Report on the National Science Foundation's
Workshop on Documenting Endangered Languages
Durham, New Hampshire, October 2007

Summary:

On 15th and 16th October 2007, 32 funders and researchers gathered to discuss the current state and future prospects in the documentation of endangered languages. Over the past decade, funding for work on endangered languages has increased significantly, with major initiatives coming from the VolkswagenStiftung (VWS) in Germany, the Hans Rausing Endangered Language Project (HRELP) in the United Kingdom, and a joint effort by the National Science Foundation (NSF) and the National Endowment for the Humanities (NEH) in the U.S. The two-day meeting was a chance for representatives from those organizations, and from other efforts, to take stock of this work and prioritize current and future needs. Representatives from universities and from indigenous language programs gave reports on the results of the current funding. Expectations for the future, both in terms of what initiatives might be forthcoming and what would be most needed, were also discussed.

Extensions of several programs were reported as accomplished or likely to be forthcoming. The NSF/NEH program is now permanent. The VWS program is in midterm evaluation; news on the specifications of a program's final phase will emerge in early 2008. The European Science Foundation is likely to launch BABEL—Better Analyses Based on Endangered Languages. New initiatives are being launched in Japan. Cultural Survival, a non-profit based in Cambridge, MA, is fostering the creation of a new, Native-run organization devoted entirely to language revitalization. All of these together, however, represent a relatively minor increase in funding, and do not seem to address the need for a major effort on the part of language communities and linguists to take advantage of the last living speakers of endangered languages.

One notable gap in funding is for revitalization efforts. Many native communities are trying to bring their languages back from the brink (or even to revive those that have already "gone to sleep"). All three major funding efforts are mandated to deal specifically with scientific issues. Although the results of those research efforts often contribute to language revitalization programs, the communities themselves have very few options for such funding. There is some funding available from the Administration for Native Americans (ANA), but this is an area where non-governmental organizations could play a bigger role.

There was great enthusiasm for developing statistics on the current level of documentation for every endangered language, in order to direct attention where it is most urgently needed. Further training opportunities would be welcome, especially given

that both senior researchers and students often need instruction in language documentation techniques. More digital archives are needed, both because some are geographically limited and because the task at hand is too massive for any one group to accomplish. Such efforts should be better integrated with existing and emerging efforts in related fields. Ensuring that data will migrate and be upgraded to new formats is essential.

The General Assembly of the United Nations has declared 2008 the Year of Languages. Language use was also recognized in the recently passed Declaration of Rights of Indigenous Peoples (September 2007). The sense of the workshop was both that a great deal of useful effort was being supported and that there is much more to be done. The consensus was that the efforts that had come to light at the meeting should be part of a larger, more inclusive and far-reaching process, to which the participants were eager to lend support.

Action Items:

- Establish a database on level of documentation.
- Foster connections between academic linguists and language practitioners.
- Generate greater awareness, especially via UN's 2008 "Year of Languages."
- Bridge interdisciplinary gaps, especially with computational colleagues.
- Train and better involve native-community linguists.
- Develop more robust and accessible tools.
- Extend the scope of existing digital archives and supplement them with new ones.
- Salvage existing data in danger of disappearing.
- Further develop consistent and harvestable metadata.
- Create ways of providing academic ("publishing") credit for online resources.
- Ensure that citation formats for online sources are recognized by citation indices.
- Implement archiving requirements on US and Canadian grants.
- Insist on more realistic archiving plans in grant proposals.
- Provide better training in documentation and archiving.
- Build coalitions with US tribes and Canadian First Nations.
- Treat this global issue globally.

Past Accomplishments:

The VWS initiative has funded 36 projects from 2002 to the present (along with planning workshops in 1999 and 2000 and several pilot projects in 2001) with worldwide coverage. A total of €15.3 million ($19.6 million) has been spent to date. These have resulted in an extensive archive, hosted at the Max Planck Institute in Nijmegen, The Netherlands. One of the goals of the initiative was to change the way that language documentation was practiced, and the corpus web site implements some of the sought-for changes. A consistent metadata was devised, along with a program to manipulate it. At present, much of the primary material remains closed to outside view; this is expected to change over the coming years as the researchers finish processing and publication.

Although the initiative is currently projected to have its last call this year (2007), there is some hope that it will be extended.

The HRELP program has awarded 108 grants totaling over £4.3 million ($8 million) since 2003. As with the VWS projects, the languages studied have been from all parts of the globe. A digital archive is under construction and will ultimately make various corpora available for study and use by native communities as well as by scientists and scholars. The initial endowment will last another ten years or so, depending on a number of factors. There is some hope that the University of London, where HRELP is housed, will assist in raising funds to extend the program, but the prospects are not currently known.

The Linguistics program at NSF has supported work on endangered languages for many years, extending back to the stewardship of Paul Chapin in the 1980s and 1990s. In 2004, a Documenting Endangered Languages (DEL) program was launched as a separate initiative, in collaboration with the National Endowment for the Humanities (NEH) and the Smithsonian Institution. NEH had also made many previous awards in the endangered language area, especially through its Preservation and Access program. The new DEL program received additional funds from both agencies, more than doubling what had been available before. The program awarded grants totaling $4.3, $4.4 and $5.1 million in the past three years (2005-2007). At the October 2007 meeting, it was announced that the DEL program had been made permanent by NSF; this means that the additional funding level will be assumed and will participate in across-the-board increases for the foreseeable future. In 2007 NEH also renewed through 2012 the agreement though which it cooperates with NSF in the Documenting Endangered Languages program.

Other organizations have made significant contributions to the worldwide effort to document endangered languages. The Social Science and Humanities Research Council of Canada has made many awards. The Foundation for Endangered Languages (UK) and the Endangered Language Fund (USA) have awarded small grants for language work for the past decade; although small in size, these grants have often positioned the recipients to apply for larger awards from other organizations. ELF just started the Native Voices Endowment: A Lewis & Clark Legacy. This endowed program will provide approximately $75,000 a year for work on languages of groups contacted by the Lewis and Clark Expedition in North America. The Grotto Foundation (Minnesota) has supported multiple language programs within its mandated geographic region. An early funding effort in Japan (Endangered Languages of the Pacific Rim) resulted in 670 million ($5.8 million) in funding from 1999-2003. Two projects are expected to revive the effort in Japan: the Global Center of Excellence, a program run by the Japanese Ministry of Education, provides support for training for and conducting descriptive linguistic research on minor languages. It is expected to last from 2007-20 12 with approximately 100 million ($0.7 million) per years. The Research Institute for Languages and Cultures of Asia and Africa (ILCAA), which is part of Tokyo U of Foreign Studies, is expected to start in 2008 and last for five years. It aims at improving

the academic infrastructure and international collaborative network for supporting descriptive linguistic research.

Although no formal evaluations of programs had been attempted, it was the consensus of the group that the funding has produced a marked increase in activity on endangered languages. This work would have been easier to accomplish thirty or fifty years ago in regards to the availability of fluent speakers, but the political climate and public awareness were not conducive to such an effort. The recording tools of today have many advantages over those of the past, and the digitization of current and even previous resources promises much wider utility of what is collected now. The urgency of the issue remains: If we are to do any further documentation of the thousands of endangered languages, we have to do it now while they are still spoken. There are promising signs that some funding will continue to be available, but many institutional roadblocks remain. It was pointed out that the percentage of linguistics dissertations dealing with endangered languages doubled between 1995 and 2006—but only from 1.1% to 2.4%. This represents a minimal response to an urgent issue.

With respect to the present state of research activity, reports were given on fieldwork, archives, tools and revitalization efforts. It was extremely clear that, despite increased attention in recent times, the mechanics of language documentation are still underdeveloped. There are not enough tools to do the job. There are not enough digital archives to properly store (and make accessible) the results. Standards for metadata and archiving are still developing and in need of a large, sustained effort. Making the tools and results last (sustainability) is another difficult challenge. The Open Language Archive Community (OLAC) has a list of archives that subscribe to their standards, but the OLAC organizers know that there are many more collections extant than are represented. Getting fuller participation is a major goal. The Ethnologue, which serves as the basis for the International Standards Organization (ISO) language codes, hopes to improve the feedback mechanisms that maintain the data quality of the collection and the codes.

Some of the large archives gave updates on their efforts. The Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), provides archiving for Australia and the Pacific. Although digital formats are used for the (currently) over 1800 hours of audio, access is currently limited to the depositor and to cultural centers in the native-language communities. The Archive of the Indigenous Languages of Latin America (AILLA) currently has material on 150 languages, including about 1,000 hours of recordings and 16,000 pages of text and pictures. Their “graded access” system has been very effective at restricting sensitive material, but it appears that more material is restricted than needs to be. A major reason seems to be that the researchers have not had time to sort out what is truly sensitive from the rest. The National Anthropological Archives of the Smithsonian Institution contains 9,000+ linear feet of manuscripts, 635,000 photographs, 8 million feet of moving images, 11,400 sound recordings, and 21,000 works of art, much of it language-related. Some of this material has been accessed with the help of DEL grants in recent years. Further work includes putting ISO-63 9-3 language codes into the catalogs and surveying currently un-catalogued material.

The HRELP Endangered Languages Archive (ELAR) currently has almost a terabyte of data and expects to double that in the next year. Over 6,000 audio files are archived already, along with video, still images, and text. The DoBeS archive at Nijmegen, funded by VWS for grantees, has established a metadata standard that is used in its own collection and by others (IMDI) and provides several ways of accessing and enriching the data. About 6,000 hours of audio and 19,000 hours of video recordings are archived, most of which are only accessible on request to others than the depositors. Many of the issues about standards and access are still quite new, and the field has yet to address them fully. The aim is an online multimedia archive for endangered languages.

Larger collaborations are likely to be needed and can certainly be useful. In the text domain, an interesting example is the Digital Repository Infrastructure Vision for European Research (DRIVER) project. This collaboration across multiple sites in Europe has found it challenging just to make simple text files sharable across a number of formats. The lessons learned for interoperability should be useful for the issues related to the highly encoded linguistic material relevant to endangered languages. The Open Society Archives have also confronted issues of access and found that many of the issues can be dealt with in a way that allows quite liberal access over the internet.

Language revitalization has taken advantage of new technologies. Podcasts are increasingly common, for example, and are used in the Mohegan language project. Language material is needed for classroom exercises for the languages attempting a revival or maintenance in the face of a small number of native speakers (typically grandparents).

Future Directions:

The VW Stiftung hopes to extend its initiative by several years. This will depend on the evaluation of the program that is currently being carried out by an international team of experts. News on this will appear in early 2008.

The European Science Foundation is in the final stages of approving an initiative entitled “BABEL -- Better Analyses Based on Endangered Languages”. If approved, the initiative will invest in work on several fundamental questions, the major one being the relationship of data (especially endangered language data) to linguistic theory. While gathering new language material is included in the mandate, the crucial next step of injecting the results of the collection efforts into theoretical discussions will be addressed directly. The level of funding will depend on the amount pledged by the signing national foundations. The US NSF has already expressed its commitment to this initiative.

There was wide enthusiasm for an addition to the Ethnologue database and/or a separate database that would indicate the level of documentation of languages. It was thought that this would help direct attention to those languages that most needed it. Whalen and Simons reported on work in progress showing that not only are languages endangered, but up to 75% of language **families** are endangered as well. Even though new languages

continue to arise, the diversity that current language families show is unlikely to appear again for many centuries, if ever (levelling influences in today's highly connected world are likely to continue; any future diversity would likely have different features). It may be possible to obtain the documentation-level data by self-report and correction, which would allow this portion of the effort to progress with a more economical investment of funding.

As stated in the summary above, the sense of the meeting was that there is great urgency but also growing accomplishment. Not only are more resources needed, but better connections should be made with similar efforts in the archiving and library worlds. Legacy material currently lacks a secure funding source. And revitalization efforts that use this material have a hard time locating needed support.

Meeting Acknowledgements:

James Herbert undertook the original conceptualization of the meeting as well as the task of creating the initial invitation list, drawing on the advice of Joan Maling and Anna Kerttula at NSF and many experts outside the Foundation. David Lightfoot, Assistant Director, Directorate of Social, Behavioral and Economic Sciences, made a promise of budgetary support that made the planning possible. The initial goal was to have the meeting at the Bellagio Study and Conference Center, through the Rockefeller Foundation. That proposal, however, was not funded.

When he joined NSF in 2006, Douglas H. Whalen took on the task of reviving the conference. A new venue was located, one which had the pleasant effect of allowing broader participation. All the original participants were re-invited and new ones were added. Elizabeth C. Zsiga and Natalie Schilling-Estes at Georgetown University kindly consented to administer the grant that would allow the non-NSF participants to have their travel and accommodations paid for. Johnny Casana at NSF provided a great deal of the coordination needed to keep all the arrangements in place. Kim Teague of Georgetown provided further logistical support.

As always with a workshop, it was the enthusiastic participation of the attendees that made the event a success.