Corpus of Austrian Dialect Recordings from the 20th Century – A Cooperation Project

Christian Huber¹, Benjamin Fischer², Bernhard Graf¹

¹Phonogrammarchiv, Austrian Academy of Sciences, ²MEi:CogSci, University of Vienna
christian.huber@oeaw.ac.at

Abstract: The Corpus of Austrian Dialect Recordings from the 20th Century comprises approx. 2450 dialect recordings from the Phonogrammarchiv’s holdings on magnetic tape from 1951 to 1995 from fieldwork by German philologists Eberhard Kranzmayer, Maria Hornung, Werner Bauer, Herbert Tatzreiter and others covering all provinces of Austria as well as linguistic varieties of German spoken in Northern Italy, Hungary, and former Yugoslavia and Czechoslovakia. In a cooperation project of the Phonogrammarchiv, the Research Department “Variation and Change of German in Austria” (both Austrian Academy of Sciences) and the Austrian Science Fund Special Research Programme “German in Austria” (F60) these recordings are now digitized, annotated and analyzed in order to ensure their lasting preservation and utilizability. By digitizing the audio recordings and by enriching and systemizing the metadata in a searchable electronic documentation it will become possible to use these recordings in a wider scientific context. The searchable corpus will further be developed into an online platform and become accessible also to the general public. In this article we will introduce and discuss the corpus, and address various digitization and metadata-related issues from an archival perspective with special attention on encountered real-world questions and problems.

1 Introduction

Languages and their dialects are dynamic systems, subject to constant change. Older varieties are gradually passing out of usage and linguistic phenomena connected to them have often vanished from the language of everyday life. They can authentically be witnessed only in contemporaneous audio recordings, which are available to varying degrees for the time from the late 19th century on. However, the accessibility and sound quality of such recordings depends on the state of preservation of their carrier media. Traditional analogue sound carriers such as wax cylinders, gramophone discs or magnetic tape are subject to natural decay: the more critical the condition of the carrier, the worse the signal quality. Once the carrier can no longer be played, the recordings on it are lost forever. Therefore it is necessary to digitize perishable sound documents as long as the carriers can still be properly played in order to preserve the recorded contents in the long term and make them available for future generations. Digital audio data are no longer bound to an individual data carrier but can be losslessly copied as often as desired. In this way they can be electronically preserved for a virtually indefinite period of time.

1.1 The cooperation

The present project is a cooperation of the Phonogrammarchiv, the Research Department “Variation and Change of German in Austria” (both institutions of the Austrian Academy of Sciences) and the Austrian Science Fund Special Research Programme “German in Austria” (F60; spokesperson: Alexandra N. Lenz). Its aim is to digitize, annotate and analyze Austrian dialect recordings on magnetic tape from the Phonogrammarchiv’s holdings from five decades (1950s–1990s) in order to ensure their lasting preservation and utilizability. Despite their high degree of scientific interest the systematic evaluation of these recordings as well as
their general accessibility has hitherto been hampered by the fact that they exist only in an analogue format that requires special storage conditions and players which are nowadays available only in specialized institutions. Transferring them to an easily accessible and easy-to-handle digital format and enhancing them with an adequate, carefully designed electronic metadata description and database will ensure the utilizability and searchability of the corpus not only for linguistic purposes but make it possible to exploit these recordings in wider scientific as well as general contexts. The data will be utilized in various subprojects of the Research Department “Variation and Change of German in Austria” and the Special Research Programme “German in Austria”. The searchable corpus will further be developed into an online platform and in this way become accessible to other interested researchers and also to the general public.

The project started in January 2019 and is presently concerned primarily with digitizing the analogue audio tapes and systematizing and enriching the descriptive metadata based on the available archival materials. A first stage (2019) is dedicated to digitizing and archiving a subcorpus of approx. 1750 dialect recordings from 1951–1983 which were also included in UNESCO’s national register “Memory of Austria” in 2018 [1]. The remainder of recordings will be dealt with in a second stage in 2020.

1.2 Road map
In section 2 we introduce the corpus. We address the phases of fieldwork that yielded the recordings and discuss the methods employed in the field as compared to previous approaches and approaches taken by other dialect survey projects. We explain what makes this corpus exceptional and sets it apart from traditional field studies in German dialectology at that time, and point out some of the possibilities it offers for sociolinguistic investigations and research into historical dialect data.

Section 3 is concerned with the procedures taking place in the Phonogrammarchiv in order to prepare the corpus for utilization by the Research Department “Variation and Change of German in Austria” and the Special Research Programme “German in Austria”. We address various digitization and metadata-related issues from an archival perspective and pay special attention to real-world questions and problems as encountered in the case at hand, and discuss the approaches we chose to overcome them.

Section 4 presents a brief outlook on the ethical and legal questions the project will have to cope with in the dissemination of the linguistic data and audio data.

2 The corpus
The corpus comprises about 2450 audio recordings (~530 hours) on magnetic tape (alongside associated archival materials) made by dialectologists Eberhard Kranzmayer, Maria Hornung, Werner Bauer, Herbert Tatzreiter and others from the 1950s up to the 1980s (and additionally a few recordings from the 1990s) in order to document (mostly dialectal) linguistic varieties of German in Austria and, in addition, Northern Italy, Hungary, and former Yugoslavia and Czechoslovakia.

2.1 The fieldwork phases
Initiated by Kranzmayer and Hornung (and later mainly directed by Hornung), the recording activities commenced in 1951 (see Figure 1) as a cooperation of the Phonogrammarchiv and the so-called “Wörterbuchkanzlei” (i.e. Kommission zur Schaffung des Österreichisch-Bayerischen Wörterbuchs, from 1969 on Kommission für Mundartkunde und Namenforschung) after the Phonogrammarchiv had received a Philips Maestro tape machine as a gift [2]. Throughout the 1950s and 1960s, large-scale fieldwork on dialectal varieties of German was conducted in numerous places in all provinces of Austria and relevant adjacent areas.
In the early years of the fieldwork, speakers from different villages were often not recorded in the villages they actually lived in but were assembled in a nearby place offering proper recording conditions. There they were the recorded one by one – to the end that the village where the recording session took place often is not the village where the sampled dialect is spoken. In accordance with the actual demographic situation, beside varieties of German also samples of other Austrian regional native languages were recorded (Croatian, Hungarian, Romani, Slovenian, and a few samples of Hebrew; not being dialects of German, they are not included in the present corpus). Some 10 to 20 years later, selected individuals were recorded again for comparative purposes, and a systematic post-survey was carried out in the province of Burgenland 20 years after the initial fieldwork had taken place.

Bauer’s and Tatzreiter’s recordings are mainly from the 1970s and 1980s. Likewise covering all Austrian provinces (with the exception of Vorarlberg) they offer comparative materials relative to the German recordings from the 1950s and 1960s at a temporal distance of 20 years. (Later, the dialectological documentation focus shifted to language islands in Romania and the Americas. These recordings have not been included in the corpus. Recordings of dialect poetry readings, folk songs or folk plays have not been included because of their non-spontaneous nature, and to avoid potential legal issues in prospective later dissemination.)

In total, samples of dialectal varieties from c. 1050 places were collected and c. 1650 speakers were recorded. In numerous cases, thus, two or more speakers were recorded for a particular local variety. Exceptionally rich documentation is available of the province of Burgenland (c. 720 recordings) whereas only a handful of recordings were made of varieties from Vienna. To counterbalance this disproportion, Wilfried Schabus’ recordings of older Viennese citizens from the 1990s have also been added to the corpus. The recordings represent the first attempt at a comprehensive nation-wide audio documentation of linguistic varieties in Austria. They provide an exceptional picture of Austria’s dialect landscape after the middle of the 20th century across all regions and constitute a source of historical dialect audio data that is unique in the German-speaking area and is a valuable primary source for various areas of research.

2.2 Field methods
Quite in the spirit of their time, Kranzmayer and Hornung aimed at capturing “die älteste erreichbare Schichte mundartlicher Ausdrucksweise [the oldest accessible stratum of vernacular expression]” by targeting old speakers who were also “auf ihre Bodenständigkeit geprüft [checked for their being rooted in their environment]”. The emerging collection of sound recordings was considered “ein Denkmal altbäuerlich-österreichischer Sprach- und Volkskultur [a monument of old rural Austrian language and folk culture]” (all quotes from Hornung [3]).
In order to go beyond the limitations of mere lexical data elicitation and to avoid the priming effects of *Wenker sentences* [4], they relied instead on recordings of spontaneous language and stimulated their informants to speak freely about a topic suggested by the inter-viewer, or recorded them in (likewise stimulated) conversation during a session. In this way, they sought to document dialectal varieties in their actual use as well as syntactic or prosodic phenomena that would not readily emerge from answering to a questionnaire. Therefore, the recordings often deal with aspects of the informants’ daily lives (e.g. agriculture, crafts), local customs and folk culture, festivals or fairs, traditional narratives or the informants’ memories of earlier times. Occasionally, lexical items are elicited in a conversational style.

Interestingly, audio recordings of free speech supplementing the traditional field records (questionnaires, etc.) are introduced around the same time also in North American dialect surveys (e.g. *Linguistic Atlas of the Upper Midwest*, *Survey of the German Spoken in Wisconsin*, see [5], [6]). Much as with Hornung and Kranzmayer, there as well “the informant is questioned only enough to get him launched on some topic that strikes his fancy” (Wilson [5]). It is currently not known if there has been a flow of ideas of some sort between Kranzmayer and the American fellow-fieldworkers or if this is a mere coincidence, facilitated by the fact that after the end of World War II inexpensive and light-weight electrical recording machines became available on the market that better allowed for such a type of recordings. Hornung attributes the idea of audio-taping spontaneous language without prior rehearsals to Kranzmayer and makes no reference to other dialect survey teams [3]. In contrast to Hornung and Kranzmayer’s approach, recording an audio sample of free speech has always remained an appendix to questionnaire-based fieldwork (which was later also accompanied by audio taping) with the American survey teams, and they never turned to it to replace traditional field records.

The informants’ social data appear to have been recorded prior to the audio recording sessions and are preserved in written form in the archival documentation (see Figure 2). It is currently unknown if the procedure of collecting the social data has generally never been taped or if this section of a session has later been edited out of the audio recordings. Only in a few rare cases it is preserved in the beginning of a recording.

![Figure 2. Archive protocol of recording B 33 (excerpt) (©Phonogrammarchiv)](image-url)
It is unclear as of yet to what extent Kranzmayer, Hornung and the other fieldworkers evaluated and analyzed the language data they collected. Hornung herself published only two short articles that are directly concerned with the recording enterprise as such or with particular recordings [3], [7], and Wilfried Schabus published the transcription of a recording from 1951, with extensive discussion [8]. Otherwise, the corpus and its recordings are only referred to more or less in passing in a handful of likewise short articles, e.g. [9], [10], [11]. While all social data and other metadata pertaining to the recordings are contained in the Phonogrammarchiv’s archival documentation, it contains only few (full or partial) transcripts. It has transpired that in the mid-1970s, Kranzmayer commissioned the transcription of a number of recordings. However, it is not known where these transcripts, if they still exist, are deposited. Finally, there is hope that transcripts or other documentary materials may be found among the papers in Hornung’s estate, currently stored at the department “Variation and Change of German in Austria”. At a later point, attempts will be made at the retrieval of such materials.

2.3 Options of analysis and data utilization

Hornung and Kranzmayer deviated from the earlier dialectological fieldwork or recording practice, which focussed on questionnaire elicitation such as the 40 Wenker sentences, and in which even recordings of purported spontaneous speech had in fact often been rehearsed many times before the actual recording session took place. Their recordings may thus be less suitable for filling gaps in a dialect atlas but in return exhibit structures and phenomena that are not readily obtainable in an interview and cannot be gleaned from questionnaire data. In this way they also preserve dialectal features that cannot be captured in elicitations and have long since vanished. Typically featuring long passages of spontaneous continuous speech, the recordings lend themselves e.g. to studies pertaining to prosody, information structure and discourse prominence-related issues, pronominal reference, narrative style and other phenomena that can fully be studied only in larger non-preconstructed contexts. Their oftentimes astonishingly good sound quality makes them suitable also for phonological and phonetic investigations.

Spanning almost five decades, the corpus likewise allows us to trace linguistic change in the recorded varieties. The large variety of covered places and recorded speakers together with the considerable chronological depth allows for multi-dimensional sociolinguistic and fine-grained geolinguistic analyses. The fact that 10 to 20 years after the initial fieldwork in the 1950s and 1960s most areas were covered again in the 1970s and 1980s, sometimes even with the same speakers, makes it possible to conduct inter- as well as intra-speaker real time analyses.

In addition, since the speakers often talked about their immediate living environment and aspects of their culture, or offered specimens of their narrative traditions, the corpus is also of great socio-historical and historico-cultural importance.

3 Preparing the corpus

In this section we address digitization and metadata-related issues and procedures in preparing the corpus for further utilization, and discuss encountered problems.

3.1 Digitization

Out of c. 2450 individual recordings (c. 530 hours) in the corpus, only 1219 recordings (c. 130 hours) had already been digitized at the time the project started. With the exception of a few DAT tapes, the remaining recordings consist of c. 400 reels of magnetic audio tape. Their digitization to 24 bit/96 kHz WAV files is in progress since the beginning of 2019. Following digitization, the digital copies of these tapes are segmented so that each recording is available as an accordingly labeled separate file. Most recordings are mono recordings. We also discovered that among the previously digitized materials, a considerable number of digital copies of archive tapes had not been segmented, or only incompletely so, and other archive tapes had only partially...
been digitized. We therefore had to include the completion of these tasks in our workflow. While the recordings in the corpus from the 1990s on digital audio tape (DAT) cassettes need not be digitized as they are recorded in a digital format, some are yet to be transferred and segmented.

3.2 1st generation vs. 2nd generation tapes

The digitization of audio tapes concerns first-generation (i.e. original) or second-generation tapes (i.e. analogue copies), which has to do with the history of analogue audio archiving in the Phonogrammarchiv. In the years 1951–1958, when the Phonogrammarchiv possessed only a single tape machine [2], the original tapes recorded in the field were included in the Phonogrammarchiv’s holdings after they were cut and interspersed with sections of leader tape to separate individual recordings from one another. The digital copies of these tapes are therefore sourced from the first tape generation. It cannot be excluded that in the course of archiving, parts of recordings that were deemed unfit or not worthy of being archived were cut away and are now lost.

Later, when also battery-powered, portable tape recorders could be acquired and were frequently lent to field researchers, there was a policy of copying the resulting recordings onto so-called archive tapes, and only the latter were fully included in the Phonogrammarchiv’s holdings (in the mid-1970s, leader tape between recordings was also replaced by reference tones). These archive copies were given priority over the original tapes, and all archival recording-specific documentation refers to them. Digital copies of these archive tapes are therefore sourced from the second tape generation only. Unfortunately, the Phonogrammarchiv never kept records of the original tapes so that it is often unclear nowadays what happened to them. Some are known to have been reused or returned to the fieldworkers whereas others remained in the Phonogrammarchiv but were not catalogued.

However, since analogue tape copies suffer a loss in audio fidelity, it is of course preferable to digitize the original tapes instead of the 2nd generation archive tapes if the original tapes can be retrieved and properly played. Of some dialect recordings, the original tapes were discovered in the Phonogrammarchiv. However, their actual contents are not clear because the tape boxes are only sparsely labeled and no documentation is available as to what recordings are found on what tape, and their condition and playability is yet to be determined. Also, it is not known if the recordings were always archived according to their order on the original tapes, and again it cannot be excluded that certain parts were edited out when copying the originals onto the archive tapes. From this it becomes clear that in order to utilize the original tapes, the archive tapes must be digitized to make it possible to identify matching recordings on the original tapes. We have decided to fall back on original tapes to the extent feasible within the project’s resources.

After Hornung’s passing in 2010, some 40 tapes from her estate that had been found in her flat were handed over to the Phonogrammarchiv. Again, no documentation is available as to their precise contents and labeling is sparse, but cursory checks have shown that some tapes contain the original versions of recordings found on 2nd generation archive tapes, and several of them have already been digitized. In some cases, the tape boxes were mislabeled and the tapes contained new dialect recordings that had not previously been known to exist. We plan to digitize also the remainder of these tapes.

A similar state of affairs must be expected also regarding other parts of the Phonogrammarchiv’s tape holdings, therefore we plan to explore the possibilities of audio fingerprinting for identifying original versions of archived recordings sourced from second-generation tapes on digital copies of original tapes for which no documentation is available.

3.3 Granularization, enrichment and re-assignment of metadata

In the original historical archival documentation (data sheets on paper, so-called protocols) there is a separate data sheet (for a long time handwritten, later typewritten) for each recording that contains the metadata pertaining to that recording (for an example from 1951 see
Figure 1 above). Metadata include e.g. the archive signature (the recording’s catalogue number), the date when and place where the recording was made, its duration, recorderes’ names and social data, involved fieldworker(s), recorded languages/varieties or musical forms, topics and other content-related indications, a time protocol, and technical metadata (e.g. technical equipment involved, original and archive format, transfer procedures), and so forth.

When the Phonogrammarchiv introduced the electronic documentation of recordings by means of a database around 1990 there were already tens of thousands of recordings with an analogue archival documentation. In order to have all recordings represented in the database quickly, in most if not all such cases only some basic metadata had been entered. An important task in our project is therefore to enrich the electronic metadata pertaining to our corpus on the basis of the available analogue documentary materials, and on this occasion also to correct possible errors.

However, when switching to electronic documentation also the fatal decision had been made to have the database not document recordings but bundles of recordings, and to lump together the metadata of the individual recordings made by a fieldworker on the same day into a single general bundled entry composed of the metadata of all recordings in the bundle, thereby dissociating the metadata from the recordings to which they pertain. As schematically illustrated in Figure 3, in bundle entries the metadata are no longer associated with individual recordings but only with the bundle.

As a consequence, a search for recordings according to a particular search criterion does not return individual recordings in the results but only bundles of recordings that contain one or more recordings to which the search criterion applies, and it cannot specify which recordings exactly these are. The search may thus return also quite a number of recordings to which the search criterion does not apply. In Figure 3, for example, a search for Croatian will not only return Recording 1 but also Recordings 2 and 3 because they are included in the bundle, despite the fact that they do not contain Croatian. Similarly, a combined search for Croatian and Traditional Song, e.g. when searching for traditional songs in Croatian, will return Bundle A with its three recordings because the bundle contains Croatian as well as Traditional Song although not a single recording in Bundle A features a traditional song in Croatian since each of the two criteria applies to a different recording in the bundle. Unambiguous and accurate search results are only possible in cases where a bundle contains just a single recording. In the case of bundles that contain more than one recording, the search criterion may apply to minimally one and maximally all recordings (and any number in between) in a given bundle in the search results, and the original protocols on paper must be consulted to determine the precise recordings to which the search criterion applies. It is evident thus that a huge number of recordings cannot be unambiguously found by a search in the database, and conversely that the database often returns search results that do not conform to the search criteria.
In our corpus, roughly 50% of the recordings are included in such metadata bundles. Since sometimes up to 20 speakers (each representing the local dialect of a different village) were recorded in a single day in the field we are faced with a number of very complex bundles. In order to make the electronic documentation usable for any search-related purposes and corpus exploitation tasks it was necessary therefore to granularize all metadata bundles and re-associate all pieces of metadata with those individual recordings to which they actually pertain. Since the problem of metadata bundles is not restricted to our corpus but extends across the database we decided that the procedures to achieve this must be applicable to the database in general.

First, we created an excerpt of the Phonogrammarchiv’s database that contains only the data sets that are relevant to our corpus, later to be re-transfered and to replace the original entries. This was done in view of the risk of inadvertently compromising the complete database by applying experimental procedures to the data in our corpus. In the next step we granularized all bundle entries that were composed of the metadata of several (i.e. more than one) recordings into as many single-recording bundles as there were recordings in the bundle, together with extending the bundle signature by a delimiter followed by internal consecutive numbering (see Figure 4).

Thus, a bundle comprising the metadata of 17 different recordings was granularized into 17 single-recording bundles (empty copies of the original bundle), each of which to contain the metadata of just a single recording. This was done in order not to alter the basic architecture of the database and to enable us later to re-transfer the granularized entries. With the help of a matrix tool, each piece of metadata from the original bundle entry is now assigned to the single-recording bundle to which it pertains (figure 5).

---

**Figure 4. Granularization of multi-recording bundles**

<table>
<thead>
<tr>
<th>multi-recording bundle signature</th>
<th>single-recording bundle signatures</th>
<th>(recording numbers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19520919.N001 (= B 181–197)</td>
<td>19520919.N001#001 (= B 181)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19520919.N001#002 (= B 182)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19520919.N001#003 (= B 183)</td>
<td></td>
</tr>
<tr>
<td>(etc.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Figure 5. Granularization, (re-)assignment of metadata**
Since in the original multi-recording entries all links between the metadata and the respective recordings were lost, this reassignment of metadata has to be done manually, using the original hand- or typewritten documentation. By granularizing a bundle of many recordings into many bundles of single recordings, the existing architecture of the database is left intact but it becomes possible for a search to find all and only the recordings to which a search criterion applies without also returning recordings to which the search criterion does not apply.

In the course of enriching the metadata we will also add the geodata of the villages or towns whose dialectal varieties are represented in the recordings, or where recording sessions took place, with the prospect of creating an interactive map of dialect places in Austria and the adjacent regions. Since more often than not, the place where a recording was made is not the town or village where the sampled variety is spoken, a feasible solution must be found to include the relevant geodata in the dataset, as the database currently allows only for entering the geodata of the place of recording.

4 Outlook

Beside questions concerning technical and IT-related aspects, the project also faces ethical and legal questions: Since the recordings in the corpus represent the Austrian dialects of German at stages of up to 70 years ago that often can no longer be encountered today, it would in principle be desirable also to make the recordings themselves available to a wider audience. However, this is where ethical and legal considerations come into play. It is evident that great care has to be taken in this matter. The legal status of such field recordings is not yet clear and it must first be examined which legal regulations actually apply to them. Moreover, even if there are no legal objections, it must still be examined for each recording individually whether a publication is ethically acceptable or whether there are reasons to better refrain from making it publically available. When dealing with corpora of historical field recordings it must be kept in mind that at the time when these recordings were made, it was not customary to have recorded speakers undersign comprehensive legal documents regulating each and every aspect of utilizing the recordings or their personal data (a comparable state of affairs holds even with present-day field recordings from remote or off-road areas of speakers perhaps not fully matching a Western academic level of literacy). In today’s frequent calls for extensive and unconstrained dissemination of field data these aspects are generally overlooked (or ignored).

While first legal consultations have already taken place, it will be an important future task of the project to do further research in this direction. In the end, these questions are relevant not only to the corpus at hand but also to historical recordings from field research in general.

5 Acknowledgments

This project would not exist without the contributions of both project partners. Alexandra Lenz, head of the Research Department “Variation and Change of German in Austria” (Austrian Academy of Sciences) and spokesperson of the Austrian Science Fund Special Research Programme “German in Austria: Variation – Contact – Perception” (F60) made it possible to employ a person for entering metadata and related work, and to cover the programming costs. The Phonogrammarchiv, in an in-kind contribution, covers the digitization of the analogue audio tapes and associated procedures. We would also like to thank Ludwig Maximilian Breuer, who takes care of this project at “Variation and Change of German in Austria” for his support, and to express our gratitude to Gernot Katzlberger of Statistik Austria for making available to us the geodata of nearly 17300 places in Austria as recorded in the “Österreichisches Ortsverzeichnis [Gazetteer of Austria] 2015”.

39
6 References


