DRAFT VERSION – TABLES AND FIGURES AT THE END OF THE DOCUMENT.

FROM THE FIELD TO THE WEB: IMPLEMENTING BEST-PRACTICE RECOMMENDATIONS IN DOCUMENTARY LINGUISTICS

1. Introduction

The imminent death of thousands of the world's languages over the next fifty years has driven linguists to search for solutions to the problems caused by this impending loss of linguistic diversity. A series of works published in the 1990s declared a state of crisis among the world's languages (e.g., Robins and Uhlenbeck (1991), Crystal (2000), Nettle and Romaine (2000)) and initiated a focus on what has become known as documentary linguistics: "a fairly independent field of linguistic inquiry and practice that is no longer linked exclusively to the descriptive framework." (Himmelmann 1998, p. 161) The crucial difference between documentary and descriptive linguistics is that the former concentrates on the *collection* of primary data (e.g., elicitation, recording, transcribing, translation) whereas the latter is concerned with the *analysis* of the primary data (see Himmelmann 1998, p. 162). Although in many cases it is impossible to stop a language from dying, refining the methodologies underlying documentary linguistics is essential to a good documentation of the language while speakers still exist. This documentation in turn provides present and future generations of linguists with empirical data for research, historians and anthropologists with information on a speech community's unique history and cultural heritage, and speakers themselves with essential material for their heritage preservation efforts (see, e.g., Nettle and Romaine (2000), Blythe and Wightman (2003)).

One major trend emerging in documentary linguistics over the last decade is the widespread use of computers for digital instead of paper- and tape-based language documentation. Bird and Simons (2003a) are concerned that "digital language documentation and description becomes inaccessible within a decade of its creation" because digital records "are often tied to software versions, file formats, and system configurations having a lifespan of three to five years." (p. 557). Following discussions in the academic community, notably among members of the Open Language Archives Community (OLAC) (Bird and Simons (2003b)), the International Standards for Language Engineering Metadata Initiative (ISLE-IMDI) (Wittenburg and Broeder (2002)) and the Electronic Metastructures for Endangered Languages Data initiative (EMELD) (Dry (2002)), Bird and Simons (2003a) construct detailed best-practice recommendations for the creation of digital language documentations and descriptions. The authors envision that once the academic community embraces a coherent set of best-practice recommendations, it will become possible to overcome the "unparalleled confusion in the management of digital language documentation and description." (p. 579). At the end of their in-depth article, Bird and Simons point out that

their recommendations are preliminary and call for "an open source revolution" in which agreed-upon data models for all of the basic linguistic types inform the development of open source tools using portable data formats, and all data are maintained in a network of interoperating digital archives (p. 580).

Although Bird and Simons' work marks a significant milestone in documentary linguistics, it leaves open three important questions: (1) How should the best-practice recommendations put forth in their paper be implemented in the field? (2) To what extent does feedback from this implementation necessitate some fine-tuning of their initial recommendations? (3) How can the successful implementation of their recommendations be used for research, teaching, and community outreach? This paper examines these questions in the context of work in the Texas German Dialect Project (TGDP) (cf. Boas (2003)), which has applied Bird and Simons' best-practice recommendations to the planning, implementation, and creation of the web-accessible Texas German Dialect Archive (TGDA). In particular, we discuss how and for what reasons Bird and Simons' recommendations for content, format, discovery, access, citation, preservation, and rights have been implemented in the context of the TGDP workflow, and examine situations where we were unable to follow these recommendations.

2. Background and Rationale

Texas Germans live mainly in a thirty-one county area of west-central Texas and are descendants of settlers who emigrated from middle and northern Germany, starting with the first large wave arriving between 1844 and 1848. Two world wars and gradual assimilation led to the loss of public institutional support for the widespread maintenance and use of German in such previously flourishing venues as German-language newspapers, schools, and churches. In the 1960s, about 70,000 speakers of Texas German remained in the central Texas area, notably in the communities of Fredericksburg, New Braunfels, Castroville, Schulenburg, and Brenham, among many others (Gilbert (1972), Salmons (1983), Nicolini (2004)). Today only an estimated 8-10,000 Texas Germans, primarily in their sixties or older, still speak the language of their forbearers fluently. Consequently, English has become the primary language for most Texas Germans in both private and public domains, whereas the reverse would have been true as late as the 1940s (Boas (2005)). With no sign of language shift being halted or reversed and fluent speakers almost exclusively in their 60s and older, Texas German is now critically endangered according to McConvell et al.'s (2002) levels of endangerment. As such, it is expected to become extinct within the next 30 years. Since the last in-depth recordings of Texas German were conducted in the 1960s (e.g., Eikel (1966), Gilbert (1972)), no detailed studies have traced more current developments of this German dialect.

At the moment, there is no data on the current state of Texas German available for linguistic, historical, and anthropological research or for heritage preservation efforts by the Texas German community. More importantly, since the 1960s there has been no effort made to document and archive this dialect.

The Texas German Dialect Project (TGDP) was founded at the University of Texas at Austin in September 2001 in an attempt to rectify this dearth of information by recording, documenting, archiving, and analyzing the remnants of the rapidly eroding dialect of Texas German.[i] The TGDP differs from similar projects in that it uses several freely available tools developed by the Max-Planck Institute for Psycholinguistics in Nijmegen that employ cross-platform standards such as UNICODE, XML, MPEG 1/2, and WAVE. The resulting archive (Texas German Dialect Archive (TGDA)) also differs from other archives—for example, the Archive for the Indigenous Languages of Latin America (AILLA)[ii] and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[iii] —in that it is not primarily concerned with digitizing and archiving existing recordings of endangered languages, but rather is the end-result of a research project whose workflow begins with data-collection in the field and ends with depositing digitized and annotated language materials in a web-accessible multimedia archive. In addition, the project workflow is driven by diverse needs for academic research, teaching, and outreach to the community, and the archive is intended for use by various groups: (1) linguists exploring the mechanisms underlying language change, language contact, and language death; (2) anthropologists focusing on the culture of Texas Germans; (3) historians trying to construct a detailed history of Texas Germans; (4) the general public interested in learning more about Texas Germans; (5) members of the Texas German community who wish to see their linguistic and cultural heritage preserved for future generations.

3. Data Collection

Bird and Simons propose to "make rich records of rich interactions, especially in the case of endangered languages." (p. 574) To achieve this goal, the TDGP developed a strategy that allows for a broad-scale collection of data representing a large number of linguistic features of current-day Texas German.[iv] After collecting and reviewing previous work published on Texas German (e.g., Eikel (1949, 1966, 1967), Gilbert (1972), Salmons (1983), Guion (1996), among others), we decided to collect three different data types, each located at different levels on Himmelmann's (1998: 178-182) 'spontaneity hierarchy' (planned vs. unplanned): Translation of English words, phrases and sentences into Texas German, open-ended sociolinguistic interviews, and conversations among Texas Germans.

3.1. TYPES OF DATA

The project collects three major types of data, which span the range of spontaneous (unplanned) to spontaneous (unplanned) speech, as outlined by Himmelmann (1998):

(1) Planned speech: Elicited pronunciation of 148 English words, phrases, and sentences contained in the Linguistic Atlas of Texas German (Gilbert (1972)). Examples include *a hairbrush*, *two goats*, *the door*, *The animal died out in the pasture*, *This chicken has long feathers*, and *Hang the picture over the bed*. The items in the list are read in English to informants, who are then asked to translate them into Texas German. Each elicitation lasts about 30 minutes. The use of word lists and sentences enables us to compare the results with data recorded over three decades ago by Gilbert, and provides a well-focused and well-controlled data set reflecting the distribution of specific phonological, morphological, and syntactic features of present-day Texas German.

(2) Moderately planned speech: Sociolinguistic interviews conducted in German, consisting of responses to questions about the informant's personal history (date and place of birth, place of origin of the informant's ancestors, etc.), together with approximately 140 questions on topics such as childhood activities, the community, religion, education, living conditions, tourism, government, language, and current activities (for example, *What was it like growing up on a farm in the 1920s*, *Why do you think Texas German is spoken less these days?*, and *How do you make sausage?*). The goal is to produce casual, relaxed conversation in which informants are given the chance to respond freely in Texas German without being asked to produce specific linguistic structures as with the word- and sentence-list translation task (see Boas (2003)). Each interview lasts between 40 to 60 minutes.

(3) Unplanned speech: Recordings in casual settings of conversations among Texas Germans. The recordings were made in two contexts: lunch and dinner conversations (during both food preparation and the meal itself), each lasting between 45-80 minutes; and performance of farm chores (e.g., cutting down trees and bushes, painting fences, digging holes and ditches), typically 30-120 minutes in length. Informants were given wireless microphones linked to a MiniDisc player that recording the conversations taking place during these activities. In these scenarios, there is almost no interference from the interviewer, as informants talk among themselves in a "natural" setting.

In addition to these types of data, additional data is collected by asking informants to fill out a ten-page questionnaire covering various demographic variables such as place and date of birth, gender, level of education, and religious affiliation.[v] The questionnaire also includes sections eliciting information about language use and language attitudes.

Questionnaires in combination with field notes describing the circumstances of the recording are collected as part of each interview, thereby fulfilling Bird and Simons' call for documenting the "multimedia linguistic field methods that were used" (p. 574).

3.2. Securing Informants' Consent for Digital Distribution

Frequently, intellectual property rights stand in the way of including older recordings in a language archive that is accessible without restrictions. In many cases, consent forms were never requested, and informants may be difficult to locate or have died in the years since the interview was conducted. Bird and Simons therefore recommend that intellectual property rights are fully documented (p. 579).

Because the TGDP is recording new interviews rather than digitizing existing ones, we are able to solicit informants' permission for the inclusion of their interviews in the archive beforehand.[vi] Before each interview, informants sign a three-page consent form explaining the nature of the project and procedures for the interview, and granting permission to use recorded interviews "as demonstrations in classrooms and on the internet." In addition, informants agree that "portions of the interview, including audio, video, and transcriptions, may be used for educational and professional purposes on the Internet." This follows recommendations for what Bird and Simons call the 'benefits' of rights, which ensure that the resource may be used for research purposes and that the use of primary documentation is not limited to the researcher, project, or agency responsible for collecting it (p. 579).

Preserving informants' anonymity is a critical issue in securing permission to archive interviews and is therefore a priority for the TGDP. Before an interview is processed and stored in the archive, it is assigned an identification number, and all mention of informants' names are dubbed over ('beeped out'). In addition, the names of specific people or information that may be used to identify the informant is removed from the transcriptions and the audio of the interview distributed over the Internet. This practice is crucial in recruiting and recording informants, as they frequently do not want the outside world to hear their personal opinions.[vii]

3.3. Recording Formats

Bird and Simons' top recommendation concerning format pertains to its 'openness.' They propose to "store all language documentation and description in formats that are open (i.e. whose specifications are published and nonproprietary)." In particular, they recommend that researchers "prefer formats supported by software tools available from multiple suppliers," and to "prefer formats with free tools over those with commercial

tools only" (p. 575). Another important issue is the quality of the recording.

Funding constraints were a major roadblock in the initial phase of the project. Due to limited resources we were not able to purchase DAT recorders, which produce uncompressed, high-quality recordings. In order to be able to begin recordings as soon as possible, we chose instead to purchase digital SONY MiniDisc (MD) recorders with super-directional SONY ECM-ZS9 zoom microphones for digital stereo recording, which are considerably cheaper and easier to use than portable DAT players. Although in principle MD's compressed ATRAC format produces recordings of lesser quality than DAT's uncompressed high-quality recordings, Campbell (2002) shows that the differences between MD and DAT are imperceptible in the frequency range of human speech and therefore interchangeable (on acoustical grounds) for most other types of linguistic analysis. Therefore, we decided not to follow Bird & Simons' recommendation pertaining to the highest quality of recording in order to be able to immediately begin with our fieldwork.

After recording interviews with MD recorders, we transfer the interviews in WAV format to our main workstation, thereby adhering to Bird & Simons' recommendation regarding open formats (see section 4.1). Since February 2002, we have recorded more than 350 hours of interviews with over 190 informants.

4. Processing of Field Recordings

The recordings go through a series of processing steps before they are stored in the Texas German Dialect Archive. An overview of the workflow is given in Table 1.

@@ Insert Table 1 here

4.1. DIGITIZATION AND LABELING

First, audio master files are transferred to our main workstation with Screenblast Soundforge in WAV format (48,000 Hz, 16-bit Stereo), which can be processed with free tools such as ELAN (EUDICO Linguistic Annotator, developed by the Max-Planck Institute for Psycholinguistics). Thus we follow Bird and Simons' recommendation to store "all language documentation and description in formats that are open" (p. 575) (see section 3.3 above).

Each audio master file is assigned a unique combination of numbers designating the interviewer, the informant, and the number of the interview conducted with that informant. Further information includes a number identifying the file as a master file and a letter showing whether the file is audio or combined audio/video. For example, the file name 1-

47-2-0-a.wav indicates that interviewer No. 1 conducted this interview with informant No. 47, and that this is the second interview with that informant. The '0' indicates that this file is a master file. When a copy of the master file is edited for transcription and translation at later stages of our workflow, each sub-section is identified by a series of consecutive numbers replacing the '0' (see section 4.2). Finally, the 'a' in the file name stands for 'audio' indicating that this is an audio master file. Subsequently, each master file is copied to the project's Linux-based file server (which is backed up daily to a secure off-site location). This procedure is influenced by Bird and Simons' recommendation to maintain language resources on digital mass-storage systems in order to enable easy backup and transfer to upgraded hardware (p. 578). It also follows their proposal to "ensure that copies of archived documentation and description are kept at multiple locations" (p. 578).[viii]

4.2. Editing of Field Recordings

Bird and Simons suggest providing "the primary recording (without segmenting it into clips)" (p. 574). At the same time, they propose to "limit any stipulations of sensitivity to the sensitive sections of the resource, permitting nonsensitive sections to be disseminated more freely" (p. 579). As noted above, to protect anonymity, informants' names are not included in the web-accessible data, and sections of interviews that could potentially be used to identify the informants are removed. However, while removing sections of the data to protect anonymity follows Bird and Simons' recommendations concerning informants' rights, it at the same time violates the recommendation to provide an unsegmented recording. It seems that these two recommendations are in conflict, and it is difficult, if not impossible, to adhere to both in many cases.

Bird and Simons propose to "publish digital resources using appropriate delivery media, e.g., web for small resources, and CD or DVD for large resources." Furthermore, they advise providing "low bandwidth surrogates for multimedia resources." (p. 576) By segmenting the field recordings into smaller sections, or 'media sessions' that vary in length between about thirty seconds and six minutes, users with low bandwidth are able to access the recordings more easily than if they had to download an entire interview of 40-60 minutes. A 'media session' is a segment of an interview that deals with a specific topic, such as the early history of New Braunfels or encounters with Native Americans during the 1860s, and may consist of a monologue, a dialogue, a song, or a poem, etc. The edited media sessions are saved in a separate folder on the project's file server,[ix] together with field notes that provide supplemental information about special circumstances surrounding the recording of the interview (number of speakers involved, location, etc.). Figure 1 illustrates the types of field notes stored in the database.

@@ Insert Figure 1 here

4.3. ANNOTATION

Student annotators in the Department of Germanic Studies at the University of Texas at Austin transcribe and translate media sessions using ELAN (EUDICO Linguistic Annotator).[x] ELAN allows for the definition of a multitude of so-called parent tiers (for each speaker in an interview) with associated sub-tiers in combination with synchronized playing of video and audio data (both for annotation and for subsequent re-playing). ELAN fulfills several of Bird and Simons' recommendations: its output adheres to their recommendation for accountability: "Transcriptions should be time-aligned to the underlying recording in order to facilitate verification." (p. 574). Second, the XML, WAV, MPEG1/2, and UNICODE formats supported by ELAN are open, thereby conforming to Bird and Simons' (2003a, p. 575) suggestion to "store all language documentation and descriptions in formats that are open." The third advantage of ELAN is the fact that it is free, thereby adhering to the proposal to "prefer formats with free tools over those with commercial tools only." (2003a, p. 575)

Annotators use a web interface to check out media sessions from the file server. When opening a new media session with ELAN, annotators also read the field notes describing the interview in order to learn more about the particular circumstances under which the interview was conducted. For example, the field notes section in Figure 1 informs annotators that interview 1-2-2 involves two speakers and two interviewers. This information helps to determine the number of parent tiers needed for annotation. Annotators define so-called parent tiers for each participant involved in a media session (interviewer(s) and informant(s)). Each parent tier is labeled with numbers to keep interviewer(s) and informant(s) apart (see Figure 2).

@@ Insert Figure 2 about here

The parent tier is used for transcribing the interview using a modified German orthography. Although we initially considered transcribing exclusively with the International Phonetic Alphabet (IPA), we soon discovered that such an endeavor is extremely time intensive and would also limit data access to people unfamiliar with the IPA. Using Standard German orthography for transcriptions does not represent Texas German adequately as it does not capture its peculiarities closely enough. For example, when two words such as *haben* ('have') and *wir* ('we') occur next to each other in fast speech, contraction occurs (see Wiese (2000)). Using Standard German orthography, this would still be transcribed as *haben wir*. Instead of employing IPA or Standard German orthography for transcriptions on the parent tiers, it was therefore decided

to use a modified German orthography, making it possible to capture different phenomena of Texas German in more detail. This choice allows us, for example, to transcribe contraction in more detail by representing our example as *hammwer* or *hammer*. Besides these practical considerations, the use of a modified German orthography also reflects the consideration of Bird and Simons' (2003a, p. 576) recommendation pertaining to scope of access. They suggest to "transcribe all recordings in the orthography of the language (if one exists)." As Texas German does not have its own orthography (but is mutually intelligible with spoken Standard German), the choice of a modified German orthography to represent the sound-form correspondences best implements Bird and Simons' recommendation. When transcribing with modified German orthography, annotators also employ a small set of basic markers in order to represent a variety of linguistic information on pauses (indicated by three dots '…'), filler sounds (indicated by 'uh' or 'hm'), or code-switching (indicated by square brackets, e.g., '[And then] geh ich nach Haus.' (And then I go home)). Following Bird and Simons' (2003a, p. 575) recommendation to document "punctuation and formatting (…) to represent the structure of information", we provide a list of markup conventions in the Texas German Dialect Archive (see section 5.1).

With the parent tiers in place, annotators define additional sub-tiers for translation, IPA, and general comments for each parent tier. These sub-tiers are time-aligned with their respective parent tiers and allow for inclusion of other types of information besides transcriptions in modified German orthography. The translation tier is used to provide a consistent word-by-word translation into English so that users not familiar with German are able to get an idea of the content and structure of each media session. The IPA tier is used in selected cases to transcribe phonological phenomena that are of interest to linguists studying Texas German. For example, over the past fifty years Texas German rounded front vowels have become progressively unrounded (Eikel (1966), Boas (2002)) (see Endnote 25). However, there is still a number of speakers whose speech exhibits variation between unrounded and rounded front vowels (see Boas et al. (2004)). The IPA tier represents such variations precisely by visualizing the differences with two distinct phonetic symbols: [i] representing the unrounded vowel and [œ] representing its rounded counterpart. A general-purpose comments tier allows annotators to note particularities about an informants' use of Texas German, if necessary.

The ELAN window includes, among other things, a waveform viewer, a subtitle viewer, and a timeline viewer aligned to the same time point (see Figure 2). Annotators first listen to sections of the media session to identify the speakers, then mark the waveform and click on the respective tier(s) for annotation. Each interview may be split into ten or more media sessions, and thus different annotators may annotate media sessions belonging to the same interview.

When a media session is saved, ELAN automatically creates an XML-compatible file with an EAF extension whose name is the same as that of its corresponding media file (WAV). As Figure 3 illustrates, the EAF file contains the annotations in combination with time stamps linking the annotation to the corresponding WAV file.[xi]

@@ Insert Figure 3 about here

4.4. Quality Control

In order to ensure consistent quality of the annotations, native speakers of Standard German who are graduate students at the University of Texas at Austin validate the annotated media sessions by correcting mistakes made by annotators and checking different media sessions belonging to the same interview for consistency. Graduate students conducting quality control use ELAN to listen to WAV files while simultaneously checking the corresponding EAF files for mistakes. While we regularly check for inter-annotator agreement by having all annotators transcribe a particular file every four weeks, quality control is still needed to correct possible inconsistencies between annotators. These procedures are influenced by the following considerations: (1) student annotators vary with respect to skill sets, largely depending on how long they have been with the project; (2) most student annotators are native speakers of English. Although their German skills are often near-native, we have found that native speakers of German will catch mistakes when conducting quality control.

4.5. Depositing Files in the Texas German Dialect Archive

In order to facilitate access to the recordings in combination with their transcriptions and translations, the Texas German Dialect Archive is structured around a MySQL database containing a variety of files whose formats are guided by Bird and Simons' best-practice recommendations pertaining to accountability, openness of format, rendering, and citation (p. 574-575). Each set of related media sessions includes the unsegmented original recording and the annotated WAV and EAF files, together with MP3 and HTML versions of each WAV and EAF file. This preserves the original recording for validation purposes and provides human-readable, low-bandwidth versions of all materials.

In addition to the primary and annotated data, the MySQL database includes a separate table for metadata information based on the informants' biographical questionnaires. The metadata includes the place and date of the recording, the place and date of the informant's birth, the gender, the childhood residence, the current residence, the level of education, the language(s) spoken in parents' home before elementary school, and the language(s) of instruction in elementary school. IN addi-

tion, each file is associated with an additional thirty-eight metadata values based on the IMDI metadata schema for endangered languages (see Johnson and Dwyer 2002). These include (1) general facts information (project, collector, content, participants, resources; (2) content subschema (interaction, explanation, performance, modality, communication context, languages, task, description, keys, register, style); (3) non-content subschema (ID, type, role, name, language, ethnic group, age, sex, education, origin, occupation); and (4) specific metadata resource schema (resource link, type, size, format, access, quality, recording conditions, position, content encoding, character encoding, software).

Bird and Simons point out that it is important to "provide complete citations for all language resources used" (p. 576) and that one should "use the metadata record of a language resource to document its relationship to other resources" (p. 577). To enable users to identify how files in the database are related to each other, each media session is assigned a unique combination of numbers identifying the informant, the interviewer, the number of the interview, and the number of the media session as it relates to the interview (for details, see section 4.2). Inclusion of the file extension serves to identify the different formats (WAV, EAF, MP3, HTML) of each file as belonging to the same media session following Bird and Simons' recommendation to "provide a formal means by which the components of a resources may be uniquely identified." (p. 577).[xii]

5. Accessing Files in the Texas German Dialect Archive

Making resources accessible to the user community is a central requirement of any archive. As of September 2006, we have conducted interviews with more than 190 informants, yielding a total of more than 350 hours of recordings. Of the 350 digitized hours, about 130 hours are publicly available, with the remaining hours at various stages of the workflow. Following Bird and Simons' (2003a, p. 576) call to "publish digital resources using appropriate delivery media, e.g. web for small resources", annotated Texas German recordings are made available by the TGDA over the World Wide Web.

5.1. Process of Access

Several existing language archives such as AILLA (Archive of the Indigenous Languages of Latin America) have a graded access system for its users. Such systems have been put in place to ensure that the rights of informants and their communities are not violated, especially when there is no information available from the depositors of the recordings on how to use the materials. This policy follows Bird and Simons' recommendation to "ensure that the intellectual property rights relating to the resource

are fully documented" (p. 579)[xiii] In the case of the TGDA, implementing this recommendation is not an issue because all informants have given permission to digitize their interviews and to use portions of them on the Internet. The TGDA also does not have to "document all restrictions on access as part of the metadata" or "document the process for access as part of the metadata, including licenses and charges" (Bird and Simons 2003, p. 576), because access to the data is (so far) unrestricted.[xiv] However, since one of the TGDA's goals in providing access to Texas German dialect materials is to ensure the ethical and responsible use of these materials, it requires users to register with the archive.

From the home page of the Texas German Dialect Project (http://www.tgdp.org), which includes a wealth of information on Texas German history, geography, and culture, users may access the TGDA after agreeing to the terms and conditions of the archive.[xv] This follows Bird and Simons' (2003a, p. 579) proposal to "ensure that there is a terms-of-use statement that clearly states what a user may and may not do with the materials." The log-in protocol fulfills four goals: (1) to make users agree to the terms and conditions of use of the archive before they access any data; (2) to exclude a user's access to the archive if the archive's conditions of use are not followed; (3) to have an inventory of users accessing the archive; and (4) to know what types of data are accessed by individual users.[xvi]

@@ Insert Figure 4 about here

5.2. Ease of Access

Users choose between two different graphical user interfaces to access recordings and the accompanying transcriptions contained in the database. The first consists of digitized maps from the Linguistic Atlas of Texas German (Gilbert (1972)). Users start by viewing a general map of Texas outlining the areas in which Texas German is spoken. After clicking on a specific area, the user is presented with a new window detailing the counties with individual locations for which Texas German recordings are available. Clicking on a specific location, e.g., Fredericksburg, displays a pop-up window containing a list of media session names with their length and formats in combination with their unique ID numbers (see Figure 4). The media sessions, which are available for download in different formats, are labeled with short titles summarizing their content (e.g., "Growing up on a farm", or "Walking to church in the winter"). Linguists interested in time-aligned transcriptions and audio files with low compression rates may download WAV formats in combination with the corresponding EAF annotation files. Alternatively, users may click on a file name, which opens a new window with an MP3 player and plays the audio portion of the media session. The same window contains a transcription and translation of the media session in

HTML (see Figure 5). Users can read the transcript and its corresponding translation while the audio file is playing to better understand the contents of the recording.

@@ Insert Figure 5 about here

The second option for accessing the contents of the database is via a query system that enables searches based on metadata associated with the sessions, thus providing access on the basis of more detailed information than provided by the general-purpose map-interface. The user may conduct searches based on any combination of nine metadata elements: place of recording, date of recording, date of birth, gender, childhood residence, current residence, languages spoken by parents before elementary school, languages spoken by teacher in elementary school, and level of education. The result is a list of files matching the search criteria, from which users may choose to download the high-bandwidth WAV file and its associated EAF file, or simply click on a file name to listen to a media session in MP3 format while reading its transcript in HTML.

5.3. Citation of Archived Materials

As one of the main goals of the dialect archive is to provide primary linguistic data, the question arises as to how users who are interested in using the data for their research, teaching, or community outreach efforts should cite archival materials. To this end, Bird and Simons suggest that linguists "furnish complete bibliographic data in the metadata for all language resources created" (p. 576). While bibliographic data is included in the metadata associated with each session, citing the electronic resource *per se* is a more complicated issue. Bird and Simons propose to "provide instructions on how to cite an electronic resource from the collection as part of the web site for a digital archive" (p. 576-577). In order to protect the privacy of our informants, we decided to adopt a modified version of the International Standards Organization's guidelines for citing online resources (ISO 690-2).[xvii] In our adaptation, the first part of a complete reference includes the name of the researcher who collected the recording, even if his or her voice does not appear in the recording, followed by the year of the recording, a descriptive title, the name of the web site, and the unique file ID identifying a media session; for example,

> Boas, Hans C. (2002): "Different types of Country Schools". [online] http://www.tgdp.org: The Texas German Dialect Project. 1-25-1-7-a.

Inclusion of the unique file ID fulfills a number of Bird and Simons' recommendations concerning citation of language documentation materials.

One of these recommendations pertains to the immutability of citations: "Provide fixed versions of a resource, either by publishing it on a read-only medium, or by submitting it to an archive that ensures immutability" (p. 577). Once deposited into the web-based archive, the contents of a media session are not changed. Therefore, there is no need to "distinguish multiple versions with a version number or date, and assign a distinct identifier to each version." (p. 577)[xviii] The unique file ID also complies with Bird and Simons' recommendations dealing with 'granularity' by providing a formal means by which the components of a resource may be uniquely identified (the file ID always points to one master file).

5.4. Using the Archive for Research, Teaching, and Outreach

Over the past thirty years, there have been no systematic efforts to gather large amounts of Texas German data to support detailed studies that trace the development of the linguistic structures of this German dialect. This has led to a serious gap in the study of Texas German, especially given that studies of other eroding dialects, such as Pennsylvania German (Raith (1992)), Brule Spanish (Holloway (1997)), and Jersey Norman French (Jones (2001)), have shown that the rate of language change in moribund dialects is unpredictable across different speech communities.

A preliminary analysis of the first thirty hours of recordings conducted in Fredericksburg, New Braunfels, and Freyburg between February 2002 and March 2003 has revealed a number of interesting linguistic features that have broad implications for research, both on the current state of Texas German and on language contact and language change in general. For example, a preliminary analysis of our data suggests that there does not seem to be a single coherent Texas German speech community across central Texas. That is, German immigrants coming to Texas between the 1830s and 1890s came primarily from four different dialectal regions in central Europe: the central west Duchy of Nassau (located in the modern German states of Hessen, Rheinland-Pfalz, and Nordrhein-Westfalen), northern Germany (from the areas around Hamburg and Bremen), eastern Germany (Thuringia and Saxony), and Alsace (now a part of France) (see Biesele (1928)). This mix of different donor dialects makes it difficult to define a coherent "Texas German Dialect". The widespread linguistic variation existing between different Texas German speech communities at the lexical, phonological, morphological, and syntactic level has been recorded by Gilbert (1972). A preliminary analysis of our recordings from the Fredericksburg, New Braunfels, and Freyburg areas confirms the widespread variation noted by Gilbert only to a certain degree.

To determine the degree of variation between different locations, we chose to analyze the first type of data, namely our informants' responses to the word and sentence lists from Gilbert's (1972) Linguistic

Atlas of Texas German (see section 3.1 above). One of the test cases considered during our pilot project focused on the different realizations of /r/ in Texas German. For example, in the rural area surrounding Freyburg (Fayette County), Gilbert (1972) shows that the pronunciation of the word *ihr* ('her' (possessive pronoun)) includes an American-English retroflex continuant [ɹ]. In contrast, in the areas surrounding New Braunfels (Comal County) and Fredericksburg (Gillespie County), Gilbert (1972) reports an American-English retroflex continuant [ɹ] as well as an apical trilled tap [r] for the same word. Our reproduction of Gilbert's (1972) data for the 2002/2003 recordings shows the same type of American-English retroflex continuant for Freyburg. We found that the apical trilled tap is now used only on rare occasions in the New Braunfels and Fredericksburg areas. Instead, the majority of our informants from these two areas overwhelmingly use the American-English retroflex.[xix] Preliminary analysis of the data demonstrates that the regional variation found in different locations some thirty years ago is no longer very distinct; that is, there is a clear trend toward emerging differences between rural and urban areas. Whereas speakers in the Freyburg area continue to pronounce their /r/ as an American-English retroflex continuant, the speech of New Braunfels and Fredericksburg informants has changed over the past three decades (see Boas et al. (2004)).[xx] Of the three possible types of change affecting dialectal speech (reduction of dialectal variety, maintenance of dialectal variety, expansion of dialectal variety (see Wagener (2002: 274))), our preliminary data on /r/ for Fredericksburg and New Braunfels suggests a reduction of dialectal variety. A more detailed investigation into the current distribution of /r/ is currently being carried out using additional data as it is added to the archive. However, our preliminary results based on data already contained in the Texas German Dialect Archive illustrate the way in which the archive can be utilized to answer research questions having to do with dialect formation, language contact, and language change.

The dialect archive has also been integral in developing and teaching linguistics courses. One of the main problems typically encountered by instructors when teaching linguistics classes is that students are asked to apply their knowledge of theoretical concepts by solving printed exercises in textbooks or provided by the instructor. Whereas these traditional exercises enable students to practice solving linguistic problems, their lack of relevancy and immediacy generally results in pedagogic problems on two levels. First, traditional exercises fail to demonstrate the pervasiveness of linguistic problems in speech communities students are exposed to in their daily lives and create the false picture of linguistics as the study of exotic and remote languages. Second, traditional exercises fail to excite and motivate students to conduct further research and learning on their own. Even when readings, class lectures, and exercises are augmented by recordings of interviews in class, students are usually left

with no chance of using these recordings by themselves outside of class to work on homework assignments or conduct research of their own.

The web-based multimedia archive of Texas German seeks to overcome these problems by giving students access to interview data in order to conduct independent research on Texas German, both in and outside the classroom. The TGDA's combination of audio clips with transcribed and translated textual data enables students to re-create the experience of sitting directly across from the Texas German informants as they talk. This high level of engagement has already resulted in an array of original student research projects on Texas German language, history, and culture.

Finally, the TGDA has played an essential role in community outreach and heritage preservation efforts. The staff of the Texas German Dialect Project is regularly invited to give guest lectures to local genealogical societies on the status of Texas German. These lectures raise awareness in the community about the current status of Texas German and enable the TGDP to connect with local schools and preservation societies eager to use TGDA materials for educational programs about Texas language, history, and culture. One of the ways in which the dialect archive will be used in the future is by setting up computer terminals in local museums to enable access to the archive. Museum visitors will then have immediate access to the archive and can listen to the stories and learn more about the history, culture, and language of the Texas German community. Although the Texas German Dialect Project has made forays into the area of language documentation, as Himmelmann (1998: 188/89) points out, one of the ways in which linguists can get involved with the community is by engaging in "language maintenance work, which may be of greater interest to the community than just a documentation." However, this interest does not seem to be shared by the Texas German community; as one informant put it: "We know Texas German is dying out, but that's the way it is. We don't need the language any more as English is more useful." However, feedback has been consistently positive regarding outreach to genealogical and preservation societies, schools, and museums.

5.5. FURTHER ISSUES

Two sets of Bird and Simons' best-practice recommendations have not been addressed in detail in the preceding sections, as they are currently being worked out. The first concerns the discovery of language resources via the Open Language Archives Community (OLAC) (see Bird and Simons (2003b)). Bird and Simons suggest to "list all language resources with an OLAC repository" (p. 576) in order to facilitate their widespread discovery.[xxi] They go on to point out that one of the most important standards for listing a resource is "a standard for identifying languages" which in turn allows OLAC metadata "to be mapped to the

more general-purpose Dublin Core metadata set and disseminated to the broader community of digital libraries" (p. 573). Since there is no SIL language code for Texas German in place, we have not yet been able to integrate our own metadata with those of OLAC. However, as soon as Texas German has its own SIL language code, we will begin with the mapping of our metadata to OLAC in order to allow the Texas German data to be discovered more easily.

The second set of best-practice recommendations currently under consideration concerns the preservation of language resources. Bird and Simons suggest committing "all documentation and description to a digital archive that can credibly promise long-term preservation and access" (p. 578). At the moment, the dialect archive is housed on an in-house file server, which is backed up daily to a larger college-wide file server. In order to "ensure that the archive satisfies the key requirements of a well-founded digital archive" (p. 578), we are planning to integrate the TGDA with the Digital Library Services at the University of Texas at Austin (see section 4.1). This step will not only provide offsite backup, but will also ensure that the materials are migrated to new formats and media/devices over time.

6. Conclusion

This paper discusses the implementation of many of Bird and Simons' (2003a) best-practice recommendations for language documentation in the Texas German Dialect Project, in particular, their recommendations concerning content, format, discovery, access, citation, preservation, and rights. We have shown that although these recommendations have resulted from years of discussions among documentary linguists, not always possible due to the need to secure informant privacy, absence of transcription conventions for dialects such as Texas German, and challenges in the areas of time and funding.

The central aim of this article, however, is to show how the infrastructure of the TGDA successfully overcomes one of Bird and Simons' main concerns, namely the fact that a "substantial fraction of the resources being created can only be reused on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of a few years" (p. 579). By using freely available cross-platform tools such as ELAN, it is possible for others to download and re-use our data for their purposes without having to resort to costly commercial tools. Employing open standards such as MPEG, WAV, XML, and UNICODE formats has the clear advantage of cross-platform compatibility and the promise for longer-term accessibility than resources that primarily rely on proprietary formats.

The advantages of the TGDA's infrastructure are not only relevant for documenting Texas German, but are also of importance to other language documentation efforts, which seek to produce data that remain

accessible for decades after their creation. The results presented in this paper represent a significant first step towards clarifying the relationships between different types of best practice recommendations, and as such are intended to spark further discussions among documentary linguists eventually leading to "a broad consensus about the design and operation of common digital infrastructure for the archiving of language documentation and description" (Bird and Simons, 2003a, p. 580).

An important point raised in this paper concerns the interdependence of data collection and data archiving. Himmelmann points out that "historically speaking at least, it has been the case that the collection activity has never received the same attention within descriptive linguistics as the analytic activity" (1998, p. 163). He goes on to offer the criticism that "methodological issues with respect to obtaining and presenting primary data have never been dealt with in depth within descriptive linguistics" (p. 164). However, in contrast to many other language archives that are primarily concerned with preserving existing recordings, the Texas German Dialect Project is concerned with the collection and annotation of primary data, a fact that has a major impact on the editing, transcription and translation of the data, as well as the presentation of and access to the data contained in the dialect archive. As such, the practices outlined here should serve to inform the design and development of future archival projects that not only preserve but create primary linguistic data.

Finally, the experiences in the TGDA show that it is necessary to consider in detail how particular implementations of best-practice recommendations at different stages in the workflow influence the structure of the resulting language archive. Consideration of these issues will hopefully lead to an enhanced set of best-practice recommendations beginning with the collection of primary data and ending with their archiving. This, in turn, will increase the likelihood that the work of documentary linguists will survive successfully in the long term.

Acknowledgements

This paper is dedicated to Walt Wolfram, who introduced me to linguistic fieldwork. His enthusiasm and thoughtful advice have broadened my linguistic interests in many ways. I thank Heidi Johnson, Keith Walters, Hans Ulrich Boas, and Jana Thompson for helpful comments on earlier versions of this paper. An anonymous reviewer provided additional suggestions that were extremely useful. The Texas German Dialect Project is grateful for the generous financial and logistical support from the Dean of the College of Liberal Arts, the Liberal Arts Instructional Technology Services, The Division of Instructional Innovation and Assessment (all of the University of Texas at Austin), and Humanities Texas (formerly the Texas Council for the Humanities) grant #2003-2950. The author also

gratefully acknowledges the logistical support of the Department of Germanic Studies at the University of Texas at Austin.

References

Biesele R.L. (1928) *A History of the German Settlements in Texas*. Ph.D. dissertation. UT Austin.

Bird S., Simons G. (2003a) Seven Dimensions of Portability for Language Documentation and Description. *Language*, 79(4), pp. 557-582.

Bird S., Simons G. (2003b) The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing*, 18(2), pp. 117-128.

Blythe J., Wightman G. (2003) The Role of Animals and Plants in Maintaining the Links. In Blythe J., Brown R. M. (eds.), *Proceedings of the Seventh Foundation for Endangered Languages Conference*, Broome, Australia, pp. 69-77.

Boas H.C. (2002) The Texas German Dialect Archive as a Tool for Analyzing Sound Change. In Austin P., Dry H. A., Wittenburg P. (eds.), *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, Spain, pp. 28.1-28.4.

Boas H.C. (2003) Tracing Dialect Death: The Texas German Dialect Project. In Larson J., Paster M. (eds.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, Berkeley, California, pp. 387-398.

Boas, H.C. (2005) A Dialect in Search of its Place: The Use of Texas German in the Public Domain. In Cravens C., Zersen, D. (eds.) *Transcontinental Encounters: Central Europe Meets the American Heartland*, Concordia University Press, Austin, pp. 78-102.

Boas H.C., Ewing K., Moran C., Thompson J. (2004) Towards Determining the Influence of Internal and External Factors on Recent Developments in Texas German Phonology. In Arunachalam S., Scheffler T. (eds.) *University of Pennsylvania Working Papers in Linguistics*, Philadelphia, Pennsylvania, pp. 47-59.

Campbell N. (2002) Recording and Storing of Speech Data. In Austin P., Dry H. A., Wittenburg P. (eds.), *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, Spain, pp. 6-1-6.3.

Crystal D. (2000) *Language Death.* Cambridge: Cambridge University Press.

Dorian N. (1973) Grammatical Change in a Dying Dialect. *Language* 49, pp. 413-438.

Dry H. (2002) E-MELD: Overview and Update. In Austin P., Dry H.A., Wittenburg P. (eds.), *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, Spain, pp. 3.1-3.8.

Eikel F. (1949) The Use of Cases in New Braunfels German. *American Speech,* 24, pp. 278-281.

Eikel F. (1966) New Braunfels German: Part II. *American Speech*, 31, pp. 254-260.

Eikel F. (1967) New Braunfels German: Part III. *American Speech,* 32, pp. 83-104.

Gilbert G. (1972) *Linguistic Atlas of Texas German*. Austin: University of Texas Press.

Guion S. (1996) The death of Texas German in Gillespie County. In Ureland P.S., Clarkson I. (eds.), *Language contact across the North Atlantic*. Niemeyer, Tübingen, pp. 443-463.

Himmelmann N. (1998) Documentary and Descriptive Linguistics. *Linguistics*, 36, pp. 161-195.

Holloway C. (1997) *Dialect Death. The Case of Brule Spanish*. Benjamins, Amsterdam/Philadelphia.

Johnson, H. (2002) The Archive of the Indigenous Languages of Latin America: Goals and Visions. In Austin P., Dry H.A., Wittenburg P. (eds.), *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, Spain, pp.13.1-13.4.

Johnson H., Dwyer A. (2002) Customizing the IMDI Metadata Schema for Endangered Languages. In Austin P., Dry H.A., Wittenburg P. (eds.),

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, Spain, pp. 5.1-5.4.

Jones M. (2001) *Jersey Norman French. A Linguistic Study of an Obsolescent Dialect*. The Philological Society, Oxford and Boston.

McConvell P., Amery R., Gale M.A., Nicholls C., Nicholls J., Rigney L., Tur S. (2002) *Keep that Language Going! A Needs-based Review of the Status of Indigenous Languages in South Australia.* AIATSIS, Canberra.

Nettle D., Romaine S. (2000) *Vanishing Voices. The Extinction of the World's Languages*. Oxford University Press, Oxford.

Nicolini, M. (2004) *Deutsch in Texas*. LIT-Verlag, Münster.

Raith J. (1992) Dialect Mixing and/or Code Convergence: Pennsylvania German? In Burridge, K., Enniger, W. (eds.), *Diachronic Studies on the Languages of the Anabaptists*, Brockmeyer, Bochum, pp. 152-165.

Robins R. H., Uhlenbeck E.M. (eds.) (1991) *Endangered Languages*. Berg, Oxford.

Salmons J.C. (1983) Issues in Texas German Language Maintenance and Shift. *Monatshefte* 75(2), pp.187-196.

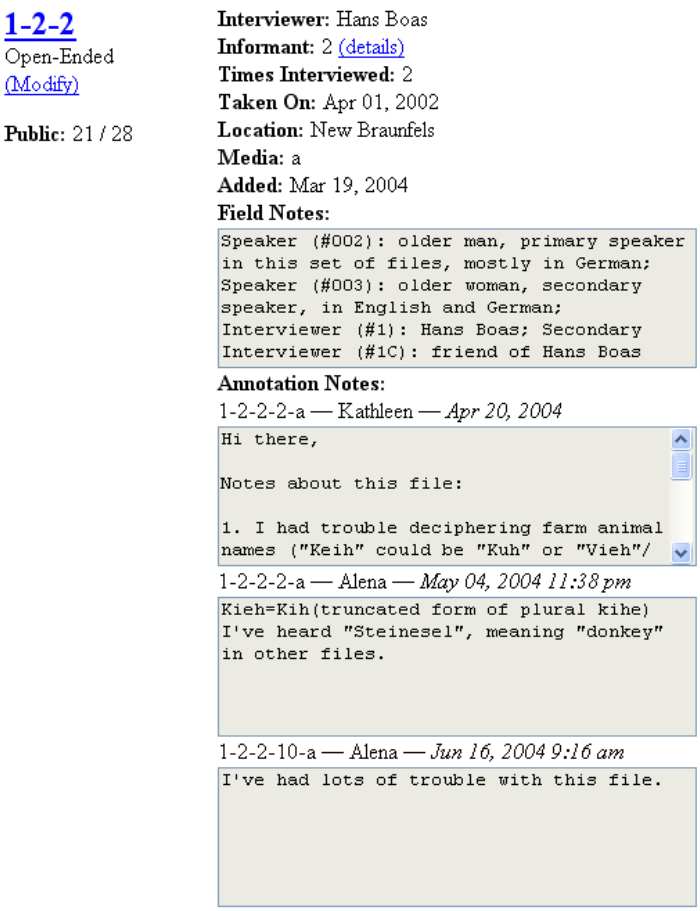
Simons G. (2002) SIL Three-letter Codes for Identifying Language: Migrating from in-house Standard to Community Standard. In Austin P., Dry H.A., Wittenburg P. (eds.), *Proceedings of the International Workshop on Resources and Tools in Field Linguistics held in conjunction with the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 22.1-22.8.

Wagener P. (2002) German Dialects in Real-Time Change. Journal of Germanic Linguistics, 14(3), pp. 271-285.

Wiese R. (2000) *The Phonology of German*. Oxford University Press, Oxford.

Wittenburg P., Broeder D. (2002) Metadata and Semantic Web. In Austin P., Dry H.A., Wittenburg P. (eds.), *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, Spain, pp. 4.1-4.14.

Figure 1: Field notes provided for annotators when checking out files



DRAFT VERSION – TABLES AND FIGURES AT THE END OF THE DOCUMENT.

Figure 2: ELAN Annotation with multiple tiers

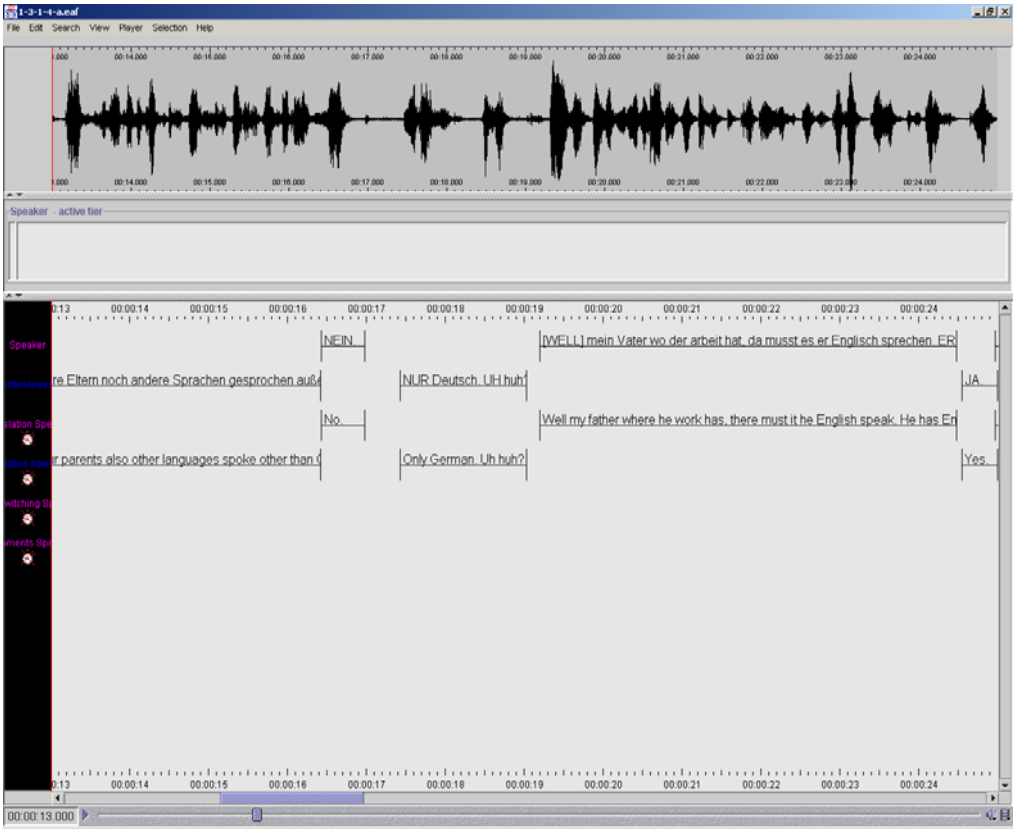
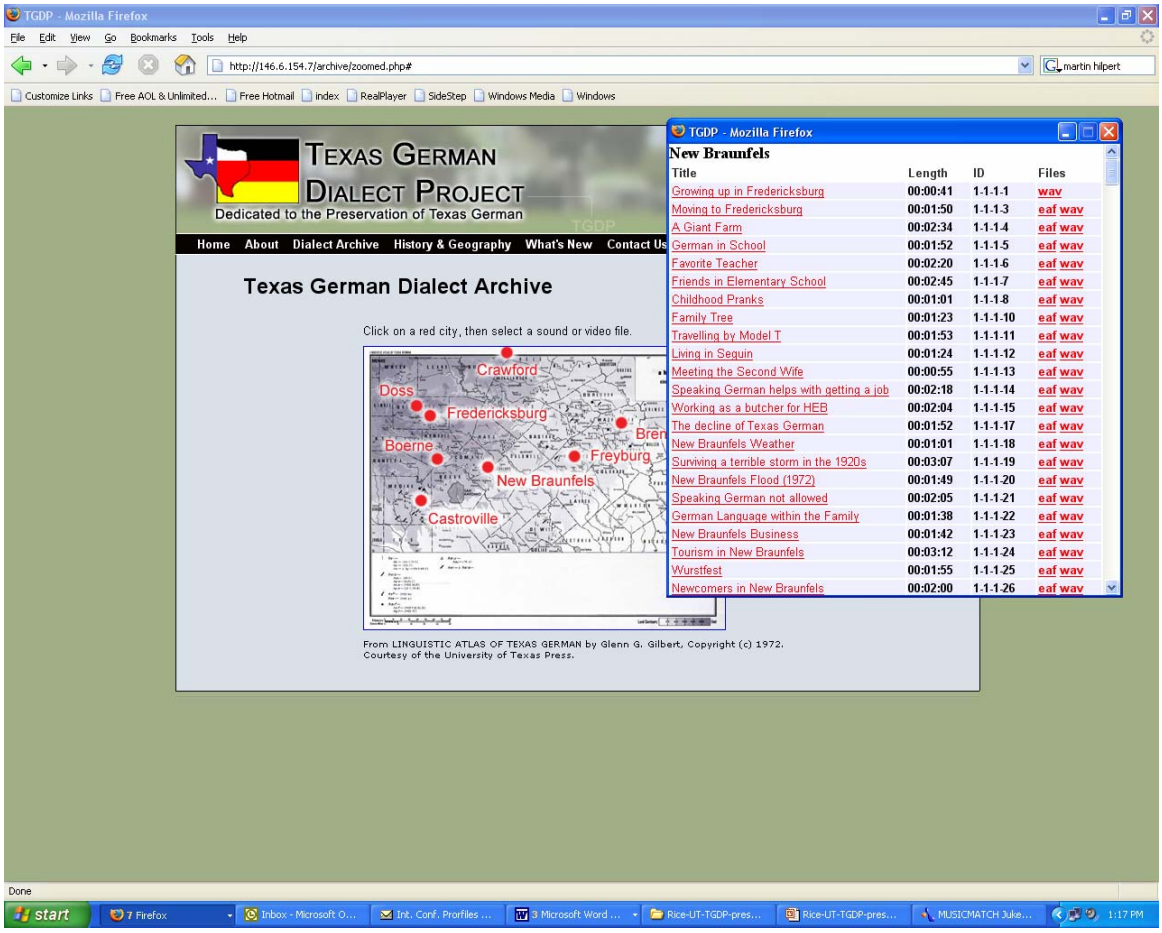


Figure 3: XML-compatible EAF transcription file produced by ELAN

```
      <TIME_SLOT TIME_SLOT_ID="ts35"
TIME_VALUE="73461"/>
      <TIME_SLOT TIME_SLOT_ID="ts36"
TIME_VALUE="76841"/>
</TIME_ORDER>
<TIER DEFAULT_LOCALE="en" LINGUIS-
TIC_TYPE_REF="Speaker 94"
 PARTICIPANT="Speaker 94" TIER_ID="Speaker 94">
        <ANNOTATION>
            <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1"
             TIME_SLOT_REF1="ts7"
TIME_SLOT_REF2="ts8">
                <ANNOTATION_VALUE> OTTO
[NAME]</ANNOTATION_VALUE>
            </ALIGNABLE_ANNOTATION>
        </ANNOTATION>
        <ANNOTATION>
            <ALIGNABLE_ANNOTATION ANNOTATION_ID="a2"
             TIME_SLOT_REF1="ts11"
TIME_SLOT_REF2="ts12">
                <ANNOTATION_VALUE>UH in Oktober
                 neunzehneinunddreis-
sig</ANNOTATION_VALUE>
            </ALIGNABLE_ANNOTATION>
        </ANNOTATION>
        <ANNOTATION>
            <ALIGNABLE_ANNOTATION ANNOTATION_ID="a3"
             TIME_SLOT_REF1="ts15"
TIME_SLOT_REF2="ts16">
                <ANNOTATION_VALUE>IN Sisterdale auf
mein Platz wo ich
                 jetzt noch wohn
hier</ANNOTATION_VALUE>
            </ALIGNABLE_ANNOTATION>
        </ANNOTATION>
        <ANNOTATION>
            <ALIGNABLE_ANNOTATION ANNOTATION_ID="a4"
             TIME_SLOT_REF1="ts19"
TIME_SLOT_REF2="ts20">
                <ANNOTATION_VALUE>MEIN Vader kam
hierüber in
                  achtzehnsiebensibzig in
uh</ANNOTATION_VALUE>
            </ALIGNABLE_ANNOTATION>
        </ANNOTATION>
```

DRAFT VERSION – TABLES AND FIGURES AT THE END OF THE DOCUMENT.

Figure 4: Accessing data in the Texas German Dialect Archive (www.tgdp.org)



DRAFT VERSION – TABLES AND FIGURES AT THE END OF THE DOCUMENT.

Figure 5: HTML transcript of media session

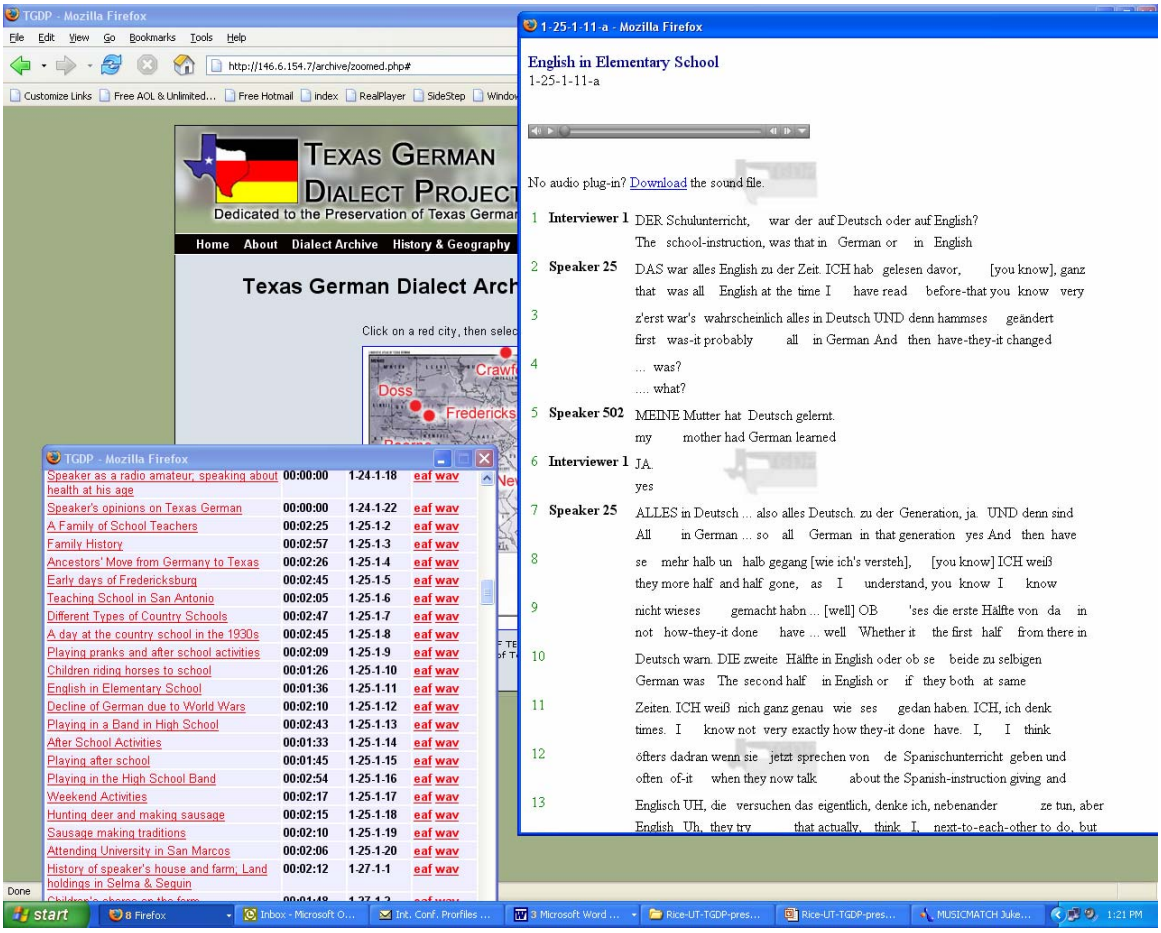


Table 1: Workflow of the Texas German Dialect Project.

Stage	Task
1. Data Collection	- fill out consent form
	- recording of interviews
	- collection of metadata information
2. Editing of Recordings	- storage of master file on main file server
	- recordings are made anonymous
	- editing of master file copies into "media sessions"
	- assigning unique file ID numbers to protect informants' privacy (e.g. 1-47-2-0-a)
3. Annotation with ELAN	- annotation (transcription and translation)
	- quality control of annotations
	- annotations are saved in XML-compatible EAF format
4. Storage in the TGDA	- WAV and EAF versions of media sessions are converted into MP3 and HTML versions
	- each media session is associated with its appropriate metadata information

[i] For this project, Texas German was chosen for three reasons. First, there exists previous work on the dialect (Eikel 1949, Gilbert 1972), which makes it easier to analyze changes that have occurred over the past century. Second, the majority of Texas German speakers live within a three hour radius from Austin. This close proximity allows us to interview a greater number of speakers than would be possible if our fieldwork sites were farther away. Finally, a large percentage of Texans are of German heritage. Working with speakers from this community has not only enabled us to obtain funding from the University of Texas, it has also allowed us to present the results of our efforts to local preservation societies and genealogical clubs (community outreach).

[ii] See Johnson (2002) and http://www.ailla.org

[iii] See http://paradisec.org.au

[iv] Thus far, fieldwork has been conducted in Fredericksburg, New Braunfels, Boerne, Comfort, Victoria, Houston, Brenham, Freyburg, Doss, Spring Branch, and Crawford, Texas. Interviews take place at informants' homes, at nursing homes, in local cafés, on their farms, or at local churches.

[v] The age range of informants as well as their proficiency in Texas German varies a great deal. The oldest informant to date is 94 years old; the youngest informant interviewed to date is 57 years old. Among the informants, there are fluent and semi-fluent speakers. Older fluent speakers are those who learned German as their first language at home (informants who are now in their 80s and 90s) and for whom German continues to be the dominant language. Younger fluent speakers typically learned English and German simultaneously as their native languages and speak fluent Texas German regularly with friends, family, and neighbors (see also Guion (1996)). In contrast, semi-fluent speakers in their 60s and 70s have never completely acquired Texas German and use it occasionally. As a result, their use of Texas German is characterized by a halting delivery (see also Guion (1996); cf. Dorian (1973)). To date, we have not been able to find any fluent or semi-fluent speakers younger than 57 years. The children of the youngest fluent and semi-fluent speakers know only a few words or phrases of Texas German.

[vi] Before data collection could begin, the Institutional Review Board of the University of Texas at Austin approved the procedures used to obtain the data. Texas German informants were found through a social network tracing process beginning with students and colleagues at the University of Texas at Austin.

[vii] As most informants are concerned about their privacy, they do typically not agree to their interview being videotaped. So far, we have only taped about 4 hours of interviews on digital video and are not planning on making them publicly available. Therefore, the remainder of this paper focuses primarily on our handling of audio recordings.

[viii] In order to fully implement Bird and Simons' (2003a, p. 578) suggestions regarding the safety of language documentation materials, we plan in the future to also "create a disaster recovery plan, such as that developed by the Syracuse University Library (1995), containing procedures for salvaging archived resources in the event of a disaster." To this end, we plan on integrating our materials with the Digital Libraries Services Division of the General Libraries at the University of Texas at Austin in the near future. See http://www.lib.utexas.edu/dlp/index.html

[ix] To facilitate collaborative work in different locations, programming staff of the TGDA developed a number of web-based tools that enable project members to access files over the web at different stages of the project's workflow.

[x] Currently, there is no open source license for ELAN, which would allow us to modify ELAN according to our needs. However, this issue has not been a problem for our project as MPI staff constantly updates ELAN and we have so far not had any issues with missing functionality. For our purposes, ELAN has a number of advantages over other transcription programs such as Transcriber (http://trans.sourceforge.net/en/presentation.php): First, it allows both audio and video annotation. Second, it interfaces directly with other software for the creation and look-up of metadata, such as the IMDI editor and IMDI browser, which are also developed by the MPI in Nijmegen. For a detailed description of ELAN, see http://www.mpi.nl/tools/elan.html

[xi] During our two-year long pilot project we have been somewhat successful at streamlining the annotation process. That is, the time it takes to annotate a media session greatly depends on a multitude of factors. Among them are (1) intelligibility of informants' speech; (2) length of time that annotators have spent with the project (ELAN takes some time to learn and annotators need to become familiar with the workflow and procedures of the project); (3) which version of ELAN was used (earlier versions tended to crash more frequently than newer ones); and (4) type of genre (monologues are typically easier to transcribe than dialogues with frequent turn taking).

[xii] As Texas German still does not have a three-letter SIL language code (cf. Simons (2002)), we have not yet mapped our metadata to other metadata schemes such as IMDI (Johnson and Dwyer (2002)) or OLAC (Bird and Simons (2003b)). Once the language code is in place, our metadata will be mapped to other metadata schemes to ensure greater discoverability of the Texas German Dialect Archive.

[xiii] Getting informants' consent for making the recordings available for research purposes by people outside of the University of Texas fulfills Bird and Simons' (2003a, p. 579) 'benefit of rights' recommendation: "a. Ensure that the resource may be used for educational purposes. b. Ensure that the use of primary documentation is not limited to the researcher, project, or agency responsible for collecting it."

[xiv] Two of Bird and Simons' (2003a, p. 576) recommendations pertaining to the process of access ("For resources not distributed over the web, publish online surrogates that are easy for potential users to access and evaluate", and "For resources not distributed over the web, publish online surrogates that are easy for potential users to access and evaluate") are difficult to implement at this point as both require planning with the help of long-term financial support. At this point, however, the project is limited by year-to-year support cycles.

[xv] New users are asked to register with the archive in order to gain access to the data. Users are asked to provide their names, affiliation, state/country, email address, and purpose for using the archive. Furthermore, users have to choose a user identification and password.

[xvi] For the terms-of-use statement of the TGDA, please see http://www.tgdp.org/archive/disclaimer.php

[xvii] http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm#5

[xviii] Technically, each file does have different versions, but these are just differences in format, not in content. That is, both HTML and EAF files contain the

same transcriptions and translations. Similarly, both MP3 and WAV files contain the same audio information, but compressed differently.

[xix] Our data show that the use of the American-English retroflex continuant is not limited to the coda of the (stressed and unstressed) syllable, but also occurs in other contexts such as the onset of syllables.

[xx] Similar variation and changes have been found for the unrounding of front rounded vowels. For example, based on an analysis of TGDA data, Boas et al. (2004) report that the variation between rounded and unrounded front vowels in New Braunfels documented by Eikel (1966) and Gilbert (1972) has changed over the past four decades. Eikel and Gilbert report for New Braunfels variations such as [y:bə□]/[i:bə□] for 'over' (Eikel (1966: 255)), [fynfciç] / [finfciç] for 'fifty' (Eikel (1966: 256)), [šø:n] / [še:n] for 'nice' (Eikel (1966: 25)]), [ky:ə] / [ki:ə] for 'cows' (Gilbert (1972: map 68)). A comparison of the Eikel and Gilbert data with the 2002/2003 TGDA data show that the unrounding of front vowels is now further progressed, with instances of rounded front vowels now extremely rare. When rounding is found among informants of the oldest generation, it appears to be random and inconsistent, instead of being distributed systematically as noted by Eikel (1966: 255).

[xxi] For more information on OLAC (the Open Language Archive Community), see http://www.language-archives.org.