

A Distributed Architecture for the Documentation of Language and Culture

Clifton Pye
Linguistics
The University of Kansas
pyersqr@ku.edu

2011 will mark the centenary of Franz Boas' Introduction to the Handbook of American Indian Languages (1911). Boas' essay addressed fundamental issues concerning the variation to be found in human biology, language and culture. Boas concluded that even though the indigenous languages of North America are structurally different from European languages they are not primitive in the sense that these languages have all the expressive powers of their European counterparts. Boas' essay appeared at the height of the eugenics movement in the United States and anticipated claims of racial superiority in Nazi Germany. Boas celebrates the variety of the human experience that exists within a common human inheritance.

While advances in genetics since Boas' time have revealed how a common genetic code can result in millions of phenotypes we know far less about the variation to be found in the human potential for language and culture. The flat earth that Thomas Friedman describes is rapidly leveling linguistic and cultural variation. Satellite broadcasts now make it possible for people in remote villages around the world to watch professional wrestling matches broadcast in English. The new millennium will see an accelerating loss of the human potential for language and culture. Ninety-six percent of the world's languages are spoken by 4% of the population. Half of the world's 6,000 languages are already moribund. There are no longer any children learning to speak these languages and they will not transmit this knowledge to a new generation. Only 600 languages may remain by the end of this century.

Only crude surveys exist mapping some of the variation to be found in the richest linguistic areas of the world. Our generation may be the last on earth to witness the variation to be found among natural and human populations. When we have yet to document the full range of linguistic expressions for English the hope of producing a similar documentation for languages with one or five or twenty or a hundred speakers seems remote. Any one of the world's moribund languages could have a linguistic structure that would revolutionize our understanding of the human capacity for language. Linguists have documented a few of these exotic structures already. The late Ken Hale documented the non-configurational structure of Walbiri, an aboriginal language of Australia. The late Dale Kinkaide documented the absence of a noun/verb contrast in Upper Chehalis, a Salish language of the Northwest Coast.

Language preserves the linguistic and cultural heritage of thousands of years of human experience. A comprehensive language survey documents the encyclopedic knowledge that resides in the lexicon of each society as well as its grammatical combinations. The Comparative Method enables linguists to amplify the record of each language by using the grammars of related languages to trace their unwritten history of linguistic and cultural innovation and borrowing. The Comparative Method is only as good as the available language data allows so the loss of each language severely limits our knowledge of whole families of languages as well as their historical contacts with other language families. Entire linguistic eco-systems such as Algonquian are vanishing before our ears.

There is a critical need to document the linguistic and cultural variation that presently exists before it disappears forever. This goal requires the development of new tools and processes to record, preserve and share as much of the human intellectual genome as possible. The era of individual investigators safeguarding their observations in a personal notebook is over. Advances in networked communications now enable investigators around the world to construct a distributed archive of linguistic and cultural data along the lines of Wikipedia. A distributed archive would allow linguistic experts to share ideas and procedures for documenting linguistic features. The individual researcher would have access to planetary resources for documentation that would provide models of the best practices in the field and which would be updated constantly. A distributed archive would allow native speakers to make their own contributions and search for new ways to preserve their language and culture.

A distributed archive for language and culture requires a major investment in linguistic and cultural infrastructure to overcome the piecemeal approach that has long characterized research in the social sciences. A language archive would enable researchers to scan the database for a variety of linguistic structures, e.g. examples of syllable types or applicative constructions. The archive would also have to document any known constraints on grammatical structure. The archive would include audio and video recordings with links to transcriptions and metadata descriptions that would allow searches at all levels of linguistic structure. This shared archive would flag areas in urgent need of documentation so that researchers could target critical features of languages and dialects rather than repeating previous research efforts.

A national archive would exhibit the best practices for documenting language and culture. In this sense, it would provide a teaching tool that would communicate not only what needed to be done, but how to do it. It would provide instructors with a wide range of examples illustrating how sounds, words or stories are realized within the full range of the human experience. Linguists and language communities have already put together a variety of language and cultural archives that are known to researchers within these individual communities. Some archives encourage contributors to add to their database, but their individual nature discourages communications across archives and slows the spread of best practices between communities. The current internet website for the Linguistic Society of America is informative in this regard. It focuses on the business of the society, but does not provide any systematic information on the languages of America. The LSA website provides a few links to other internet resources for language description, but these are hard to find on the website.

There are a few initiatives around the world that provide an idea of what a national archive would contain as well as technological limitations to avoid. The World Atlas of Language Structures (wals.info) is a joint project of the The Max Planck Institute for Evolutionary Anthropology and the Max Planck Digital Library with a book version published by the Oxford University Press. WALS provides a database of structural properties of languages combed from published materials. WALS contains 141 maps with descriptions of such features as vowel inventory size, noun-genitive order, passive constructions, and "hand"/"arm" polysemy. WALS shows how languages can be taken apart and compared feature by feature. WALS is limited by its top-down architecture. Its webpage only displays information put together by a select group of authors and does not allow speakers of individual languages to add new information or correct old information. The linguistic features described in WALS provide an initial template for language documentation that could be followed for every language.

The Electronic Metastructure for Endangered Languages Data (E-MELD) webpage provides a forum for discussing the best practices for language documentation. It tackles the critical topic of identifying how current technology can be used to record language data. This is a critical issue for any researcher who realizes that a recording device purchased three years earlier is no longer supported by the manufacturer. Anyone who wants to record their community should know that many consumer electronic recorders do not record a high quality audio signal. E-MELD has changed its project goal to supporting the Open Language Archives Community. It provides an Open Repository Editor which enables any researcher to submit information about the language data that they have collected to the Linguist database without making the actual dataset available. The Linguist database contains information on 7555 languages including the 7270 languages in the Ethnologue database. Most of this information simply lists the languages and number of speakers. E-MELD also contains a link to the Online Database of Interlinear Glossed Text (ODIN) which lists texts and publications for 1274 languages. This information is extremely useful, but lacks an organized template which would point to the linguistic features displayed in the texts.

The Hans Rausing Endangered Languages Project (www.hrelp.org) provides support for individual investigators interested in documenting endangered languages. It is linked to a program documenting endangered languages run by the School of African and Oriental Studies at the University of London. The program offers graduate training and funding to support work on language documentation. The website has a link to an Endangered Language Archive with material on 70 languages. The archive includes recordings of discourse, dialect surveys, stories and word lists. It does not provide a systematic organization for specific features contained in this data.

The internet encyclopedia Wikipedia (www.wikipedia.org) also has a significant amount of information about languages and language families including much information from Ethnologue. The information for individual languages varies tremendously and focuses on the number of speakers and regional variants. Information on the grammars of the languages is mixed. Many entries only provide examples of noun or pronoun forms. Wikipedia does not provide much information on the grammars of endangered languages. The key advantage that Wikipedia has over other archives is its open architecture which encourages a collaborative approach to language documentation. While the linguistic information that Wikipedia provides is limited, it allows indigenous speakers of these languages to edit the information about their own language in their own language.

Each of these resources provides an aspect of what a comprehensive inventory of the world's languages should include. There is an urgent need for technical information about recording and archiving data on language and culture. There is an equally urgent need for technical information about the various grammatical components that need to be documented. Above all else, there is a critical need to record the many variants of each language while they can still be found in their original contexts. This work is beyond the capacity of the world's social sciences and this is precisely why linguists need to create a distributed technology which would allow them to show how anyone in the world can document their own language. It is even possible to create an iPhone app for language documentation.

Organization will be a critical part of this endeavor. It is the linguists' responsibility to create a template for language documentation which will alert speakers to those aspects of their

language and culture in need of documentation. The Comparative Mayan Grammar wiki (pyersqr.org/Maya/doku.php) provides an example of what such a template might look like. It lists a number of grammatical features in need of description with examples to provide a model for new contributions. While this website currently lacks many features, it provides an example of an open approach to linguistic research. Its open architecture allows anyone to implement a new feature such as the addition of an audio or video archive.

To this point I have provided examples from the discipline I know best, but an open archive has obvious applications across the social sciences. Its chief advantage is that it allows social scientists to document a wider array of human behavior while giving subjects a role in documenting their own lives. To be sure, there are enormous institutional obstacles to overcome in constructing such archives. Not the least of these is the traditional culture of the social sciences built around the individual investigator. It is time for the social sciences to learn from the hard sciences how to implement collaborative research on a national or international scale.

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.