A paper presented at the
LSA Workshop on Sociolinguistic Archive Preparation, Portland, Oregon, 4–5 Jan 2012
Final revision for proceedings volume: 29 Nov 2013
Proceedings to appear in *Language and Linguistics Compass*

The role of metadata in the infrastructure for archival interoperation

Gary F. Simons

SIL International and
Graduate Institute of Applied Linguistics

Abstract:

Sociolinguists want to be able to share and compare datasets, and they want to do so now and far into the future. Achieving this dream will require that sociolinguistic corpora are archived in a sustainable way and that those corpora are encoded in such a way that interoperation among them is possible. This paper first sets the stage by describing broadly the requirements for sustainable archiving. It then focuses on the role of metadata in constructing an infrastructure that will support the kind of interoperation that is envisioned—both corpus-level metadata for facilitating the discovery of relevant corpora and observation-level metadata for facilitating the comparison of observations that are comparable.

1. Introduction

The fundamental problem addressed in the Workshop on Sociolinguistic Archive Preparation was the problem of sharing. Sociolinguists are asking each other: "How do we archive our corpora so that they can be shared?" We need to be able to compare current findings with previous findings to describe change over time. We need to be able to compare contemporaneous findings from multiple speech communities to describe synchronic variation. We also need to be able to study other scholars' data in order to confirm the conclusions of their analyses.

And we need to be doing this sharing with sustainability. We want to perform the above mentioned operations with corpora not just at the time they are made available; we also need to keep doing those things far into the future. But given the relentless entropy that degrades digitally stored information, the relentless pace of innovation that makes hardware and software obsolete even before they stop functioning, and the relentless development of sociolinguistic practice that keeps coming up with new ways of approaching our discipline, how do we keep our corpora from falling into disuse, and ultimately slipping into oblivion? (Bird and Simons 2003b, Simons 2006)

The goal of the workshop organizers was to launch a process whereby the sociolinguistics community could develop protocols and engineering standards for data sharing. In the call for participation, they envisioned a day in which a whole community of practitioners would be archiving their work in such a way that all "the resulting corpora could be subsumed under a uniform archival 'umbrella' permitting the resulting studies to be compared." This paper first sets the stage by synthesizing findings in the archiving field to describe broadly the characteristics of such an archival umbrella. It then focuses on the role of metadata in constructing an infrastructure that will support the kind of interoperation that is envisioned.

2. Foundations of sustainable sharing

There are five necessary conditions for the sustainable sharing of any corpus (Simons and Bird 2008). In order for a corpus to be shared today, it must be *discoverable, accessible, interpretable,* and *portable*. And for this to continue far into the future, it must also be *preserved*. The next five paragraphs expand on these five conditions.

A corpus must be *discoverable;* that is, it cannot be used unless the prospective user is able to find it, both discovering that it exists and learning where it is located. The key to making this possible is descriptive metadata. The description of the corpus must be published in such a way that the user to whom it would be relevant is able to discover that the corpus exists when searching for potentially relevant data. The description of the corpus should also contain enough information so that the user to whom it is relevant is able to judge it as being relevant without having to first obtain a copy.

A corpus must be *accessible;* that is, it cannot be used unless the prospective user is able to access it. Accessibility has two major facets. First, the potential user must have the right to access and use the corpus. To facilitate this, it is essential that the corpus creator sorts out the rights when the data is being collected and then states them clearly when it is archived (Dwyer 2006). The Open Access[1] strategy fosters the most widespread long-term use. Creative Commons[2] licenses provide an off-the-shelf legal framework for implementing this strategy. Second, the potential user must know the procedure for gaining access. Direct access via a URL is the mechanism that offers the most widespread use; a persistent URL will ensure that use is long-term.

[1] http://www.eprints.org/openaccess/

[2] http://creativecommons.org/

A corpus must be *interpretable;* that is, it cannot be used if the user is not able to make sense of the content. The basic standard in digital preservation is ISO 14721 — the Open Archival Information System (OAIS) reference model. This is an international standard that defines how an institution must act in order to function as a trustworthy archive. Among the mandatory responsibilities of a conforming archive is that it must: "Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information" (CCSDS 2012, page 3-1). This means that to be fully usable a sociolinguistic corpus must document the context of the study, the methodology that was followed, the terminology that is used, and all the details of the data encoding (including the character set, the file format, the definitions of all the data categories, and the conventions for expressing their values).

A corpus must be *portable;* that is, it cannot be used if it does not interoperate in the user's working environment. A corpus must work with the user's hardware and operating system, with the software tools that are available to the user, and with the best practices of the targeted user community. Maximizing portability means using formats that are open and transparent and widely supported by many software suppliers. By contrast, explicitly implementing conversions for proprietary or home-brew formats is not sustainable in the long term as platforms keep changing. Maximizing portability also entails following best practice markup and terminology. See Bird and Simons (2003b) for a full discussion of the dimensions of portability for digital language documentation and description.

Finally, a corpus must be *preserved;* that is, use of a corpus cannot be sustained if a faithful copy of the original resource ceases to exist. The archiving institution must follow procedures to ensure that resources are preserved against all reasonable contingencies (such as by keeping up-to-date offsite backups), to ensure periodic migration to fresh and current media, to ensure that all copies are authenticated as matching the original, and to maintain preservation metadata (such as for provenance and file fixity). Care to preserve the original resource must also be taken when correcting or otherwise improving a corpus, since subsequent users may want to ignore the improvements so as to have identical data for comparing their analysis with prior competing analyses. See Chang (2010) for a full discussion of practices involved in the preservation of digital language resources.

Individual sociolinguists can create corpora that are portable and interpretable, but they cannot by themselves preserve those corpora long into the future or provide long-term access to future users. That is the role archives play for the corpora entrusted to their keeping. Neither can individual sociolinguists by themselves make their corpora discoverable by any potential user; nor can archives do this by themselves. For this they depend on aggregating services.

Aggregators do not curate a collection themselves. Instead, they harvest information curated by many others and offer the convenience of a single point of entry for accessing those many sources. A general internet search engine like Google is a prime example. If archives expose the contents of their collections on web pages, then it becomes possible for potential users to discover relevant material by using Google search. The language-resource-specific search service developed by the Open Language Archives Community[3] is another example. If archives expose their catalogs in OLAC's standard format (see below), then potential users can discover relevant material through precise search on facets like language, language family, country, linguistic data type, medium, format, and more.

Thus the total infrastructure for corpus archiving involves four key players: creators, archives, aggregators, and users. Creators create the language resources and then deposit them into archives. The archive is an institution that curates language resources for long-term preservation and makes at least the metadata for those resources available to aggregating services. The aggregator is an institution that gathers resources (or at least metadata about resources) from many archives in order to provide interoperation across resources in all the archives. Finally, the user, who wants to find relevant language resources, needs to search in only one place (namely, at the aggregation service) in order to find resources from a host of archives and creators, including archives and creators that were previously unknown to the user.

3. Foundational terminology

There are three terms—archive, metadata, and interoperation—that are foundational for the issues discussed in this paper. These are also terms that may be misunderstood due to variation in usage. In this section, I define these terms before proceeding.

The term *archive* is polysemous in common usage. For instance, *Wikipedia* (on January 4, 2012) identified two primary senses in its definition: "An archive is a collection of

[3] http://search.language-archives.org/

historical records, or the physical place they are located." In the title of the workshop, "Workshop on sociolinguistic archive preparation," the first sense is in focus; but the new emphasis on archiving in the linguistics community, puts the focus on the second. The usage in the workshop title thus presents a problem in that if we call a collection of information an archive, sociolinguists will think they have "archived" when they have created an "archive." If, as a result, they fail to deposit their work with an archiving institution, their work cannot be shared sustainably and will eventually slip into oblivion. We can avoid this misunderstanding with a more careful use of terminology. The outcome we are looking for is that sociolinguists will create *archivable* corpora and that they have *archived* when these have been deposited in an *archive*.

Another term with multiple senses is *metadata*. Literally, it means "data about data." Just as we have data at many levels, so also with metadata. When librarians and archivists talk about metadata, they mean data about the items they curate. When sociolinguists use the term, they often mean data about the individual observations they are making, but for the archivist, that is data rather than metadata. To avoid confusion, I will speak of corpus-level metadata versus observation-level metadata.

A final foundational term is *interoperation*. Two or more systems interoperate when they can exchange information or services and then make satisfactory use of what is exchanged. Two levels of interoperation (corresponding to corpus-level and observation-level) are distinguished. Macrointeroperation operates at the level of the corpus and has to do with helping users to discover corpora that are relevant to their research interests. By contrast, microinteroperation goes inside of each corpus and has to do with helping users to compare points of content between relevant corpora.

4. Corpus-level metadata and macrointeroperation

There is no need for the sociolinguistics community to develop an infrastructure for macrointeroperation since a suitable one is already functioning. The Open Language Archives Community[4] (OLAC) is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources. They have done so by developing consensus on best practices for sharing descriptive metadata of archived language resources, and then implementing a network of interoperating repositories and services for accessing those metadata and language resources. Since its founding in 2000, the OLAC virtual library has grown to include over 190,000 language resources housed in 46 participating archives.[5] For example, some of the major participating archives are The Language Archive (Max Planck Institute for Psycholinguistics, The Netherlands), PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures, Australia), Linguistic Data Consortium (Philadelphia), Collections de Corpus Oraux Numeriques (Paris), and Endangered Languages Archive (SOAS, London).

The community has defined standards for the encoding and exchange of corpus-level metadata to permit discovery and sharing of language resources. There are three foundational standards: *OLAC Process*[6] defines the governance and standards process; *OLAC Metadata*[7] defines the XML format used for the exchange of metadata records; and *OLAC Repositories*[8]

[4] http://www.language-archives.org
[5] http://www.language-archives.org/archives
[6] http://www.language-archives.org/OLAC/process.html
[7] http://www.language-archives.org/OLAC/metadata.html
[8] http://www.language-archives.org/OLAC/repositories.html

defines the requirements for implementing a metadata repository that can be harvested by an aggregator using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting.[9]

A usage note adopted through the OLAC process, *OLAC Metadata Usage Guidelines,*[10] explains the available metadata elements and how to use them. The OLAC metadata scheme is based on Dublin Core (Bird and Simons 2004). This is a standard originally developed within the library community to address the cataloging of web resources. Dublin Core has fifteen basic metadata elements: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type.[11] OLAC uses an enriched variant of Dublin Core known as Qualified Dublin Core which supports the incorporation of application-specific vocabularies for describing resources. The OLAC community has used its process to define five metadata extensions (Bird and Simons 2003a) that are tailored to language resources:

- Subject language: for identifying precisely (with a code from the ISO 639 standard[12]) which language(s) a resource is "about";
- Linguistic type: for classifying the structure of a resource as primary text, lexicon, or language description;
- Linguistic field: for specifying a relevant subfield of linguistics;
- Discourse type: for indicating the linguistic genre of the material; and
- Role: for documenting the parts played by specific individuals and institutions in creating a resource.

[9] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[10] http://www.language-archives.org/NOTE/usage.html
[11] http://dublincore.org/documents/dces/
[12] http://www.sil.org/iso639-3/

An example of a sociolinguistic corpus that is discoverable through OLAC is *SLX Corpus of Classic Sociolinguistic Interviews* (Strassel et al. 2003). The following listing shows the metadata record as published by the Linguistic Data Consortium in the XML format prescribed by the *OLAC Metadata* standard:

```
<olac:olac>
   <dc:title>SLX Corpus of Classic Sociolinguistic
      Interviews</dc:title>
   <dc:creator xsi:type="olac:role" olac:code="author">Stephanie
      Strassel, Jeffrey Conn, Suzanne Evans Wagner, Christopher
      Cieri, William Labov, Kazuaki Maeda</dc:creator>
   <dc:date xsi:type="dcterms:W3CDTF">2003-11-25</dc:date>
   <dc:description>http://www.ldc.upenn.edu/Catalog/docs/
      LDC2003T15</dc:description>
   <dc:description>Application: sociolinguistics</dc:description>
   <dc:description>Data source: field recordings</dc:description>
   <dc:format>Sample rate: 22050Hz; Sample type: pcm</dc:format>
   <dcterms:extent>Corpus size: 1572864.000 KB</dcterms:extent>
   <dcterms:medium>Distribution: 1 DVD</dcterms:medium>
   <dc:identifier>LDC2003T15</dc:identifier>
   <dc:identifier>ISBN: 1-58563-273-2</dc:identifier>
   <dc:rights>Non-member license:
      http://www.ldc.upenn.edu/Catalog/nonmem_agree/generic.license.h
      tml</dc:rights>
   <dc:language xsi:type="olac:language" olac:code="eng"/>
   <dc:subject xsi:type="olac:language" olac:code="eng"/>
   <dc:type xsi:type="olac:linguistic-type"
      olac:code="primary_text"/>
   <dc:type xsi:type="dcterms:DCMIType">Sound</dc:type>
</olac:olac>
```

Note the use of domain-specific code values on the Language, Subject, and Type elements at the end of the record. These make it possible to support precise search for specific languages and resource types across the combined collection of all 46 participating archives with the use of OLAC's search service.[13]

The above metadata record in XML format is intended to give the reader a sense of how the macrointeroperation works behind the scenes. It is not an example of what the corpus creator must create. Rather, the responsibility falls to the archive to collect the metadata from the creator by whatever means it wishes, to save and maintain the metadata by whatever means it wishes (such as in a database), and then to output the metadata in the standard XML format for sharing with the OLAC community. For sociolinguists who have not yet deposited their corpora into an archive, there is a MetaMaker[14] service which allows a user to describe a corpus by filling in a web form. Upon pressing the Submit button, the metadata is assembled into the correct XML format and submitted to OLAC.

The OLAC infrastructure can be used as is to foster discovery and access of sociolinguistic corpora. However, the sociolinguistics community could achieve even greater precision in searching by coming to agreement on additional metadata conventions specifically for sociolinguistic corpora. The simplest step would be to agree on the use of a

[13] http://search.language-archives.org
[14] http://talkbank.org/resources/metamaker/

standardized label within the Type element in order to make it possible to retrieve all known sociolinguistic corpora. For instance,

```
<dc:type>Sociolinguistic corpus</dc:type>
```

Similarly, labels for widely used data formats could be standardized and included in metadata records. For instance, a corpus that uses the CHAT transcription format (MacWhinney 2000) could include the following metadata element:

```
<dc:format>CHAT transcription format</dc:format>
```

Even more ambitious would be to use the extension mechanism[15] defined in the *OLAC Metadata* standard to create controlled vocabularies for resource types and data formats that are of special interest to the sociolinguistics community.

5. Observation-level metadata and microinteroperation

Whereas a usable infrastructure is already in place for corpus-level metadata and for macrointeroperation that supports corpus discovery across dozens of archives, the same cannot be said for observation-level metadata. It is not yet possible to build services that would find comparable observations across dozens of corpora or would perform statistical analysis across dozens of corpora. To achieve such microinteroperation across sociolinguistic corpora, the sociolinguistics community will need to develop standardized ways of encoding the observation-level metadata. Standards would be needed at three levels: standard names and definitions for a wide variety of demographic and situational factors, standardized ways

[15] http://www.language-archives.org/OLAC/metadata.html#Defining%20a%20third-party%20extension

of expressing the possible values for those factors, and standardized formats for encoding the association of factors with values.

A common thread that runs through all kinds of sociolinguistic data collection—whether one is recording speech events, observing language choice in social context, or measuring language attitudes—is to record characteristics of the people involved. This makes it possible in the analysis to discover correlations between these characteristics and features of observed language behavior. Commonly referred to as demographic factors, such characteristics include things like gender, year of birth, age cohort, educational level, social class, ethnicity, religious affiliation, level of religious activity, immigrant generation, age at arrival, first language, heritage language, language proficiency, and even relationship to interviewer (see other papers in this issue for examples of these and other demographic factors). In addition to characteristics of the participants, the researcher may also encode characteristics of the situational context (such as setting, topic, or purpose) or of the communication itself (such as tone, modality, register, or genre).

A minimal step in the direction of supporting interoperation at the observation level would be to include a listing of the factors recorded in observational metadata as part of the corpus-level metadata. In this way, researchers searching for observational data of a particular type would be able to discover corpora to look in. Using the Dublin Core metadata scheme, this would be a kind of Description of a corpus, and using the OLAC extension mechanism, it would be possible to define a refinement of Description with a label like “sociolx:factor”. One could then encode a corpus-level metadata element like the following to indicate that one of the demographic factors encoded in the corpus is the age of arrival for participants who are immigrants:

```
<dc:description xsi:type="sociolx:factor">Age at
      arrival</dc:description>
```

In the OLAC metadata schema, the content of the XML element is freeform, so that each researcher could use any phrase to describe any factor.

The first step toward standardization would be for the sociolinguistics community to develop a list of factors and achieve a consensus regarding their names and definitions. In this way if two different researchers were to say that they had encoded a factor with the same standradized name, the community of potential users could be assured that they had encoded comparable things. Such a list of standardized factor names and definitions would then serve as a controlled vocabulary for metadata encoding, with each name having a corresponding encoding token for use in metadata. For instance, if the factor named "Age at arrival" were encoded as "age_at_arrival", then the following would be the standardized way of signaling in an OLAC corpus-level description that the observations use this factor in their metadata:

```
<dc:description xsi:type="sociolx:factor"
      olac:code="age_at_arrival"/>
```

If all archived corpora used the standardized list of sociolinguistic factor names in this way in their corpus-level metadata, it would be possible for a researcher who is looking for datasets that encode age at arrival for immigrants to actually find all such corpora.

The next step in achieving microinteroperation would be to develop standard ways of expressing the possible values of the standardized factors. This step is needed to ensure that observations recorded using the same factor are in fact directly comparable. For instance, if one researcher encodes gender as "M" or "F" and another as "male" or "female" and yet another as 0 or 1, then their datasets cannot yet interoperate, even though they contain the very same information at a conceptual level. Statistical analysis across all the datasets would require that all the data first be recoded to a consistent set of values for the factors.

Achieving the dream of interoperation across sociolinguistic corpora therefore requires that the sociolinguistics community not only agree on standardized names for factors, but also on standardized ways of representing their values. Note, however, that this particular standardization effort is not a matter for the average sociolinguistic practitioner to engage in. Rather, it is a matter for the technologists who support the sociolinguistics community to work out. The semantics of what needs to be represented is a problem that belongs to sociolinguists, but standardizing the format of representation is a problem that belongs to those who will be implementing the software systems.

A standardized means for representing a data value is called an encoding scheme.[16] For factors that have an open-ended number of possible values, the encoding scheme must specify the rules for constructing a valid value. *XML Schema Datatypes*[17] is a widely used standard that provides names and rules for encoding a large number of primitive data types like boolean, integer, decimal, string, and date. An example of an encoding scheme for a complex type is *DCMI Point*[18] which specifies how to express the location of a point on earth in terms of its longitude, latitude, and elevation. For factors that have a closed set of possible values, an encoding scheme enumerates the list of possible values along with a definition for each. Such an encoding scheme is typically called a controlled vocabulary. One means of standardizing a controlled vocabulary is represented by the *Recommended Metadata*

[16] http://wiki.dublincore.org/index.php/Glossary/Encoding_Scheme
[17] http://www.w3.org/TR/xmlschema-2/
[18] http://dublincore.org/documents/dcmi-point/

Extensions[19] of OLAC in which each vocabulary is defined both with a human-readable HTML document and a formally encoded XML schema[20] that supports machine validation. Another means of defining standardized vocabularies is followed by the Dublin Core Metadata Initiative. They formally encode their vocabularies and documentation in the machine-readable Resource Description Framework (RDF) schema language;[21] this is in turn transformed to HTML for human readers.

To go all the way to automated microinteroperation, one more step is required. It is necessary that the complete set of observations comprising a corpus (including the association of the factors with their values) be encoded in a standardized format. Again, devising such standards for sociolinguistic corpora is a problem for the technologists to tackle on behalf of the whole community. For observations that can be expressed in spreadsheet-like tables, the most ubiquitous format for archiving and data interchange is tab-delimited (TAB) or comma-separated value (CSV) format. In this format, each line of the file contains all the data and metadata values for a single observation (separated by tabs or commas), with factor names as the column headings in the first line of the file. This format can be loaded into virtually any spreadsheet, database, or statistics program. For datasets that are more complex, involving hierarchical structures or network structures, more complex formats are needed. The *TEI Guidelines* (Burnard and Bauman 2013) provide an example of a widely used standard for the encoding of linguistic corpora with XML markup. The Linked Data[22] approach, based on RDF, is also beginning to find traction for linguistic data

[19] http://www.language-archives.org/REC/olac-extensions.html
[20] http://www.w3.org/TR/xmlschema-0/
[21] http://dublincore.org/schemas/rdfs/
[22] http://www.w3.org/standards/semanticweb/data

(Chiarcos et al. 2012). When a dataset is encoded in a known format and it is declared in a standard way in the corpus-level metadata, then an automated process can find the corpus and load the data into a common database or transform the data into the format needed by an analysis program. After multiple datasets have been processed in this way, automated search and analysis across a whole set of corpora is then possible.

6. Conclusions

Sociolinguists can share their corpora long into the future if they deposit them with archival institutions that will preserve them, make them accessible to potential users, and make them globally discoverable through good corpus-level metadata that is fed into an aggregation infrastructure like OLAC. In order for potential users to actually use these corpora, the observation-level data and metadata need to be expressed in encoding schemes that are well-documented so that the values are readily interpretable and in file formats that are portable across a variety of hardware and software platforms. Once the sociolinguistics community can develop a consensus concerning standardized labels for relevant factors along with associated encoding schemes for their values, then it will be possible to achieve the dream of automated search and comparison across the wealth of known sociolinguistic corpora.

References

Bird, Steven and Gary Simons 2003a. Extending Dublin Core metadata to support the description and discovery of language resources *Computers and the Humanities* 37(4):375-388. <http://arxiv.org/abs/cs.CL/0308022>

Bird, Steven and Gary Simons. 2003b. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79:557–582. <http://www.language-archives.org/documents/portability.pdf>

Bird, Steven and Gary Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In D. I. Hillmann and E. L. Westbrooks, eds., *Metadata in Practice*, pp. 203–222. Chicago: American Library Association. <http://www.ldc.upenn.edu/sb/home/papers/mip.pdf>

Burnard, Lou and Syd Bauman (eds.). 2013. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, version 2.5.0. Charlottesville, Virginia: Text Encoding Initiative Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

CCSDS. 2002. *Reference Model for an Open Archival Information System (OAIS).* Recommended Practice CCSDS 650.0-M-2. Consultative Committee for Space Data Systems. <http://public.ccsds.org/publications/archive/650x0m2.pdf>

Chang, Debbie. 2010. *TAPS: Checklist for responsible archiving of digital language resources.* MA thesis, Graduate Institute of Applied Linguistics. Dallas, TX. <http://www.gial.edu/images/theses/Chang_Debbie-thesis.pdf>

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann (eds.). 2012. *Linked data in linguistics: Representing and connecting language data and language metadata.* Berlin: Springer.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Gippert, Jost, Nikolaus Himmelmann and Ulrike Mosel, eds. *Fundamentals of Language Documentation: A Handbook.* Berlin: Mouton de Gruyter, pp. 31–66.

MacWhinney, Brian. 2000. *The CHILDES Project (3rd ed.). Volume I: Tools for Analyzing Talk: Transcription Format and Programs*. Mahwah, NJ: Lawrence Erlbaum Associates. <http://childes.psy.cmu.edu/manuals/chat.pdf>

Simons, Gary. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. *SIL Electronic Working Papers 2006-003*. < http://www-01.sil.org/silewp/2006/003/SILEWP2006-003.htm>

Simons, Gary and Steven Bird 2008. Toward a global infrastructure for the sustainability of language resources. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, 20–22 November 2008, Cebu City, Philippines,* pp. 87–100. <http://www.aclweb.org/anthology-new/Y/Y08/Y08-1008.pdf>

Strassel, Stephanie, Jeffrey Conn, Suzanne Evans Wagner, Christopher Cieri, William Labov, and Kazuaki Maeda. 2003. *SLX Corpus of Classic Sociolinguistic Interviews.* Philadelphia: Linguistic Data Consortium. < http://catalog.ldc.upenn.edu/LDC2003T15>