Department of Spanish Research   Spanish

2006

Frustrations of the Documentary Linguist: The State of the Art in Digital Language Archiving and the Archive that Wasn't

Robert E. Vann
robert.vann@wmich.edu

Frustrations of the documentary linguist:
The state of the art in digital language archiving and the archive that wasn't

Robert E. Vann
Western Michigan University

1.0 Introduction

This paper is a qualitative review and critique of existing electronic language archives from the perspective of the documentary linguist. In today's day and age there are many online archives available for storage of and access to digital language data, among others: AILLA, ANLC, ASEDA, DOBES, ELRA, E-MELD, LACITO, LDC, LPCA, OTA, PARADISEC, Rosetta, SAA, THDL, and UHLCS. It is worth pointing out that I use the term *language archive* loosely in this paper, referring to a variety of organizations and resources such as those listed above that store and provide digital language data to different degrees and for different purposes. Technically speaking, however, not all of these organizations may be considered language archives by all involved (Anthony Aristar, personal communication, June 14, 2006). Indeed, there may not even be consensus agreement among linguists and others who work with language resources about what exactly constitutes a language archive. For example, E-MELD and Rosetta generally provide material as exemplary linguistic data, but they do not archive material as a rule. Moreover, ELRA and LDC may be more dedicated to accumulating and selling electronic linguistic resources than they are to preserving them for posterity. DOBES may store data and research studies for its project researchers, but, to my knowledge, DOBES does not normally accept data from other projects. Not all archives are created equal. My paper assesses digital language archives in terms of the current lack of standardization regarding different aspects of archival content and access.

Given the IT revolution begun in the late 20th century and the advent of widespread digital technology at the dawn of the 21st century, it comes as no surprise that planning for digital archive development has been largely technology driven, with an emphasis in many projects on archiving materials in nonproprietary, open source digital formats for preservation and future consultation (Bird & Simons, 2002). The E-MELD Project itself is an excellent example of such a project. Many technological advances have come out of this project, particularly in the standardization of electronic formats for archiving linguistic data and metadata and in the development of an infrastructure for collaboration among electronic archives. Indeed, the strong focus on technology in modern linguistic archive development has led us to the state of the art in digital language archiving. Unfortunately, this well-placed emphasis on standardizing technology has also led us to spend relatively less time and energy on standardizing other important aspects of linguistic archives.

As we will see below, archival practices concerning content and access vary from project to project and can often be ad hoc. Such inter-archival variation can be a source of deep frustration for documentary linguists. There are many wonderful digital linguistic archives in existence today, each with unique attributes to discover and appreciate. Nevertheless, three years after digitizing my linguistic corpus of spontaneous group conversation in Spanish from naturally occurring social networks in the Paisos Catalans (Vann, 2003), due to the current state of the art in digital language archiving, I still dream of depositing my materials in the archive that wasn't.

2.0 Six ways that documentary linguists can assess electronic language archives

Documentary linguists interested in archiving and accessing speech data are wise to comparison shop for the most appropriate and cost-effective digital archive. My own personal experience looking for the best place to archive my own linguistic corpus has led me to conclude that the following six criteria, among others, may be of interest to other documentary linguists in their consideration of potential archives with which to deposit their linguistic data and metadata: (1) the actual languages archived (open, restricted to language families, restricted to indigenous or endangered languages, restricted to a particular language, etc.); (2) the type of language data archived (audio/video recordings, transcriptions, field notes, lexicons, literature, etc.); (3) the type of speech data available, if any (spontaneous conversational speech, elicited speech, interview speech, readings of word lists, etc.); (4) the levels of access available; (5) the fee for service; and (6) audience design and user-friendliness of linguistic archives. The first three criteria fall under archival content, whereas the last three fall under archival access.

2.1 Languages archived

Surprising as it may seem, it is not necessarily easy to find projects suitable for archiving language data in language X, especially if language X is not endangered. For obvious reasons priority is given in many linguistic archives to endangered languages. Some large and well-known projects, such as DOBES and E-MELD itself, archive only data related to, or proceeding from, endangered languages. Others restrict even the endangered languages they archive, limiting themselves by region. Prominent examples of regional language archives include AILLA, ASEDA, and PARADISEC, which will only accept materials representing endangered indigenous languages of Latin America, Australia, and the Pacific region, respectively. Other projects are restricted to archiving a single language and culture, such as Sinica (Chinese) and CCDSP (Portuguese). Such archives simply will not accept materials representing language X. Of course, there are open projects such as Rosetta, whose stated goal is to accumulate comparable linguistic data on 1000 different languages, and LDC, a vast archive that, to my knowledge, does not maintain any restrictions on the languages in which linguistic data may be deposited.

My own experience in selecting an appropriate archive for my linguistic corpus of spontaneous group conversation in Spanish from naturally occurring social networks in the Paisos Catalans has been frustrating due in part to restrictions on languages accepted for archiving or the lack thereof. To my knowledge, there is no online digital language archive specifically dedicated to the preservation and distribution of Spanish language data nor does the Spanish language neatly fit into any of the regional archives given that it is a major world language spoken on four continents. Though I am considering some of the open archives, such as LDC among others, I fear that my corpus could easily get lost among the many other languages that such projects archive and among the many types of language data that such projects often archive for varying purposes. The latter issue is discussed in §2.2.

2.2 Type of language data archived

The E-MELD School of Best Practices in Digital Language Documentation (http://emeld.org/school/classroom/archives/index.html) states that "A corpus is any set of linguistic materials relating to a language; it might consist of audio recordings, video recordings, transcriptions, field notes, a grammar, a lexicon, grammatical analyses, or ethnographic information. Any linguist that works in the field creates a corpus of language data." The LDC web page for data providers and corpus authors (http://www.ldc.upenn.edu/Providing/) specifies further: "Any substantial body of information rendered in a human language can serve as language data." The OLAC Working Group on Linguistic Data Types (http://www.language-archives.org/REC/type-20060406.html) divides all

linguistic resources into the following three types of language data: primary texts, lexicons, and language descriptions.

Linguistic archives vary greatly in the type of language data that they predominantly house and in the purposes for which these data are predominantly contributed by depositors.[1] Some projects emphasize metadata language descriptions (cf. e.g., E-MELD) while others emphasize written primary texts, including lexicons (cf. e.g., OTA). Many archives now emphasize primary texts of digital audio data where possible (cf., e.g., ATM). Both within and across archives, however, linguistic corpora may be deposited for different purposes, even when the types of language data archived are similar. For example, ATM content in the form of digital audio files is generally deposited for purposes of academic research in ethnomusicology, anthropology, and linguistics. ELRA, like ATM, primarily maintains audio corpora, however, much of ELRA's content is deposited for commercial purposes in language engineering, telecommunications, etc. Sorting out these differences can be a source of great frustration for documentary linguists looking to deposit linguistic resources that they have created. Linguistic resources created for the purposes of academic research might be out of place if deposited in an archive whose other resources were predominantly deposited for purposes of commercial exploitation.

The relationship between the type of language data archived and the purposes for depositing such data can be complex and easily underappreciated by field linguists whose purpose in archiving linguistic materials may simply be dialectological and/or sociolinguistic. Many of the large, well-known archives do not serve this purpose explicitly, though they may do so indirectly as their multipurpose holdings often contain various types of language data. For example, the LDC maintains corpora from authors in academia, government, and private enterprise, representing different data types and depositor purposes. These authors deposit linguistic resources in many forms: interviews, customer service interactions, lectures, radio and television broadcasts, news wires, web sites, books, magazines, newspapers, court transcripts, telephone conversations, etc. Sociolinguists with linguistic corpora to deposit may be disappointed perusing such archives looking for a good fit for both their data type and their purposes for depositing linguistic resources. To complicate matters, among those archives that emphasize digital audio recordings of speech there is still significant variation in the type of speech data archived. This issue is discussed further in §2.3.

2.3 Type of speech data archived

Most linguists would agree that preserving speech data in digital audio sound files for archival purposes is indeed a worthwhile goal; however, many might disagree as to the types of speech data to minimally include. The Rosetta project, for example, presently only includes speech data when native speakers submit recordings of themselves reading legacy texts (http://www.rosettaproject.org/about-us/archive/resource-types). The project's hope is that in the future, audio and video documentation of languages will become standard practice for linguistic researchers, and, consequently, resources will enter the Rosetta collection primarily as audio or video files. The Speech Accent Archive limits itself to speech data produced by different people reading the same paragraph in English. In contrast, the Audio Archive of Linguistic Fieldwork at Berkeley lists availability of digitized speech files by various content types labeled as stories, linguistic data, ethnographic data, and songs/chants.

Like many sociolinguists, I prize spontaneously occurring conversational speech above all other forms of speech when the goal is to describe the vernacular (Labov, 1972). The corpus that I have gathered (Vann, 1996), digitized and transcribed (Vann, 2003) reflects the rarest of coveted speech data in this domain: naturally occurring group conversations in pre-existing social networks. Ideally, I would like to deposit my corpus with an archive dedicated to collecting and preserving just this sort of speech data but, to my knowledge, no such archive exists. Perhaps to the dismay of sociolinguists, most linguistic archives do not seem to be organized by the type of speech data that they archive nor do they necessarily present the same speech data types.

---

[1] The related matter of variation in archival audience design is discussed in §2.6.

This issue is crucial in the relationship between linguists and archives. What I am looking for in an archive in terms of type of speech data collected and made available is simply not the focus of most linguistic archivists. As discussed above, descriptive linguists such as myself may not share the focus of linguistic archivists in terms of languages archived or types of corpus data archived either. In sum, there appears to be a disconnect between the content organized and presented by linguistic archives and the data that field linguists might wish to archive.

2.4 Access to archived data

In today's academic world, most linguistic researchers at US universities are bound by Institutional Review Boards that oversee any and all investigations involving human subjects. In such research projects, human subjects are often promised that their anonymity will be protected and that any speech data they provide will not be made available to the general public, but rather, only to investigators for research purposes. Thus, many field linguists who wish to archive their linguistic corpora consider the potential for restricted access to their data to be a sine qua non of an acceptable linguistic archive. Yet many linguistic archives are entirely open to the public with no graded access whatsoever. Such is the case of LACITO, Rosetta, and the Speech Accent Archive to name just a few.

Other archives offer graded access, but the different grades are not consistent across archives. Compare, for example, DOBES and AILLA. DOBES maintains four levels of access: (1) open resources, with unlimited access granted to everyone via the web; (2) restricted open resources, open for users that register online and sign a code of conduct via forms available on the web site; (3) protected resources, made available only after usage requests have been submitted via the web forms and approved by either the archivist or the responsible researcher; and (4) closed resources, where access is not available due to the explicit wish of the participants. AILLA likewise maintains four levels of graded access, but the restrictions are different: (1) free public access, no permission required; (2) protocol access by password with hint, by time limit for devolution to free public access, or by written terms and conditions that must be agreed to; (3) depositor controlled access to the resource (direct permission); and (4) creator controlled access to the resource (direct permission).

The lack of consistency across archives in the potentials for restricted access is frustrating for documentary linguists. Though DELAMAN has taken up this issue recently, with the intent to regularize access to data sources and give them enduring identifiers, to my knowledge, no organization yet tracks archives online by access restrictions and so linguists concerned about the matter must check out every potential archive one by one in their quest to find a suitable archive with which to deposit their corpora.

2.5 Fees for access to archived data

A similar issue facing linguists concerns the fee that some archives charge for the services that they provide. In essence, this issue is one of access as well. Many archives are free. Some of the higher profile linguistic archives treat their linguistic resources as commodities, charging for their distribution, though not for profit. Such is the case of LDC and ELRA. Both of these organizations distribute linguistic resources for fees that can at times range into tens of thousands of US dollars. The fees that these organizations impose can at times be cost prohibitive for linguists interested in accessing such resources. Documentary linguists interested in the widest possible dissemination of their resources might think twice about deciding to deposit their corpora with archives that charge user fees.

2.6 Audience design and user friendliness of linguistic archives

Organization, clarity, transparency, efficiency, precision, and detail are the characteristics that I admire in a good web site, no matter what the topic of the site. In a perfect world, online linguistic

archives would all openly state and describe in the introductory pages of their web sites the audience(s) for which they have been designed. In our world, the web sites of many linguistic archives do not even identify the audiences for which they have been designed.[2] This lack of audience design has been a source of great frustration for me personally in my quest to find a suitable archive for my corpus of Spanish in the Països Catalans. Do I really want somebody accessing my data for purposes other than sociolinguistic research (Bill Labov, personal communication, April 7, 2006)? To put it another way, where can I archive my data so that sociolinguists, linguistic anthropologists, ethnographers of speaking and others will find it when they want to analyze corpus data of Spanish in the Països Catalans for the purposes of social, cultural, and linguistic analysis?

Again this is an issue related to access but, rather than a matter of restricting access (§2.4), this is one of planning for easy discoverability.[3] Colleagues cannot access linguistic resources that they cannot find or do not know about. As we have seen above (§2.1), there are linguistic archives that concentrate only on certain language regions, language families and even on specific languages such that, to some degree, linguistic archives can be classified based on the language(s) that they archive. Most archives, however, are not easily classifiable based on audience design because this information is not readily apparent in most cases.

The closest that most archives come to representing audience design is a mission statement, where at least one can discern whether or not the intended audience is academic in nature. The mission statement of the Archives of Traditional Music (http://www.indiana.edu/~libarchm/mission.html), for example, states that its collections are cataloged and preserved "for use by educators, researchers, and interested members of the public, including the people from whom the material was collected." The DOBES statement (http://www.mpi.nl/DOBES/dobesprogramme/) reads (in part) as follows: "The aims of the DOBES programme are to document languages in their cultural setting and present them in such a way that they are useful for linguistics, anthropology, history, comparative literature, and other disciplines." In comparison, ELRA's homepage (http://www.elra.info/) states that "ELRA is the driving force to make available the language resources for language engineering and to evaluate language engineering technologies." Nevertheless, ELRA's audience design clearly includes both academic and nonacademic potential clients. It is the only archive I know of whose online catalog itself (ELDA) actually distinguishes its linguistic resources by audience design. Three distinct designs are given: (1) for academic use; (2) for research use by a commercial organization; and (3) for commercial use only. Such information is highly useful for documentary linguists considering different archives as potential outlets for their data.[4]

Related to the question of audience design is the matter of a linguistic archive's user friendliness. Due to a lack of audience design, many archives may appear too technical for the average user. Indeed, the average user may not understand nor have any use for the terminology employed by many computational linguists and language engineers. When such technical jargon finds its way into archive web sites and search engines, the user friendliness quotient goes down. Even E-MELD's use of terms like XML and linguistic ontology might be seen as arcane to nontechnical users. Nevertheless, there is a much bigger problem that results from poor audience design. Practices of proprietary formatting, annotation tools, and platforms can be confusing to users.[5] For example, the CHILDES project corpora are all transcribed in CHAT and CA/CHAT formats such that downloading and installing proprietary software (CLAN or TalkBank Viewer) is necessary to enable web browsing. Similarly, Max Planck projects, including DOBES, require the IMDI browser to access corpora containing

---

[2] Of course, there is no guarantee to a documentary linguist considering a resource deposit that the intended audience of web site X representing archive Y will actually coincide with the eventual end users of the site.

[3] Unlike variation in type of language data, which is clearly a matter of content that may be archived for varying depositor purposes as discussed in (§2.2), the matter of audience design is an access issue that falls squarely in the laps of archivists and the webmasters they employ to represent their archives online for resource discovery.

Discoverability is an issue that has already been discussed at length at prior E-MELD conferences. Both the OLAC and GOLD initiatives address this problem, though in my opinion, neither goes far enough from the perspective of audience design.

[4] A caveat, of course, is that archives' best intentions are in practice constrained by funding.

[5] Worse yet, competing annotation formats may not even be interoperable (Good, 2006, p. 7).

metadata, media, annotation and info files. In contrast, AILLA corpora are browsable and downloadable from common, every day internet browsers such as Netscape, Internet Explorer, etc.[6]

Given that archives cannot possibly predict the full range of users of a language documentation, unless audience design is transparently represented, it would seem that complicated tools and discipline-specific archive interfaces exist in an inverse relationship with user friendliness. As a documentary linguist looking to deposit a corpus with a linguistic archive, I prefer archives whose web sites represent audience design transparently and also strive for a positive and friendly user experience by maximizing ease of access to linguistic resources for the end user. One way to maximize ease of access is to minimize proprietary interface tools/platforms/formats and specialized software. As Trilsbeek and Wittenburg (2006, pp. 315-6) point out, researchers are often unwilling to use tools that they are unfamiliar with, and preferences for certain archive tools and user interfaces will influence the choices that depositors make in language documentations. After all, the primary focus of the depositor is on language documentation, and the culmination of the documentation process is the eventual presentation and/or publication of the data. The unfortunate state of the art of digital language archiving is that, amazingly, there is still no standard for things as seemingly simple as browsing archived collections or creating printouts of materials deposited in an archive (Trilsbeek & Wittenburg, 2006, pp. 330-331).

3.0 Conclusion

For the documentary linguist whose purpose is to create, archive, and review resources to facilitate research into the unique characteristics of a given linguistic variety, there exists an important gap in existing support for digital language preservation. Essentially, there is no centralization of digital language archives under one umbrella, no master plan relating the various archives and standardizing their practices. While online catalogs such as OLAC act as a sort of metadata clearinghouse, linking to many projects with independent agendas, and while many projects themselves link to other related projects, no particular project has been conceived with one-stop shopping in mind for documentary linguists. As a documentary linguist who works on a particular language, the most important linguistic archive I can imagine would be an online archive openly dedicated to preserving audio recordings and transcriptions of spontaneous speech for my language (among others) gathered from naturally-occurring social groups, providing for various levels of web access free of charge, and enabling linguists to easily register/reference corpus-based examples of speech in my language. I dream of the archive that wasn't.

This is the final year of the E-MELD project and its goals are largely realized. Yet, the fear that a common standard for the digitization of linguistic data may never be agreed upon and that, consequently, the resulting variation in archiving practices and language representation will seriously inhibit data access, searching, and cross-linguistic comparison continues to be rational. The primary stated goal of E-MELD is that, with guidance from descriptive linguists, consensus must be reached about certain aspects of archive infrastructure in order to offer the widest possible access to the data and provide it in a maximally useful form (http://emeld.org/documents/E-MELD.html). I propose that some organization officially encourage and oversee the standardization of archiving practices by offering overt accreditation of linguistic archives. This organization could perhaps be the LINGUIST list through its school of best practice though, alternatively, it could be some new organization representative of linguistic archives around the world that has yet to be created. In any case, I submit that only through centralized certification of archival content and access will true standardization of archiving practices fully be achieved. Certification through the LINGUIST list would be a win-win for archivists, depositors, and users alike. Archives that wished to be endorsed by the largest linguistic organization in the world and the discipline's central electronic publication would actively accommodate their practices to fall in line with approved standards. Linguists, as both depositors and

---

[6] Good (2006, p. 5) calls such tools *non-custom* and states as a goal of documentary linguistics to determine to what extent such tools are suitable for the tasks of linguistic archives. Obviously, the fewer custom tools the better.

users, would appreciate the immediate value inherent in linguistic archives accredited by the LINGUIST school of best practice. Users outside the fields of linguistics would feel secure when visiting an archive for the first time upon seeing the official LINGUIST seal. Most importantly for the future, however, no matter which organization were to offer official accreditation, certification of digital linguistic archives would at least provide a measure of consistency in terms of content and access across archives.

References

Bird, S. & Simons, G. (2002). Seven dimensions of portability for language documentation and description. In *Proceedings of the Workshop on Portability Issues in Human Language Technologies, Third International Conference on Language Resources and Evaluation* (pp. 23-30). Paris: European Language Resources Association.

Good, J. (2006). The ecology of documentary and descriptive linguistics. In *Proceedings of the 2006 E-MELD Workshop on Digital Language Documentation (Tools and Standards: The State of the Art), June 20-22, 2006, Michigan State University, East Lansing, MI* (web publication available at http://emeld.org/workshop/2006/papers/ToolEcology-2001.pdf).

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Trilsbeek, P., & Wittenburg, P. (2006). Chapter 13--Archiving challenges. In J. Gippert & N. P. Himmelmann & U. Mosel (Eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Vann, R. E. (1996). Pragmatic and cultural aspects of an emergent language variety: The construction of Catalan Spanish deictic expressions (Doctoral dissertation, The University of Texas at Austin, 1996). *Dissertation Abstracts International* (UMI No. 9633318).

Vann, R. E. (2003). Digitizing and transcribing field recordings of Catalonian Spanish. In *Proceedings of the EMELD Language Digitization Project Conference 2003: Workshop on Digitizing & Annotating Texts & Field Recordings, July 11-13, 2003, Michigan State University, East Lansing, MI* (web publication available at http://emeld.org/workshop/2003/paper-Vann.html).

Online archives and resources for linguistic documentation

AILLA: Archive of the Indigenous Languages of Latin America
(http://www.ailla.utexas.org/site/welcome.html)
ANLC: Alaska Native Language Center
(http://www.uaf.edu/anlc/)
ASEDA: Aboriginal Studies Electronic Data Archive
(http://www1.aiatsis.gov.au/aseda/specialproj/aseda/index.html)
ATM: Archives of Traditional Music
(http://www.indiana.edu/~libarchm/index.html)
AALF: Audio Archive of Linguistic Fieldwork
(http://www.mip.berkeley.edu/blc/la/)
CCSP: Comparative Corpus of Spoken Portuguese
(http://www.ime.usp.br/~tycho/)
CHILDES: CHIld Language Data Exchange System
(http://childes.psy.cmu.edu/)
CLAN: Child Language ANalysis
(http://childes.psy.cmu.edu)
DELAMAN: Digital Endangered Languages And Musics Archive Network
(http://www.delaman.org/)
DOBES: DOkumentation BEdrohter Sprachen
(http://www.mpi.nl/dobes/)
ELRA: European Language Resources Association
(http://www.elra.info/)

E-MELD: Electronic Metastructure for Endangered Languages Data
   (http://emeld.org/)
LACITO: LAngues & CIvilisations à Tradition Orale
   (http://lacito.vjf.cnrs.fr/archivage/index.html)
LDC: Linguistic Data Consortium
   (http://www.ldc.upenn.edu/)
LINGUIST: The Linguist list
   (http://www.linguistlist.org/)
LPCA: Language and Popular Culture in Africa
   (http://www2.fmg.uva.nl/lpca/)
MPI: Max Planck Institute
   (http://www.mpi.nl/world/)
OLAC: Open Language Archive Community
   (http://www.language-archives.org)
OTA: Oxford Text Archive
   (http://ota.ahds.ac.uk/ota/)
PARADISEC: Pacific And Regional Archive for DIgital Sources in Endangered Cultures
   (http://paradisec.org.au/)
Rosetta: The Rosetta project
   (http://www.rosettaproject.org/)
SAA: Speech Accent Archive
   (http://accent.gmu.edu/)
Sinica: Academia Sinica Balanced Corpus of Modern Chinese
   (http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1)
THDL: Tibetan and Himalayan Digital Library
   (http://www.thdl.org/)
UHLCS: University of Helsinki Language Corpus Server
   (http://www.ling.helsinki.fi/uhlcs/)