Why it Can be Difficult to Make Historic Language Recordings Accessible: A View from a Corpus of Historic Dialect Recordings

Christian Huber
Phonogrammarchiv
Austrian Academy of Sciences
Vienna, Austria
christian.huber@oeaw.ac.at

ABSTRACT

There is a growing demand to make historic linguistic field recordings accessible not only to the scientific community but also to the language communities as well as the interested public. However, when dealing with a corpus of historic language recordings, a number of challenges must be faced before dissemination issues can even be addressed. The present paper reports the experiences made in preparing a corpus of historic Austrian dialect recordings from the Phonogrammarchiv's holdings and the real-life issues encountered in the process and discusses what needs to be done with such a corpus before something can be done with that corpus.

CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Security and privacy → Human and societal aspects of security and privacy → Privacy protections • Applied computing → Law, social and behavioral sciences → Anthropology → Ethnography

KEYWORDS

Historic dialect recordings, historic language corpora, metadata structuring, granularisation, geodata, archiving, digitisation, ethical issues, legal issues

ACM Reference format:

Christian Huber. 2023. Why it can be difficult to make historic language recordings accessible: A view from a corpus of historic dialect recordings. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114302

1 INTRODUCTION

The Phonogrammarchiv of the Austrian Academy of Sciences has been engaged in making linguistic recordings from its inception [1], its first recording of an Austrian dialect of German dating from 1901 [2]. Over the decades, a collection of several thousand recordings of German dialects of Austria and adjacent areas has been created [3][4]. However, historically grown collections of language recordings pose challenges that are rarely discussed, as they do not arise in modern corpora that are generated within a specific research context and infrastructure. In such collections, the recordings were made not only at different times, but also with different objectives, according to different methods, with different recording technologies, and using different documentation practices [5]. Therefore, before such corpora can be exploited in linguistic or other research, one must deal with questions of data organisation as well as the preservation of their sonic content.

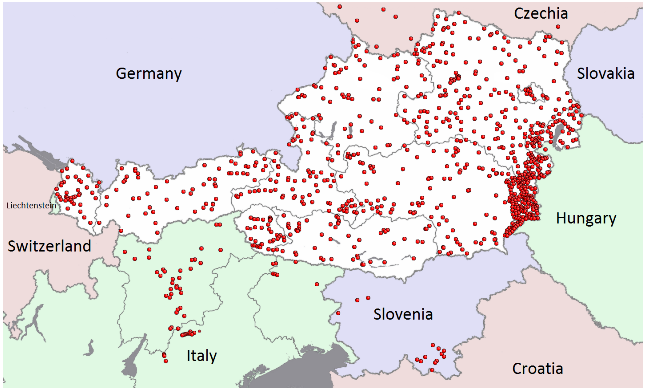


Figure 1. Locations of documented dialect points (audio recordings) (© OpenStreetMap contributors)

In accordance with the then-prospective budget, we selected approximately 2450 recordings of spontaneous language on magnetic tape (and some digital audio tape cassettes) from, roughly, 1000 places and 2500 speakers (see Figures 1 and 2), covering almost five decades (early 1950s to mid-1990s). In a cooperation of the Phonogrammarchiv with the Austrian Science Fund Special Research Programme F60 *German in Austria* and the *Austrian Centre for Digital Humanities* that started in 2016, we digitised these recordings and provided a structured and searchable description building on the Phonogrammarchiv's database and aim to annotate them utilising the corpus-linguistic structures developed in the *German in Austria* programme, and finally to present the results in a common platform.

2 DIGITISING THE TAPES

Traditional analogue sound carriers, e.g., wax cylinders or magnetic tape, are subject to natural decay. Once the carrier can no longer be played, the recordings on it are lost forever. Therefore, perishable sound documents must be digitised as long as the carriers can still be properly played in order to preserve the recorded contents in the long term. 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.

At the start of the project, less than half of the recordings had already been digitised. The remainder was contained on around 400 tape reels that were digitised to 24bit/96kHz .wav files and subsequently segmented, so that each recording is now available as a separate file. We also discovered that among the previously digitised materials, a considerable number of digital copies of tapes had not been segmented, or only incompletely so, and other tapes had been digitised only partially. We therefore had to include the completion of these tasks in our workflow.

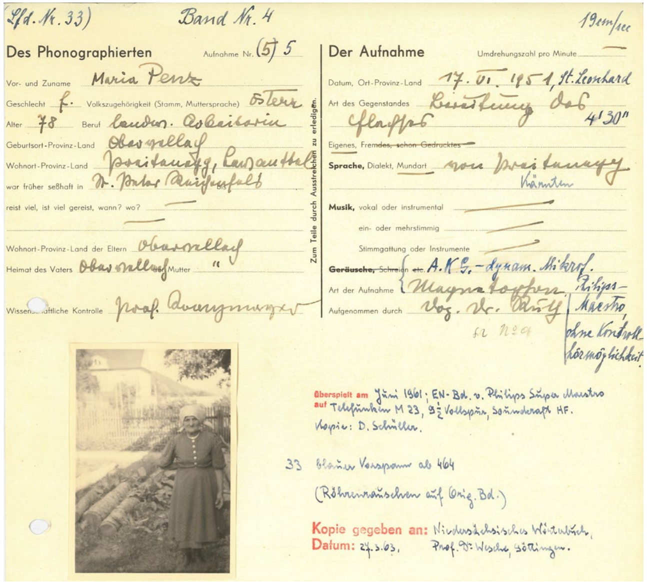
Figure 2. Fieldwork in Carinthia (1951) (©Phonogrammarchiv)

3 METADATA

The original historical archival documentation consists of data sheets on paper for each recording (for a long time handwritten, later typewritten) that were already available in a scanned format (.pdf files; for an example see Figure 3). Metadata include, e.g., the archive signature, the date and place of the recording, its duration, the recordees' names and social data, the involved fieldworker(s), recorded languages/varieties or musical forms, topics and other content-related indications, a time protocol detailing the contents of the recording, and technical metadata (e.g., technical equipment involved, track positions, tape speed).

3.1 Metadata enrichment

For handling the metadata, we utilised the pre-existing, very fine-grained, structures provided by the Phonogrammarchiv's relational database, and the metadata entries already available in it. However, these entries were often incomplete and in need of granularisation. When the Phonogrammarchiv introduced the electronic documentation of recordings in a database around 1990, there were already tens of thousands of recordings with archival documentation on paper. To save time and to have all recordings represented in the database quickly, most often only some basic metadata had been entered. An important task in the project was therefore to enrich the electronic metadata pertaining to our corpus based on the available analogue documentary materials (to be typed out or subjected to optical character recognition), and also to correct possible errors.



(Lfd. Nr. 33) Band Nr. 4 19cm/sec

Des Phonographierten Aufnahme Nr. (5) 5

Vor- und Zuname Maria Penz

Geschlecht f. Volkszugehörigkeit (Stamm, Muttersprache) Österr.

Alter 78 Beruf landw. Arbeiterin

Geburtsort-Provinz-Land Obervellach

Wohnort-Provinz-Land Preitenegg, Lavanttal

war früher seßhaft in St. Peter Reichenfels bis

reist viel, ist viel gereist, wann? wo? —

Wohnort-Provinz-Land der Eltern Obervellach

Heimat des Vaters Obervellach der Mutter "

Wissenschaftliche Kontrolle Prof. Kranzmayer

Zum Teile durch Ausstreichen zu erledigen.

Der Aufnahme Umdrehungszahl pro Minute —

Datum, Ort-Provinz-Land 17. VI. 1951, St. Leonhard

Art des Gegenstandes Bearbeitung des Flachses 4'30"

Eigenes, Frem~~des, schon Gedrucktes~~

Sprache, Dialekt, Mundart von Preitenegg Kärnten

Musik, vokal oder instrumental —

ein- oder mehrstimmig —

Stimmgattung oder Instrumente —

~~**Geräusche,** Schreien etc.~~ A.K.G.-dynam. Mikrof.

Art der Aufnahme Magnetophon Philips-Maestro, ohne Kontrollhörmöglichkeit.

Aufgenommen durch Doz. Dr. Graf

 Nr. 9

Überspielt am Juni 1961; EN-Bd. v. Philips Super Maestro
auf Telefunken M 23, 9½ Vollspur, Sonderkraft HF.
Kopie: D. Schüller.

33 blauer Vorspann ab 464

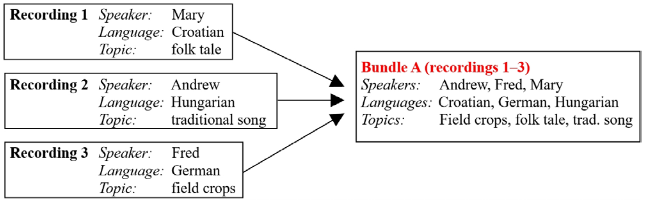
(Röhrenrauschen auf Orig. Bd.)

Kopie gegeben an: Niedersächsisches Wörterbuch,
Datum: 27.3.63, Prof. Dr. Wesche, Göttingen.

Figure 3. Archive protocol of recording B 33 from 1951 (excerpt) (©Phonogrammarchiv)

3.2 Granularisation

However, when switching to electronic documentation in the 1990s, it had also been decided to set up the database in such a way that it does not document individual recordings but only bundles of recordings: the metadata of the individual recordings made by a fieldworker on the same day were collapsed and lumped together into a single general bundled entry composed of the metadata of all recordings in the bundle, thereby dissociating the metadata from the actual recordings to which they pertain, as schematically shown in Figure 4. In such bundle entries the metadata are no longer associated with individual recordings but only with the bundle as such. Thus, from *Bundle A* in Figure 4 it can no longer be told whether *Mary*, or *folk song*, or *Croatian*, or any of the other entries, pertains to *recording 1*, *2*, or *3*.



Recording 1 *Speaker:* Mary *Language:* Croatian *Topic:* folk tale

Recording 2 *Speaker:* Andrew *Language:* Hungarian *Topic:* traditional song

Recording 3 *Speaker:* Fred *Language:* German *Topic:* field crops

Bundle A (recordings 1–3) *Speakers:* Andrew, Fred, Mary *Languages:* Croatian, German, Hungarian *Topics:* Field crops, folk tale, trad. song

Figure 4. Lumping together metadata in a bundle entry (schematically)

Therefore, search results can be severely contaminated, since a particular search criterion does not return individual recordings in the search results but only bundles of recordings that contain one or more recordings to which the search criterion applies. In addition, the search results cannot specify which recordings these are, and the search may also return a number of recordings to which the search criterion does not apply. Similarly, a combined search, e.g., a search involving two search criteria, may return bundles in the search results in which one or more recordings correspond to one of the search criteria at a time but with no recording to which both criteria apply. Since with bundles containing more than one recording, the search criterion may apply to minimally one and maximally all recordings in the bundle in the search results, the original protocols on paper must be consulted to determine the precise recording(s) to which the search criterion applies. Thus, a huge number of recordings cannot be unambiguously found by a search in the database, and the database often returns search results that do not conform to the search criteria.

In our corpus, roughly 50% of the recordings were included in such metadata bundles. Since sometimes up to 20 speakers (each representing the local variety of a different village) were recorded on a single day in the field, we were faced with a number of very complex bundles. To make the electronic documentation usable for any search-related purposes and corpus exploitation tasks, it was therefore necessary to granularise all metadata bundles and re-associate all pieces of metadata with those individual recordings to which they actually pertain. Since the problem is not restricted to our corpus but extends across the Phonogrammarchiv's database, we decided that the procedures to achieve this must be applicable to the database in general. For practical reasons we created an excerpt of the Phonogrammarchiv's database that contains only the data sets relevant to our corpus. Later, these data sets will be re-transferred and will replace the original entries.

In the next step we granularised all bundle entries composed of the metadata of several 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 (schematically shown in Figure 5).

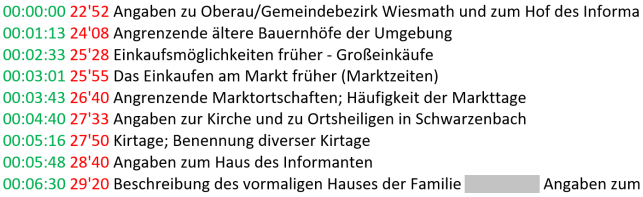
multi-recording bundle signature		*single-recording bundle signatures*	*original recordings (archive numbers)*
19520919.N001	=>	19520919.N001#001	(= B 181)
(= B 181–B 197)		19520919.N001#002	(= B 182)
		19520919.N001#003	(= B 183)
		(etc.)	

Figure 5: Granularisation of multi-recording bundles

With the help of a matrix tool, each piece of metadata from the original bundle entry was then assigned to the single-recording bundle to which it pertains. Since in the original multi-recording entries all links between the metadata and the respective recordings were lost, this reassignment of metadata had to be done manually by falling back on the original hand- or typewritten documentation.

3.3 Timelines in protocols

Since the timelines in the original protocols of recordings (indicating what happens when in a recording) often do not start at the beginning of the respective recording but at the beginning of the tape reel containing it (which usually contains several other recordings), we had to correct the time markers in about 900 protocols and align them with the sound files (as, e.g., in Figure 6), later to be linked to the sound files in the database.



00:00:00 22'52 Angaben zu Oberau/Gemeindebezirk Wiesmath und zum Hof des Informa
00:01:13 24'08 Angrenzende ältere Bauernhöfe der Umgebung
00:02:33 25'28 Einkaufsmöglichkeiten früher - Großeinkäufe
00:03:01 25'55 Das Einkaufen am Markt früher (Marktzeiten)
00:03:43 26'40 Angrenzende Marktortschaften; Häufigkeit der Markttage
00:04:40 27'33 Angaben zur Kirche und zu Ortsheiligen in Schwarzenbach
00:05:16 27'50 Kirtage; Benennung diverser Kirtage
00:05:48 28'40 Angaben zum Haus des Informanten
00:06:30 29'20 Beschreibung des vormaligen Hauses der Familie [redacted] Angaben zum

Figure 6: Adapted and original time markers in a protocol

4 GEODATA

Due to the large number of villages and towns covered in the corpus it was necessary to implement a uniform and unambiguous representation of geographical information using a controlled list of places and converting mentions of toponyms (recording site, a speaker's place of birth or residence, etc.) into references to entries in the list of places. A local authority, *Statistik Austria*, provided us with an up-do-date and official dataset of all towns in Austria, including their official administrative names and geodata as well as the larger administrative units (municipalities, districts, provinces). With the help of this data set, it was possible to set up a representation of place names in such a way that they are not only identified by their official designations and reference numbers (beside geographical coordinates) but also are embedded in the hierarchy of the respective administrative units, where each level is embedded under the next higher level (i.e., PLACE < MUNICIPALITY < DISTRICT < PROVINCE < STATE), with the option of also adding alternative names of a toponym (e.g., potential historical names, or its name in other languages), or other information.

5 DISSEMINATION: LEGAL AND ETHICAL QUESTIONS

While it is a noble goal to make historic dialect recordings accessible to all interested parties (researchers, communities, or also the interested taxpayer who often financed the fieldwork and archiving), legal regulations have still to be obeyed, and ethical questions must be considered.

The recordings in the corpus were generally made under the stipulation that they would be used only for research purposes but would not be made publicly accessible. Thus, the recordings at times also feature sensitive or rather personal content (identified as such by the fieldworkers, the informants themselves, or also archivists), and great care must be taken when considering what should be made accessible to whom, even if several decades have passed since the recordings were made.

On the legal side, it must be kept in mind that the recordings were made at a time before it became common practice to record an agreement with the speaker as to how a recording could be used. A crucial question is whether what a speaker utters on a recording surpasses the threshold of originality and is protected by copyright law. In most cases this question cannot be decided outside a court of law, and permission to publish a recording had to be obtained from the speakers or their legal successors. However, in most cases, the personal data given in the protocols is not sufficient to track down speakers or their heirs (e.g., no date of birth is mentioned but only the year of birth, or the age at the time of the recording). If speakers or their heirs are not known or cannot be located, the respective recordings might be registered as orphan works.

In some types of research, e.g., sociolinguistics, certain personal data may be relevant. Since according to the General Data Protection Regulation the protection of an individual’s personal data expires with the individual’s passing (see, e.g., [8]), the personal data of speakers who are known, or can safely be assumed, to be deceased (e.g., if a speaker had reached a higher age than the oldest living individual in Austria, or on Earth), could in principle be shared. However, ethical considerations may come into play here as well. While it is standard practice to anonymise (or rather, pseudonymise [7]) personal data in written accounts, sound recordings pose the problem of the human voice. Whether or not a speaker’s voice counts as personal data is still a matter of debate (see, e.g., [8] vs. [9]).

For such reasons, making the recordings openly accessible is not a trivial matter, and affordable solutions generally applicable not only to isolated recordings but to larger portions of the corpus, or to the entire corpus, are not yet in sight.

6 CONCLUSION

The preparation of a corpus of historic language recordings can be laden with more complications than first meets the eye. The dire funding situation in Austria for such projects often requires dividing the work between several cooperation partners contributing their respective expertise, and requires the partners’ goodwill, and much in-kind work. A high degree of flexibility is asked for, since a change of priorities on the part of a cooperation partner (or even one’s own department) may soon have the consequence that the project’s objectives cannot be achieved according to the original planning, so that alternative ways must be found. Thus, we hope to be finally able to tackle the issue of merging the Phonogrammarchiv’s metadata with *German in Austria*’s corpus-linguistic structure and to start annotating selected recordings, which has been delayed for several years. So far, about 70 transcripts in various formats (from the 1970s) are available. For increasing the number of transcripts, we have meanwhile decided to also include contributions from parties external to the cooperation who work on recordings from the corpus in other contexts. It is clear, however, that it will still take time until a substantial number of fully described and annotated recordings become available.

ACKNOWLEDGMENTS

Thanks go to both cooperation partners, to Gernot Katzlberger of *Statistik Austria* for making available the geodata of nearly 17300 places in Austria, and to Regula Sutter for her help in selecting the appropriate CCS concepts.

REFERENCES

[1] Leo Hajek. 1928. Das Phonogrammarchiv der Akademie der Wissenschaften in Wien von seiner Gründung bis zur Neueinrichtung im Jahre 1927 (58. Mitteilung der Phonogrammarchivs-Kommission). *Sitzungsberichte der Akademie der Wissenschaften in Wien, philosophisch-historische* Klasse 207(3), Hölder-Pichler-Tempsky, Vienna, 1–22. Online: https://www.oeaw.ac.at/fileadmin/Institute/PHA/PDF/Hajek_1928.PDF (accessed June 8, 2023)

[2] Ph 105: Attergauer Dialekt https://catalog.phonogrammarchiv.at/sessions.php?id_sessions=3794&action=view&sortieren=signatur&vonBis=0-9 (accessed May 14, 2023)

[3] Maria Hornung. 1961. Tonaufnahmen im Dienste der Mundartforschung. Zum 60jährigen Bestehen des Phonogrammarchivs der österreichischen Akademie der Wissenschaften in Wien. *Zeitschrift für Mundartforschung* 28(2), 183–191.

[4] Wilfried Schabus. 1999. Die Bestände des Phonogrammarchivs an Sprachaufnahmen. *Das audiovisuelle Archiv* 45, 23–32. Online: https://www.oeaw.ac.at/fileadmin/Institute/PHA/PDF/schabus_1999.pdf (accessed June 8, 2023)

[5] Christian Huber and Benjamin Fischer. 2021. Digitising a corpus of Austrian dialect recordings from the 20th century. In *Digital Lexis and Beyond*, ed. by Ch. Katsikadeli, M. Sellner & M. Gassner, Verlag der ÖAW, Vienna, 38–65. DOI: https://doi.org/10.1553/OE_Phonogrammarchiv

[6] Wirtschaftskammer Österreich. 2023. EU-Datenschutz-Grundverordnung (DSGVO): Wichtige Begriffsbestimmungen. https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Wichtige-Begriffsbestimmu.html (accessed May 14, 2023).

[7] Caroline Schwabe. 2021. Was sind pseudonymisierte Daten? Pseudonymisierte Daten nach DSGVO. https://www.robin-data.io/datenschutz-akademie/wiki/pseudonymisierte-daten/ (accessed May 14, 2023).

[8] David Vasella. 2021. VK Berlin: Identifikation über die Stimme nicht möglich; Begriff der “Verarbeitung”. https://datenrecht.ch/vk-berlin-identifikation-ueber-die-stimme-nicht-moeglich-begriff-der-verarbeitung/ (accessed May 14, 2023).

[9] Datenschutz.org. 2023. Biometrische Daten: Besondere Schutzwürdigkeit bei sensibelsten Daten! https://www.datenschutz.org/biometrische-daten/ (accessed May 14, 2023).