented structures. The latter can for example happen with legacy data as they arrived at the IDS from Craig Volker on German in Oceania or, as discussed in the following section, from Michael Clyne on Australian German.

3.1 Curation of legacy data: example Australian German

The “Monash Corpus of Australian German”, as described above (section 2.1 “AD”), was compiled by Michael Clyne between 1966 and 1973. As early as 1965, Michael Clyne had, in preparation of the project, attempted to establish a close collaboration with the archive (then called the “Deutsches Spracharchiv – DSAv”, and not yet part of the IDS) with the aim of integrating the recordings into its inventory. However, the plan was not realized at the time because the format of the recordings (more specifically: the tape speed) did not conform to the archive's requirements and its rather narrow focus on specific types of phonetic analyses. More than 40 years later, around 2008, Michael Clyne, then already officially retired, renewed his offer to donate the corpus to the archive, and this time, the AGD accepted. Sadly, the transfer of the material could not be accomplished before Michael Clyne's unexpected death in 2010. The IDS then contacted his family and former colleagues at Monash University who sent a first box with digitized recordings on CD to Mannheim in mid-2012. After closer inspection, it turned out that these data were incomplete (i.e. a substantial number of recordings was missing) and their documentation (basically not more than a code written on the CD covers) too sparse for a reliable curation. The AGD therefore set out on a quest to retrieve not only the original recordings, but also any other accompanying material such as transcripts or documentation of speakers. Luckily, Claudia Riehl, who had worked with the data before, was able to provide contact details of three former doctoral students of Michael Clyne who, after some detective work of their own, located the original tapes at Monash University and shipped them to Mannheim in early 2013, accompanied by a fairly large collection of transcripts in MS Word format. Only then could the actual curation of the data begin. The effort comprised the following steps:

- The reel-to-reel tapes were digitized and compared to the corresponding files on CD, if they existed. In all but a few cases, the newly digitized files turned out to be of better quality than the old ones.
- By triangulating between information in existing (published) descriptions of the corpus, information recorded on tape or CD covers and the transcripts, and information in the recordings themselves, a definite version of the corpus was established. This involved cutting and merging recordings as appropriate, assigning them to one of the four sub-corpora and establishing matching pairs of recordings and transcripts.
- Metadata on speakers and interviews were extracted, again using all available information sources, in particular biographic details mentioned by the speakers themselves in the recordings. These metadata were then represented in XML files according to the archive's standards.
- The MS Word versions of the transcripts were first transformed to plain text files and then converted to the XML format of the EXMARaLDA system. In that form, they were ready for alignment with the audio recordings, using the EXMARaLDA Partitur-Editor. While the transformation step could be accomplished automatically, the alignment had to be carried out manually by student assistants as shown in Figure 1. In the process of alignment, we also closed larger gaps in the transcriptions – most importantly the interviewers' utterances which had been transcribed only fragmentarily so far.



EXMARaLDA Partitur-Editor 1.6.1 [untitled.exb]
File Edit View Transcription Tier Event Timeline Format CLARIN SFB 538/632 SiN Help
00:11.45
2.117
00:13.56
00:08
00:09
00:10
00:11
00:12
00:13
00:14
00:15
00:16
00:17
Add event...
Append interval
Done.

Äh, Sie sind beschäftigt jetzt?

	0 [00:00.0]	1 [00:06.4]	2 [00:11.4]	3 [00:13.5]	4 [00:16.4]
MC	Sollen wir, sollen wir das vielleicht ((…))? Wo sitzen Sie Mrs Wolf, oder ((…))?	((???))	Äh, Sie sind beschäftigt jetzt?		Oh, you got some preserves
AW		((???))		No it's kinda ((???)) now. Ich bin fertig	
OW					

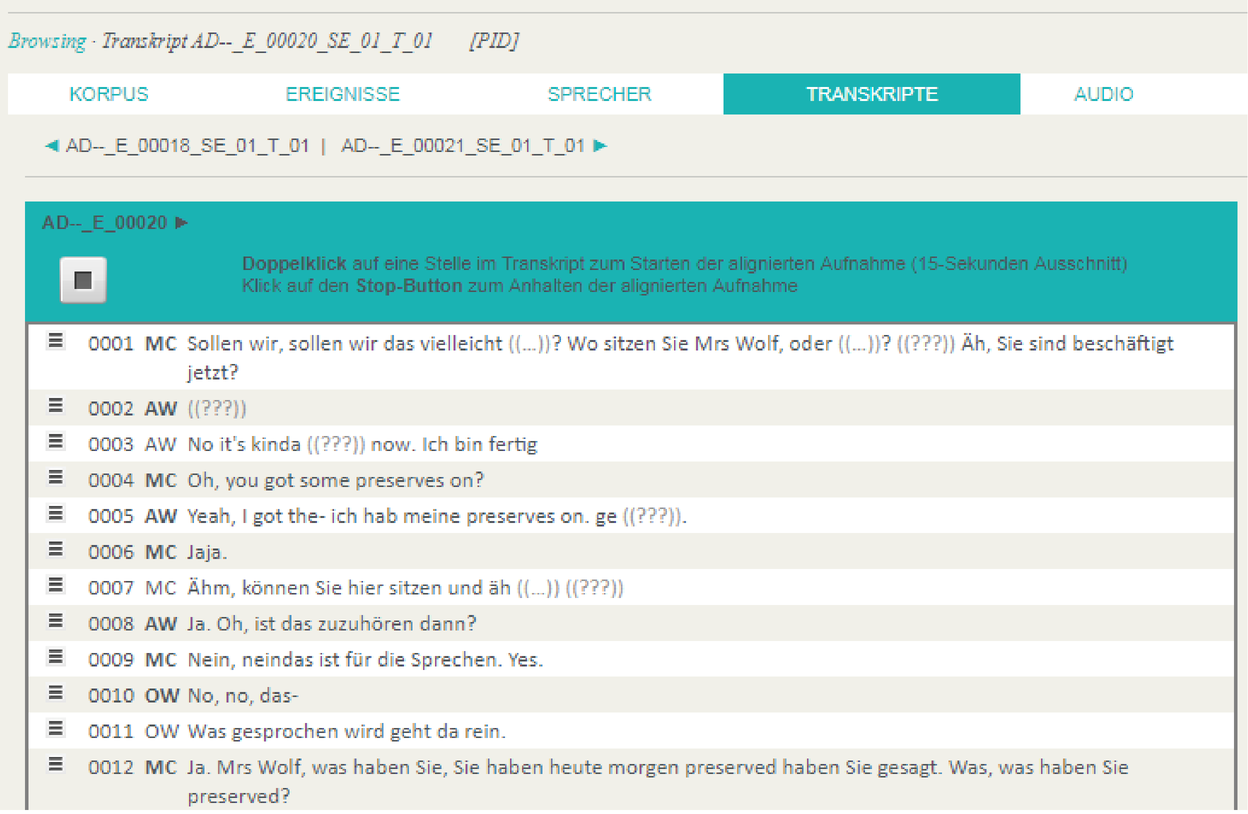
[13:18:31] File F:\AGD-DATA\dgd2_data\transcripts\AD\AD--_E_00020_SE_01_T_01_DF_01.fln imported

Segmentation: cGAT_MINIMAL Player: BAS-Audio-Player

Figure 1: Aligned transcript and recording in the EXMARaLDA Partitur-Editor.

- Recordings and transcripts were checked for mentions of person names and other details that would allow direct identification of speakers. Whenever such a mention was found, the corresponding part of the recording was replaced with a Brownian noise, and the name in the transcript replaced with a pseudonym, e.g. “Mrs. Wolf” in Figure 1. This type of masking of personal information was considered necessary because, although participants had given their consent (documented mostly on the tapes themselves) to the recordings and their use for scientific purposes, they had of course not foreseen that the data might be made available to a larger audience through an internet platform. To protect the participant's privacy (and to comply with the legal requirement of “data parsimony”) we therefore judged it necessary to remove information that would allow a direct identification of speakers.
- Aligned transcripts were again transformed from the EXMARaLDA format to the XML format currently used as the standard inside the archive. In that process, the transcripts were also tokenized and supplemented with lemmatization and Part-Of-Speech information on additional annotation layers. These steps could be carried out fully automatically.

With version 2.8 of the Database for Spoken German, the complete Corpus of Australian German, comprising 220 audio recordings (around 65h) and 168 aligned transcripts (around 330,000 tokens), could be made available to the scientific community in April 2017, more than 50 years after Michael Clyne had started the project. Figure 2 shows an excerpt of a transcript from the AD corpus in the current version of the DGD.


Browsing · Transkript AD--_E_00020_SE_01_T_01 [PID]
KORPUS
EREIGNISSE
SPRECHER
TRANSKRIPTE
AUDIO
◄ AD--_E_00018_SE_01_T_01 | AD--_E_00021_SE_01_T_01 ►
AD--_E_00020 ►

Doppelklick auf eine Stelle im Transkript zum Starten der alignierten Aufnahme (15-Sekunden Ausschnitt)
Klick auf den **Stop-Button** zum Anhalten der alignierten Aufnahme

0001 MC Sollen wir, sollen wir das vielleicht ((...))? Wo sitzen Sie Mrs Wolf, oder ((...))? ((???)) Äh, Sie sind beschäftigt jetzt?
0002 AW ((???))
0003 AW No it's kinda ((???)) now. Ich bin fertig
0004 MC Oh, you got some preserves on?
0005 AW Yeah, I got the- ich hab meine preserves on. ge ((???)).
0006 MC Jaja.
0007 MC Ähm, können Sie hier sitzen und äh ((...)) ((???))
0008 AW Ja. Oh, ist das zuzuhören dann?
0009 MC Nein, neindas ist für die Sprechen. Yes.
0010 OW No, no, das-
0011 OW Was gesprochen wird geht da rein.
0012 MC Ja. Mrs Wolf, was haben Sie, Sie haben heute morgen preserved haben Sie gesagt. Was, was haben Sie preserved?

Figure 2: A transcript from the Australian German corpus as presented in the Database for Spoken German. Registered DGD users can access the excerpt with audio directly via the following URL: http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=AD--_E_00020_SE_01_T_01_DF_01&cID=c1&wID=&textSize=200&contextSize=4

The process of depositing the data at the archive and compiling all the information necessary for its curation took a lot of time and detective work. For the actual curation of the data, we estimate that around four person months for technical staff at the archive and little less than 800 hours of student assistant's time had to be invested. Added to this is the administrational overhead to organize the work and the technical overhead for integrating the curated data into the database.

While we have no doubt that this investment is justified – Michael Clyne's studies on Australian German have been highly influential, there will not be another opportunity to document this variety of German, and the data have a lot of potential for further investigation – we still need to ask ourselves what can be learned from this effort to avoid unnecessary curation work in the future and to make the step from the compilation of a speech island corpus to its dissemination via the archive a less time-consuming and costly one. Ongoing collaborations, in which the archive works closely together with projects collecting and transcribing speech island data, give us the opportunity to apply what we see as the current best practice in the compilation of such resources. The next section discusses one such case in more detail.

3.2 Collaborations with ongoing corpus projects

The Rabaul Creole German (Unserdeutsch) project, headed by Péter Maitz at the University of Augsburg (see Maitz et al. 2016; Maitz, 2016 and Maitz, this volume), aims at "(i) documenting Unserdeutsch as thoroughly as possible, (ii) reconstructing and presenting the development and history of the language as accurately as possible, and (iii) systematically describing the structure of the language." (quote from the project's website[5])

To ensure that the collection and processing of the data would be done according to the technical and methodological state-of-the-art and that the resulting corpus could be easily integrated into and disseminated via the Archive for Spoken German, Péter Maitz contacted the archive already in the project's planning phase. This led to a binding agreement (documented also in the proposal to the funding agency) between the IDS and the Unserdeutsch project to collaborate closely in the construction of the corpus and to make it available to the scientific community – after a "period of grace" in which the project keeps the exclusive right to grant or deny access to the data – via the archive's platforms.

At the present point in time, the first phase of data collection has been completed. The archive has supported the project in this phase by:

- giving advice on legal issues, most importantly on procedures and forms for informed consent which ensure that the participants give valid authorization for the dissemination of recordings and transcripts for research purposes,
- proposing a scheme for organizing, documenting and naming the corpus data, compliant with the archive's storage and metadata systems,
- giving advice on technical issues for the recording situation to ensure an optimal quality of audio files for further processing,
- providing a server on which to upload raw recordings directly from the field, which is especially important when there is little technical infrastructure on location and a real risk of losing data before it can be transferred to a permanent storage,
- preparing audio data for transcription (by cutting and merging files and applying appropriate filters to improve its quality), and giving feedback on possible improvements for the next round of recordings,
- giving advice and training for transcription of the data using a tool (EXMARaLDA) which produces time-aligned transcripts that fit immediately into the archive's workflows,
- Giving advice on further annotation steps (such as normalization, PoS tagging etc.) to be carried out on completed transcripts.

The project has recently also completed the second phase of data transcription and is now embarking on the third phase in which (a) transcribed text will be supplemented with a normalized version on a second annotation layer and (b) lemmatisation and POS tagging information will be added on further annotation layers. Continuing our collaboration by providing advice and training for the ongoing work, we are confident that the project will manage to bring the corpus compilation to a timely conclusion, and that the effort to integrate the resource into the archive will be minimal, when compared to cases as the one described above.

[5] https://www.philhist.uni-augsburg.de/en/lehrstuehle/germanistik/sprachwissenschaft/rabaul_creole_german/

A similar collaboration has been initiated with the NamDeutsch project (PIs: Heike Wiese and Horst Simon, see Wiese et al. (2017)), and we aim at making this kind of early and intensive partnership between research projects and the archive the default case for obtaining new data in the future.

4 Comparative Research and Comparable Data

Speech island data, by their very nature, lend themselves to comparative or contrastive studies with other language varieties (Eichinger, 2003). The growing availability of such data in digital form opens up many new opportunities for comparison, but also brings up new methodological questions which shall be discussed in the following sections.

4.1 Comparative Research

Taking the Australian German data as an example, several comparative approaches are possible. Australian German can be viewed and analyzed under at least the following perspectives:

1. as a variety of German
2. as a variety based on Silesian varieties
3. as a variety in contact with English.

According to the first perspective, the Australian German variety of German can be compared with standard German. Here, a synchronic as well as a diachronic comparison can be envisaged. Keeping the recording years roughly constant, the Australian German data, recorded in the 1960s, could be compared with recordings of standard German from the same time period. The AGD contains such a corpus, the Pfeffer-Korpus (PF) of German colloquial speech. The colloquial speech style of the interviews made by Michael Clyne in AD and the colloquial speech recorded in PF are a further feature the two corpora would have in common. Diachronically, AD could be compared to German from today. An established and still growing corpus of contemporary, colloquial talk is FOLK (research and teaching corpus of spoken German).

According to the second perspective (Australian German as a variety based on Silesian varieties), the AD data could be compared with other data from Silesian speakers. Recordings of such speakers who were displaced after World War II can be found in dialect corpora, such as the Zwirner Corpus (ZW) and the Corpus of German Dialects from the Former Eastern Areas of Germany (OS).

According to the third context (Australian German as a variety in contact with English), the AD data could be compared with other German-to-English contact varieties. German has been in direct contact with English elsewhere in the world, for example in North America. Those speech island data are available through e.g. the MV corpus (Mundarten und Varia, see above) in the AGD containing North American speech islands, or through the recordings of the Texas German Dialect Project (TGDP, Boas et al., 2010).

4.2. Comparable Data

In order to perform comparative speech island research, it is necessary to have data from different speech islands or from the language origins in comparable structures. “Comparability” on the data level encompasses at least three distinct dimensions, which we will term “Content”, “Technology” and “Infrastructure”.

4.2.1. Content

In the content dimension, methods of *corpus design* and *data collection* are fundamental. Ideally, two comparable sets of data have a common approach to *speaker sampling* – for instance by controlling the selection of speakers for the same set(s) of sociobiographic variables such as gender, age, regional provenance or educational background. Of course, the situation of a specific speech island may make it impossible to reproduce a corpus design that worked well for another scenario. For instance, the decision taken in the Zwirner corpus to select speakers from three different age ranges (around 20, around 40, around 60) at each location may prove impossible to realize in a speech island community where only older competent speakers remain.

Similarly, inter-corpus comparisons profit from a common approach to *data elicitation*. The famous Wenker sentences (which also figure in several corpora of the AGD including the speech island data in the MV, OS and ZW collections, see above) – i.e. standard German sentences carefully selected for interesting linguistic properties given to dialect speakers for “translation” into their variety – are probably the best-known and most widely used method of eliciting directly comparable data. Kaufmann (2007) uses a similar method (though a different set of sentences) for his work on German speaking Mennonite communities in the Americas. Word lists, reading texts, naming the days of the week or the numbers from 1 to 10 are further elicitation methods that fall into the same category. Approaches which exert a less direct and strict control on the speakers’ utterances, leading to more spontaneous and authentic, but still comparable, language productions, are picture description tasks (used, for instance, in Michael Clyne’s AD data) or map tasks (used in the “Deutsch heute” corpus and quite a few other resources of spoken German, but, to our knowledge, not in speech island corpora so far). Sociobiographic interviews, finally, are the most common means (present in almost all corpora mentioned here) of eliciting speech with maximal authenticity and spontaneity. Obviously, however, such data are harder to compare because topics will vary greatly within and between corpora. Generally speaking, better comparability is “paid for” with reduced authenticity. The most reasonable approach to data collection may thus be one that mixes highly controlled elicitation methods with freer forms of talk.

Further aspects of the content dimension that have an impact on comparability are subordinate or subsequent to decisions about corpus design and data collection. For example, for one and the same data type, differences and commonalities in the details of the *recording setup* (audio or video? one microphone or several?) will also determine which comparable analyses are possible. Furthermore, corpus design and speaker sampling have to be adequately reflected in the *metadata* that is collected on-site and documented as part of the corpus. And, maybe most importantly, comparability of corpora depends crucially on decisions about *transcription and annotation* of the data. A standard orthographic transcription is a convenient means for establishing a common basis for comparison across different language varieties, but, by definition, it also levels out differences in pronunciation which may be the very object of study. Furthermore, in varieties that differ greatly from standard German, decisions about which standard orthographic form to choose for a given item can be anything but straightforward. As an example, consider a form like ‘du hat’ in Unserdeutsch, i.e. the phenomenon common in Creole languages of reduced verb morphology. Should this be transcribed as ‘du hast’ (the standard German form that would be used in this position) or as ‘du hat’ (the form respecting the grammatical system of the Creole)? Similar problems may arise with pseudo-cognates such as ‘Deern’ in Low German whose form is related to the standard German ‘Dirne’ (meaning ‘prostitute’) but whose meaning is better described by the word ‘Mädchen’ (‘girl’). From the perspective of an archive, the ideal solution is to create two interdependent transcription levels – one that is maximally faithful to the actual words uttered (i.e. a “literal” transcription in modified orthography or even a phonetic transcription in IPA), and one that maps these forms onto a standard orthographic equivalent following a documented guideline for difficult cases as the ones exemplified.[6]

4.2.2. Technology

Commonalities and differences on the content level decide on the potential of two corpora for being used in comparative research. Whether and how this potential can be realized then depends to a large extent on practical technological issues. “Technology” here encompasses, first, a series of decisions on how to represent the different parts of the corpus in digital form, more specifically:

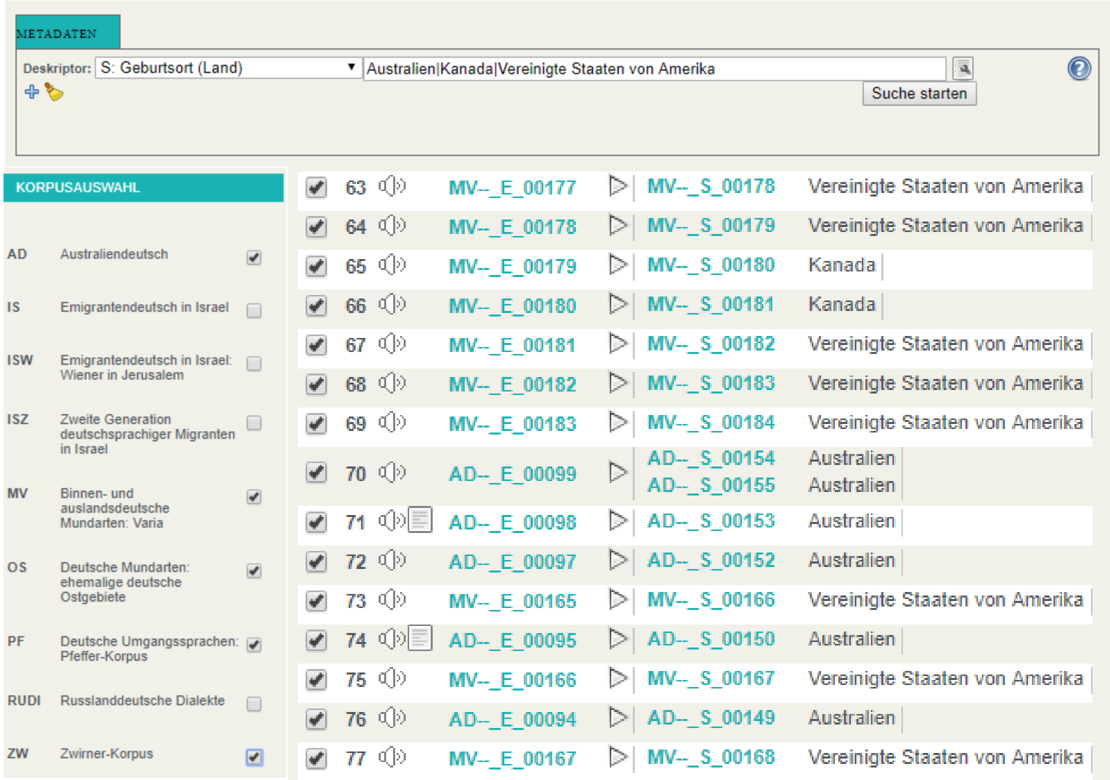
- formats (such as WAV or MPEG-4) and technical parameters (such as sampling rate, number of channels, frame rate) for audio or video recordings,
- schemas (such as IMDI or TEI) and the corresponding formats (such as various XML formats) to document metadata,
- tools (such as ELAN, Praat or EXMARaLDA) and the formats (such as plain text or, again, XML-based formats) they use to represent transcriptions and annotations.

Second, technology is also needed for accessing the data and carrying out analyses on them. This includes decisions about how a corpus is distributed (on CD, for download on a website, or integrated into a platform for online analysis), and which tools (if any) are provided to view transcripts or carry out queries on a corpus.

At the AGD, we have established a workflow for oral corpora which is based on existing official standards (such as MPEG for video recordings, ISO 24624:2016 for transcriptions of speech), de-facto standards (such as the STTS tagset for POS tagging of German) or documented best practices wherever possible. Schmidt (2016a and 2016b) discuss different parts and aspects of this workflow in more detail, DFG (2013) and Schmidt et al. (2013) document concrete recommendations for researchers building up their own corpus or

[6] Evidently, this shifts the responsibility for comparability of the data to the transcription-normalisation mapping, which is anything but trivial. Such guidelines will have to take into account the characteristics of the specific variety and thus be corpus specific to a considerable degree. Still, providing such guidelines seems to us to be a reasonable first step towards "semantic interoperability" of transcripts.

submitting a corpus to the AGD for archiving. In that way, we have achieved a relatively high degree of homogeneity at least for those AGD's speech island corpora that have already found their way into the DGD platform. Differences in corpus design and other content aspects notwithstanding, most of the data described in the previous section thus rests on a common technological basis. This can be exploited in various ways for more efficient processing of the data. For instance, since we represent all transcriptions in the same format, we can apply the same mechanisms for lemmatization and part-of-speech tagging to all corpora instead of having to adapt them for each individual case. And since metadata and annotations all adhere to the same schema, users of the DGD platform can browse and query different corpora in one and the same interface, carry out analyses involving more than one corpus, or assemble parts of different corpora into a virtual corpus (see Figure 3). Last but not least, we can offer a fair amount of interoperability for our data making sure, for instance, that all transcriptions can be converted to the most common tool formats for annotation.



Deskriptor: S: Geburtsort (Land) ▾ Australien|Kanada|Vereinigte Staaten von Amerika

AD	Australiendeutsch	☑
IS	Emigrantendeutsch in Israel	□
ISW	Emigrantendeutsch in Israel: Wiener in Jerusalem	□
ISZ	Zweite Generation deutschsprachiger Migranten in Israel	□
MV	Binnen- und auslandsdeutsche Mundarten: Varia	☑
OS	Deutsche Mundarten: ehemalige deutsche Ostgebiete	☑
PF	Deutsche Umgangssprachen: Pfeffer-Korpus	☑
RUDI	Russlanddeutsche Dialekte	□
ZW	Zwirner-Korpus	☑

☑	63	MV--_E_00177	MV--_S_00178	Vereinigte Staaten von Amerika
☑	64	MV--_E_00178	MV--_S_00179	Vereinigte Staaten von Amerika
☑	65	MV--_E_00179	MV--_S_00180	Kanada
☑	66	MV--_E_00180	MV--_S_00181	Kanada
☑	67	MV--_E_00181	MV--_S_00182	Vereinigte Staaten von Amerika
☑	68	MV--_E_00182	MV--_S_00183	Vereinigte Staaten von Amerika
☑	69	MV--_E_00183	MV--_S_00184	Vereinigte Staaten von Amerika
☑	70	AD--_E_00099	AD--_S_00154 AD--_S_00155	Australien Australien
☑	71	AD--_E_00098	AD--_S_00153	Australien
☑	72	AD--_E_00097	AD--_S_00152	Australien
☑	73	MV--_E_00165	MV--_S_00166	Vereinigte Staaten von Amerika
☑	74	AD--_E_00095	AD--_S_00150	Australien
☑	75	MV--_E_00166	MV--_S_00167	Vereinigte Staaten von Amerika
☑	76	AD--_E_00094	AD--_S_00149	Australien
☑	77	MV--_E_00167	MV--_S_00168	Vereinigte Staaten von Amerika

Figure 3: A metadata query across different corpora in the DGD. The query selects recordings with speakers born in Australia, Canada or the US. Results from the AD and the MV corpus are returned.

4.2.3. Infrastructure

We have argued in the previous section that a common technical approach to speech island corpora is a necessary prerequisite for comparative research **within** the Archive of Spoken German. Speech island data worthy of comparative approaches, however, can also be distributed across different sites. Thus, as mentioned in section 2.1.3, the data of German in Romania are made accessible through an online interface and archived at a data center at the LMU in Munich. Likewise, the Texas German Dialect Corpus has its own website at the

University of Texas in Austin which is also home to http://speechislands.org/ - a portal in which many other collections of German speech island data will be made available. Extending the view beyond German speech islands will bring even more sites into play - such as, for instance, the LANCHART Centre in Copenhagen which hosts corpora on Danish in the Americas (http://danishvoices.ku.dk/).

We think it is neither possible nor desirable to completely “centralize” these resources. Organizing access to speech island data in a network of archives, data centers and platforms instead of a single institution has many advantages. First, it stimulates discussion and competition about the best approaches. Second, it allows a division of labor where no single organization will have the know-how or capacities to deal with all the data. Third, some resources (take Texas German as an example) are best hosted in close proximity to the speech community they represent, because this ensures that the speaker community identifies with the scientific investigation of “their” language variety. Finally, the continuity of speech island research also depends on legal issues concerning data protection. These in turn depend on political systems in respective countries, or other political unions, e.g. differences in EU and US legislation.

The questions discussed in the previous sections in the context of the AGD thus become questions also about digital research infrastructures. Especially on the technological dimension, standards and best practices must be negotiated and coordinated between different institutions and on an international level. Ideally, i.e. if a common approach can be agreed upon, speech island corpora hosted in different places will then be accessible to researchers in similar ways, and solutions developed at one site can be reused by others. Some elements in CLARIN, the Common Language Resource Infrastructure (see https://www.clarin.eu/), convey a first idea of how such an infrastructure can work in practice: on the basis of a common approach to metadata and a metadata domain shared by many CLARIN centers in Europe, CLARIN provides the “Virtual Language Observatory” (VLO, https://vlo.clarin.eu/) as an instrument for discovering language corpora and tools independently of the place at which they are actually hosted. Likewise, “Federated Content Search” (FCS, https://spraakbanken.gu.se/ws/fcs/2.0/aggregator/#) is an approach to making the content of corpora at different sites searchable through a single interface. If the underlying specifications are correctly applied to speech island data, a researcher could thus discover through the VLO that data on German in the USA is available at UT Austin as well as at the AGD, and she could use a single FCS on the respective corpora, for example in order to find instances of the *prepositional calques* (‘mitaus’, ‘mitohne’) mentioned in Boas (2002b). The integration of oral corpora (including speech island and other types of data) into such a digital infrastructure is one of the challenges that the AGD is currently working on.

5 Conclusions and future directions

We have discussed in this contribution the current state of speech island corpora in the Archive for Spoken German and highlighted a number of aspects which we perceive as important challenges in our work on acquiring, archiving and processing such resources and disseminating them to interested researchers.

The state of affairs clearly reflects the long and complex history, not only of the archive itself, but also of the individual data sets and the sometimes adventurous routes via which they have found their way into the archive. While the speech island portfolio of the AGD is certainly large and varied, it is also incomplete and imperfect in many ways - some German speech islands are not represented at all, for others, metadata are too fragmentary or recordings and transcriptions (where they exist) of insufficient quality. For others still, an uncleared legal situation is an obstacle to reuse. Finally, some corpora simply await further curation work (i.e. digitization, documentation, etc.). We will continue to invest a part of the archive's resources into such work, thereby hopefully increasing the amount and quality of speech island data in the archive.

We have identified two approaches which may help to reduce the observed shortcomings in the work with speech island corpora in the future. The first approach (as discussed in section 3 and illustrated in Figure 4) is a closer collaboration between the creators of speech island corpora and the institutions taking responsibility for their long-term availability. If the phases of corpus compilation and corpus curation are seen as more tightly coupled - meaning that producers and archives work together and exchange knowledge from the outset of a project - questions of comparability in content and technology can be addressed early on and extra work to adapt corpora to an archive's requirements reduced or altogether avoided. Likewise, if we do not wait too long before publishing a given corpus, user feedback can further help to identify errors made during compilation or curation and thus inform further curation cycles without causing too much extra work.

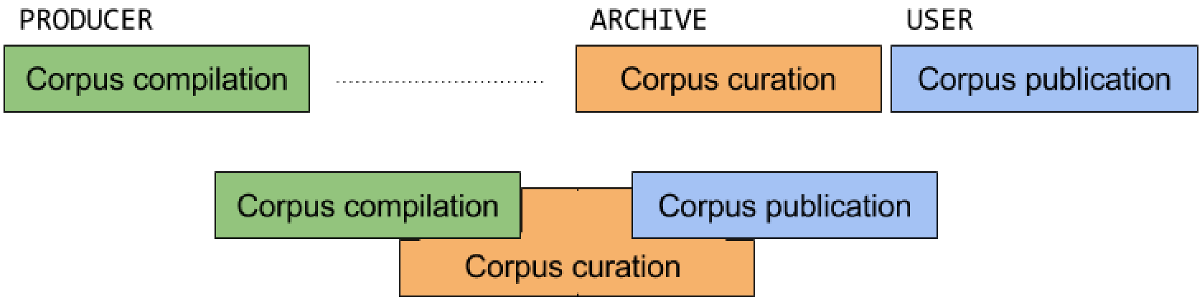


Figure 4: Phases of corpus compilation, curation and publication separated (top) or tightly coupled (bottom).

The second approach (as discussed in section 4) consists in further refining and developing common standards and best practices for speech island corpora on the content and technological level. Each successful effort in identifying and documenting an element of common ground between different speech island corpora, such as a shared elicitation methodology or a common format for digital files, can help to make curation processes more efficient and methods for working with speech island data easier to use. The infrastructure aspect is then an extrapolation of this approach from a single archive to a network of interconnected data centers.

These connections can even fuel interdisciplinary research, as the commonly recorded speech-biographic interviews can also be analyzed with methodologies that have been employed less on speech island data but could enhance speech-island research, i.e. conversation analysis, as suggested by de Ruiter and Albert (2017). In general, the steps of data curation and dissemination described above not only save the remains of unique speech islands, but also make comparative linguistic research possible on solid grounds.

6 References

Bärnert-Fürst, Ute. "Conservation and Displacement Processes of the German Language in the Speech Community of Panambi, Rio Grande do Sul, Brazil." Berend, Nina and Klaus J. Mattheier. *Sprachinselforschung. Eine Gedenkschrift für Hugo Jedig*. Frankfurt am Main: Peter Lang, 1994. 273-287.

Bellmann, Günter and Joachim Göschel. "Tonbandaufnahme ostdeutscher Mundarten 1962-1965: Gesamtkatalog." *Deutsche Dialektgeographie (DDG), vol. 73* 1970.

Berend, Nina and Hugo Jedig. *Deutsche Mundarten in der Sowjetunion. Geschichte der Forschung und Bibliografie*. Marburg: N.G. Elwert Verlag, 1991.

Berend, Nina. "Die Aufnahme deutscher Siedler und die Bildung von Sprachinseln in Russland seit Katharina II." *Die deutsche Sprache in Russland: Geschichte, Gegenwart, Zukunftsperspektiven*. Ed. Ulrich Ammon and Dirk Kemper. München: iudicium, 2011. 60-72.

—. "Sprachliche Anpassung. Eine soziolinguistisch- dialektologische Untersuchung zum Rußlanddeutschen." *Studien zur deutschen Sprache. Forschungen des Instituts für Deutsche Sprache* 1998.

Betten, Anne and Miryam Du-nour. "Sprachbewahrung nach der Emigration. Das Deutsch der zwanziger Jahre in Israel. Teil II: Analysen und Dokumente. Unter Mitarbeit von Monika Dannerer." *PHONAI 45*. Tübingen: Max Niemeyer Verlag, 2000.

Betten, Anne. "Sprachbewahrung nach der Emigration. Das Deutsch der zwanziger Jahre in Israel. Teil I: Transkripte und Tondokumente. Unter Mitarbeit von Sigrid Graßl." *PHONAI 42*. Tübingen: Max Niemeyer Verlag, 1995.

Betten, Anne. "Sprachbiographien deutscher Emigranten. Die „Jeckes“ in Israel zwischen Verlust und Rekonstruktion ihrer kulturellen Identität." Deppermann, Arnulf. *Das Deutsch der Migranten*. Berlin: de Gruyter, 2013. 145-191.

Boas, Hans C., et al. "The Texas German Dialect Archive: A Multimedia Resource for Research, Teaching, and Outreach." *Journal of Germanic Linguistics, 22(3)* 2010: 277-296.

Boas, Hans. "The Texas German Dialect Archive as a Tool for Analyzing Sound Change." Austin, P., H. A. Dry and P. Wittenburg. *Proceedings of the International Workshop on Resources and Tools in Field Linguistics held in conjunction with the Third International Conference on Language Resources and Evaluation*. Las Palmas, 2002a. 28.1-28.4.

Boas, Hans. "Tracing Dialect Death: The Texas German Dialect Project." *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. Linguistic Society of America, 2002b. 387-398.

Bond, Anatole. "Deutsche Siedlung am Schwarzen Meer Lustdorf bei Odessa. Geschichte und sprachliche Studien." *Deutsche Dialektgeographie (DDG)*. Vol. 104. Marburg: N.G. Elwert Verlag, 1978.

Buchheit, Robert H. "Language Maintenance and Shift among Mennonites in South Central Kansas." *Yearbook of German-American Studies*. Vol. 17. 1982. 111-121.

Clyne, Michael. *Deutsch als Muttersprache in Australien*. Wiesbaden: Franz Steiner Verlag, 1981.

Costello, John R. "Syntactic Change and Second Language Acquisition: The Case for Pennsylvania German." *Linguistics, 16(213)* 1978: 29-50.

de Ruiter, J. P. and Saul Albert. "An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology." *Research on Language and Social Interaction, 50(1)* 2017: 1-18.

DFG. *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf, 2013. <http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf>.

Droescher, Werner O. "Puhoi. Eine egerländer Mundart in Neuseeland." *PHONAI 15: Monographien 7*. Tübingen: Max Niemeyer Verlag, 1974. 195-235.

Eichhoff, Jürgen. "A Word Atlas of Pennsylvania German." Seifert, Lester W. J. *Monatshefte für deutschsprachige Literatur und Kultur*. Ed. Mark L. Louden, Howard Martin and Joseph C. Salmons. Madison, WI: Max Kade Institute, 2001. viii + 121 pages, 173 maps: University of Wisconsin Press, 2003.

Eichhoff, Jürgen. "Der deutsche Einfluß auf das Amerikanische Englisch." Stanforth, Anthony W. *Deutsche Einflüsse auf den englischen Wortschatz in Geschichte und Gegenwart*. Tübingen: Max Niemeyer Verlag, 1996. 173-189.

Eichhoff, Jürgen. "The German Language in America." Trommler, Frank and Joseph McVeigh. *America and the Germans: An Assessment of a Three-Hundred-Year History*. Philadelphia: University of Pennsylvania Press, 1985. 223-240.

Eichinger, Ludwig M. "Deutsch in weiter Ferne: die Verbreitung der deutschen Sprache außerhalb des zusammenhängenden deutschen Sprachgebiets: deutsche Minderheiten." Stickel, Gerhard. *Varietäten des Deutschen: Regional- und Umgangssprachen. Jahrbuch / Institut für Deutsche Sprache 1996*. Berlin, New York: de Gruyter, 1997. 155-181.

Eichinger, Ludwig M. "Island Hopping : Vom Nutzen und Vergnügen des Vergleichens von Sprachinseln." Ziegler, Evelyn and Jannis K. Androutsopoulos. *"Standardfragen": Soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation; [Festgabe zum 60. Geburtstag von Klaus Jochem Mattheier]*. Peter Lang Verlag, 2003. 83-107.

Eller-Wildfeuer, Nicole. *Sprecherbiographien und Mehrsprachigkeit. Deutschbasierte Minderheitensprachen in Osteuropa und Übersee*. Stauffenburg Linguistik, vol. 96, 2017.

Fausel, Erich. *Die Deutschbrasilianische Sprachmischung: Probleme, Vorgang Und Wortbestand. Mit Einem Geleitwort von Hugo Moser*. Berlin: Erich Schmidt, 1959.

Fiess, Dietrich. "Siedlungsmundart -- Heimatmundart: Studien zur Entwicklung der Mundart von Sarata in Bessarabien aus ihren verschiedenen Herkunftsmundarten." *Idiomatica, vol. 4*. Tübingen: Max Niemeyer Verlag, 1975.

Fuller, Janet M. "Morpheme types in a matrix language turnover: The introduction of system morphemes from English into Pennsylvania German." *International Journal of Bilingualism, 4(1)* 2000: 45-58.

Geršić, Slavko. "Hodschag / Batschka." *PHONAI 15: Monographien 7*. Tübingen: Max Niemeyer Verlag, 1974. 7-194.

Grubačić, Emilija. "Knićanin / Banat." *PHONAI 9: Monographien 3*. Tübingen: Max Niemeyer Verlag, 1971. 7-94.

Grubačić, Emilija. "Kriva Bara / Banat." *PHONAI 6: Monographien 1*. Tübingen: Max Niemeyer Verlag, 1970. 129-187.

Jedig, Hugo H. *Lepel, Laumptje, Lostichkeit. Gesammelte Beiträge zu deutschen Mundarten in der Sowjetunion, edited by Nina Berend*. Mannheim: Institut für Deutsche Sprache, 2014.

Karch, Dieter and Wolfgang W. Moelleken. "Siedlungspfälzisch im Kreis Waterloo, Ontario, Kanada." *PHONAI 18: Monographien 9*. Tübingen: Max Niemeyer Verlag, 1977.

Kaufmann, Göz. "The Verb Cluster in Mennonite Low German: A new approach to an old topic." *Linguistische Berichte 210* 2007: 147-207.

Keiser, Steve H. "Sound Change across Speech Islands: The Diphthong /aI/ in Two Widwestern Pennsylvania German Communities." *Ohio State University Working Papers in Linguistics, 54* 2000: 143-170.

Krefeld, Thomas, Stephan Lücke and Emma Mages (eds.). *Korpus im Text vol. 2. Zwischen traditioneller Dialektologie und digitaler Geolinguistik: Der Audioatlas siebenbürgisch-sächsischer Dialekte (ASD)*. Münster: Verlagshaus Monsenstein und Vannerdat OHG, 2015.

Ladilova, Anna. *Kollektive Identitätskonstruktion in der Migration: Eine Fallstudie zur Sprachkontaktsituation der Wolgadeutschen in Argentinien*. Frankfurt: Peter Lang, 2013.

Laur, Wolfgang. "Riga." Spracharchiv, Deutsches. *Lautbibliothek der Deutschen Mundarten*. Vol. 9. Göttingen: Vandenhoeck & Ruprecht, 1958.

Lipold, Günter. "Gottschee in Jugoslawien – System, Stil, Prozeß Phonologie einer Sprachinselmundart 1. Teil: Suchen, Hinterland, Zentralgebiet." *PHONAI 26: Monographien 16*. Tübingen: Max Niemeyer Verlag, 1984.

Maitz, Péter. "Unserdeutsch. Eine vergessene koloniale Varietät des Deutschen im melanesischen Pazifik." Lenz, Alexandra N. *German Abroad – Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung*. Göttingen: V & R unipress (Wiener Arbeiten zur Linguistik; 4), 2016. 211-240.

Maitz, Péter, Werner König and Craig A. Volker. "Unserdeutsch (Rabaul Creole German): Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea." *Zeitschrift für germanistische Linguistik 44(1)* 2016: 93-96.

McGraw, Peter A. "Dane County Kölsch, Wisconsin USA." *PHONAI 21: Monographien 12*. Tübingen: Max Niemeyer Verlag, 1979.

Moelleken, Wolfgang W. "Diaphonic Correspondences in the Low German of Mennonites from the Fraser Valley, British Columbia." *Zeitschrift für Mundartforschung, vol. 34(3/4)* 1967: 240–253.

—. "Die rußlanddeutschen Mennoniten in Kanada und Mexiko: sprachliche Entwicklung und diglossische Situation." *Zeitschrift für Dialektologie und Linguistik* 1987: 145-183.

Moelleken, Wolfgang W. "Niederdeutsch der Molotschna- und Chortitzamennoniten in British Columbia/Kanada." *PHONAI 10: Monographien 4*. Tübingen: Max Niemeyer Verlag, 1972.

Mühlhäusler, Peter. "Bemerkungen zum „Pidgin Deutsch“ von Neuguinea." Molony, Carol, Helmut Zobl and Wilfried Stölting. *German in Contact with other Languages*. Kronberg: Scriptor, 1977. 58–70.

—. "Tracing the roots of pidgin German." *Language and Communication, 4/(1)* 1984: 27–57.

Paul, Peter. *Barossadeutsch: Sprachliche Untersuchungen des ostmitteldeutschen Dialektes des Barossatales (Südaustralien)*. BA thesis: University of Adelaide, 1962.

Paul, Peter. "Barossatal: Südaustralien." *PHONAI 6: Monographien 1*. Tübingen: Max Niemeyer Verlag, 1970. 189-317.

—. *Das Barossa Deutsche*. MA thesis: University of Adelaide, 1965.

Popadić, Hanna. "Deutsche Siedlungsmundarten aus Slawonien / Jugoslawien." *PHONAI 19: Monographien 10*. Tübingen: Max Niemeyer Verlag, 1978.

Protze, Helmut. "Die Zipser Sachsen im sprachgeographischen und sprachhistorischen Vergleich zu den Siebenbürger Sachsen." *Zeitschrift für Siebenbürgische Landeskunde 29.2* 2006: 142-151.

Protze, Helmut. "Siebenbürgen und die Zips: Ein sprachgeographischer und siedlungsgeschichtlicher Vergleich." Schmitt, L. E. *Verhandlungen des zweiten Internationalen Dialektologenkongresses, Marburg/Lahn, 5.-10. Sept. 1965/2 (=Zeitschrift für Mundartforschung, Beiheft; NF 4)*. Wiesbaden, 1968. 673-686.

Rowley, Anthony R. "Fersental (Val Fèrsina bei Trient/Oberitalien) – Untersuchung einer Sprachinselmundart." *PHONAI 31: Monographien 18*. Tübingen: Max Niemeyer Verlag, 1986.

Ruoff, Arno. "Grundlagen und Methoden der Untersuchung gesprochener Sprache: Einführung in die Reihe “Idiomatica” mit einem Katalog der ausgewerteten Tonbandaufnahmen." *Idiomatica, vol. 1*. Tübingen: Max Niemeyer Verlag, 1973.

Salmons, Joseph. "Naturalness and Morphological Change in Texas German." Berend, Nina and Klaus J. Mattheier. *Sprachinselforschung: Eine Gedenkschrift für Hugo Jedig*. Frankfurt am Main: Peter Lang, 1994. 59-72.

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

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

—. "The Database for Spoken German – DGD2." *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC’14)*. Reykjavik: European Language Resources Association (ELRA), 2014a. 1451-1457.

—. "The Research and Teaching Corpus of Spoken German – FOLK." *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC’14)*. Reykjavik: European Language Resources Association (ELRA), 2014b. 383-387.

Schmidt, Thomas, et al. *Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache*. Mannheim: Institut für Deutsche Sprache, 2013.

Stift, Ulf-Michael and Thomas Schmidt. "Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch." Sprache, Institut für Deutsche. *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Ed. Melanie Steinle and Franz Josef Berens. Mannheim, 2014. 360-375.

Volker, Craig A. "Rabaul Creole German Syntax." *Working Papers in Linguistics, 21*. University of Hawaii, 1989. 153–189.

Wagener, Peter and Karl-Heinz Bausch. "Tonaufnahmen des gesprochenen Deutsch: Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven." *PHONAI 40*. Tübingen: Max Niemeyer Verlag, 1997.

Wiebe, Grace E. *The segmental phonemes of Swift Current Mennonite Low German*. MA thesis: University of Alberta, 1983.

Wiese, Heike, et al. "German in Namibia: A vital speech community and its multilingual dynamics." Maitz, Péter and Craig A. Volker. *Language & Linguistics in Melanesia (Sonderheft: Language Contact in the German Colonies: Papua New Guinea and beyond*. 2017. 221–245.

Wildfeuer, Alfred. *Sprachenkontakt, Mehrsprachigkeit und Sprachverlust: Deutschböhmisch-bairische Minderheitensprachen in den USA und in Neuseeland*. Mouton: De Gruyter, 2017.

Zwirner, Eberhard and Wolfgang Bethge. "Erläuterungen zu den Texten." Spracharchiv, Deutsches. *Lautbibliothek der Deutschen Mundarten*. Göttingen: Vandenhoeck & Ruprecht, 1958.