LSA 2005 Summer Institute Proposal
Best Practice in Language Documentation Technology: Course and Workshop

Overview

This proposal is for a joint course and workshop for the 2005 Linguistics Summer Institute on tools and standards for producing and archiving language documentation materials. The course is being organized by the Open Language Archives Community (OLAC) Outreach Working Group, and the workshop is being organized by the Electronic Metastructure for Endangered Languages Data (E-MELD) project. While both the course and the workshop could function as stand-alone events, the two sets of organizers believe each event would be enhanced by the other. Because of this, the two proposals have been put together into one document. We realize, of course, that the Institute organizers must take a number of factors into account in arranging the schedule, and we are prepared to be flexible in organizing these two events in order to best accommodate their requirements.

Immediately below, we give a general rationale for incorporating a course and/or workshop on issues in language documentation technology into the Institute program. We then give separate overviews of the course and the workshop followed by a brief description of the advantages to be found in coordinating the two.

Rationale

Over the last several decades, the rise of new technologies has drastically changed the nature of linguistic resources. Formerly, primary documentary evidence for a language consisted of notebooks and analog recordings, whereas today linguists doing documentary work have a bewildering array of choices of media, tools, and standards to select from in creating materials.

The available options for producing textual materials offer a good example of the sorts of complications that confront linguists in the digital age. While there is no shortage of programs for the creation of texts, the most popular ones, such as Microsoft Word, are aimed at the business world where documents typically need to be preserved for only a few years. Language documentation materials, especially those concerning endangered languages, ought to be preserved for centuries. Furthermore, much linguistic documentation consists of highly structured information, such as entries in a lexicon or morphological paradigms. Word processors are chiefly designed for unstructured prose. Structured linguistic data should be entered into a database where its structure can be properly encoded. There are a variety of database programs which can be used for this purpose, but most of them have the same limitations as word processors—they produce proprietary, short-lived formats that are difficult to migrate forward as technologies and methodologies evolve.

A number of initiatives have recently been launched to assist linguists in creating archive-quality digital resources. These include, among others, the International Standards for Language Engineering Metadata Initiative[1], the Digital Endangered Languages and Musics Archive Network[2], the Open Language Archives Community [3], and the Electronic Metastructure for Endangered Language Data initiative[4]. These organizations have been working collaboratively over the past three years to develop best practice recommendations for linguists, especially for documentary linguists working with endangered languages. The course and workshop proposed here would bring together a number of individuals with expertise in linguistic resource creation and archiving to present recommendations in these areas—to students, in the case of the course, and to experienced field linguists, in the case of the workshop. In addition, each seeks to give documentary linguists an opportunity to see how well existing recommendations apply to a range of linguistic resources, thereby allowing them to suggest revisions to the standards which would facilitate their work.

While much of the current work in the area of language documentation technology is focused on endangered languages, the tools and standards being developed will be of value to any linguist who works with corpora of texts or recorded speech. The potential audience for the course and workshop, therefore, encompasses not only field linguists but also sociolinguists and others working with corpora of better-described languages.

[1] IMDI; http://www.mpi.nl/IMDI/
[2] DELAMAN; http://www.delaman.org/
[3] OLAC; http://www.language-archives.org/
[4] E-MELD; http://www.emeld.org/

Course proposal

Course structure: The three-week course will meet four days a week, Monday through Thursday. We would prefer the course to take place in the first three weeks of the program, so that it will be convenient for participants in the E-MELD workshop (discussed below) to stay on for the second section of classes, if desired. We are requesting a computer-equipped classroom, since the course will include direct application of the principles and methods being taught. Students will be encouraged to bring their own data, in the form of recordings and/or digital texts, if they have any. They will also be encouraged to bring any recording equipment that they are planning to use in their work, since the course will include evaluation of various recording tools. Archives such as the Archive for Indigenous Languages of Latin America (AILLA) and the Alaska Native Language Center (ANLC) will contribute legacy materials, such as analog recordings and unanalyzed texts, that students can work with as well.

The course will cover the following topics:

- Tools and standards for corpus management, including metadata standards;
- Tools and standards for the creation of audio resources;
- Tools and standards for the creation of linguistic texts, including interlinear text;
- Tools and standards for the creation of linguistic databases, with a focus on lexical databases;
- The use of ontologies to create interoperable resources;
- Managing rights and access to digital resources, including addressing the needs and sensitivities of native-speaker communities;
- Organizations involved in the creation and development of standards for the production of linguistic resources.

Course Instructors: There is no one individual who has the expertise to teach all aspects of the course. Instead, we aim for a team-taught course with a series of different instructors.

Possible instructors (and the topics they will cover):

- Helen Aristar-Dry, Eastern Michigan University (survey of standards-setting initiatives and current proposals)
- Steven Bird, University of Melbourne (tools and standards for the creation of linguistic texts and databases)
- Heidi Johnson, University of Texas at Austin and the Archive of the Indigenous Languages of Latin America, and Gary Holton, University of Alaska, Fairbanks and the Alaska Native Language Center (tools and standards for the creation of audio resources and for corpus management.)
- Gary Simons, SIL International (tools and standards for the creation of metadata for linguistic resources)
- Terry Langendoen, University of Arizona (the use of ontologies in linguistic resources to promote interoperability)
- Peter Wittenburg, Max Planck Institute for Psycholinguistics, Nijmegen, or Mark Liberman, University of Pennsylvania (managing rights and access to digital resources)

Workshop proposal

Background: The E-MELD project (http://emeld.org/) is an NSF-funded project with a dual objective: (1) to preserve endangered languages documentation and (2) to aid in the development of electronic infrastructure for linguistic archives. One mandate associated with the second objective is to involve a large segment of the linguistics community in the development of archiving standards. To that end, E-MELD is funded to hold annual workshops for the duration of the five-year project. The first workshop, held in 2001 in conjunction with the Santa Barbara LSA Linguistic Institute, focused on the need for standards in digitizing linguistic data. The second in 2002 in Ypsilanti focused on standards for digitizing lexical information. The third, held in 2003 in conjunction with the LSA Linguistic Institute in East Lansing, focused on digitizing texts and field recordings. And the fourth will be held at Wayne State University July 15–18 2004; it will focus on linguistic databases and best practice. These workshops have provided a forum for language engineers to get together with working linguists involved in language documentation to discuss archiving issues and design future initiatives. They have also allowed the E-MELD project to discuss its progress with members of the discipline and to evaluate emerging standards in light of the needs of the linguistics community. The two workshops held in conjunction with the LSA Linguistic Institutes have been

particularly successful, drawing on average twice the number of participants expected and allowing E-MELD to identify numerous individuals and projects hitherto unknown to us who are involved in the creation of digital linguistic resources. Although the 2005 workshop will have a different format and purpose, it will also be open to the public; and our prior experience suggests that holding the workshop in conjunction with the LSA Linguistic Institute would serve a burgeoning interest in the discipline in digital language documentation.

Purpose and Format: The 2005 E-MELD Workshop will be a week-long "technical institute" designed to give working linguists hands-on experience applying recommendations of best practice to their own language documentation. The format planned for the workshop includes five working days in a computer lab, followed by a one-day conference designed to report and evaluate participants' experiences and results.

In the fall of 2004, we will advertise this opportunity internationally and choose ten to twelve participants to be funded by the workshop, as well as four to five recognized experts in digital language documentation who will serve as workshop staff. The invited participants will be expected to bring lexical material on an endangered language which does not currently exist in a best practice format. During the first five days of the workshop, the invitees will work with the language engineering experts to design and implement a conversion path for this legacy material. The experts will also offer one tutorial session each day based on the identified needs of the participants. These will differ from the lectures planned for the course in that they will treat very specific topics like "Converting IPA Kiel to Unicode" or "From Filemaker Pro to Best Practice" and they will include hands-on work for attendees. These tutorials will be advertised in advance and will be open to all LSA Institute participants, though they are expected to be of particular interest to students enrolled in the archiving course. The workshop staff will also be available for a specified period each day to answer questions and give advice to walk-ins on electronic archiving issues.

By the end of the five-day work period, the invited participants will have produced some portion of an archive-quality lexicon of an endangered language, thus adding to our small store of enduring digital resources based on irreplaceable legacy documentation. The participants will also have gained experience working with the standards and software which support best practice, so that they can not only finish the work begun on their own documentation but also serve as mentors for other descriptive linguists.

On the sixth day of the workshop, participants and staff will report on their experiences in an open forum; the purpose of this mini-conference will be to expose additional members of the linguistics community to emerging digital standards, to involve additional linguists in the standards-setting enterprise, and to evaluate the work to date. workshop participants will be ideally suited to evaluate the practicality of existing standards, to identify software needs, and to suggest future initiatives.

Participants: As noted above, the field linguists to be invited as funded participants will be chosen in Fall 2004 as the result of a competitive application procedure. The choice of workshop staff will depend to some extent on the nature of the documentation to be brought by the participants. However, we anticipate that we will need at least:

- A computational linguist with experience in corpus conversion. Possibilities include Baden Hughes (U. of Melbourne), Steven Bird (U. of Melbourne), Mark Liberman (U. of Pennsylvania).
- An expert in the archiving of audio materials. (Since the 2005 workshop will focus on lexical information, we do not expect to deal with video-archiving.) Possibilities include Daffyd Gibbon (U. of Bielefeld), Peter Wittenberg (Max Planck Institute of Psycholinguistics), Bartek Plichta (Michigan State U.), Chilin Shih (U. of Illinois).
- A linguist/language engineer with specific experience in the design of lexicon schemas and metaschemas. Possibilities include Gary Simons (SIL International), Mike Maxwell (U. of Pennsylvania), Terry Langendoen (U. of Arizona), Will Lewis (California State U., Fresno), Scott Farrar (U. of Bremen)

Requested Support: As noted above, the E-MELD grant can fund the travel and accommodations of the invited participants and the workshop staff during the week of the workshop. What is requested from the Institute is

(a) A computer lab which will accommodate fifteen to twenty people, with Internet access, Windows machines, and the ability to install software (Days 1–5)
(b) A meeting room with internet access, projection facilities, and seating for 50 attendees (Day 6)
(c) Institute housing for participants and staff

Prospects for course and workshop coordination

Assuming a three-week course and a one-week workshop, the intention of the organizers is to hold the workshop concurrent with the final week of the course. By this point, the students in the course will have had enough exposure to issues in language documentation technology that they will be able to interact productively with the E-MELD advisors and the invited field linguists. In particular, in the third week of the course, the lectures will continue as in the first two weeks, but the applied aspects of the course will be expanded and include participation of the students in some of the activities of the workshop—including participation in the workshop working groups.

This coordination will benefit the workshop organizers by giving them a wider range of participants than they would otherwise have access to. It will benefit the students in the course by giving them hands-on experience in using and evaluating standards for language documentation beyond what would be possible in the course itself. In addition, since many of the students in the course will, no doubt, have enrolled in order to prepare for their own field work, they will have the added benefit of becoming acquainted with expert field linguists who will be able to give them advice based on their experiences collecting data in the field.

Beyond the intellectual benefits of coordinating the course and the workshop, it is worth pointing out that there are also financial benefits. Several of the workshop advisors will also be able to serve as course instructors. Since E-MELD has some funding for advisors, this will make a team-taught course of the sort described above more economical than it would otherwise be since each instructor will not need to be independently funded by the Institute. In addition, this coordination will permit E-MELD to obtain feedback from a larger group of individuals than would be possible if they were limited solely to working with invited field linguists.

Endorsements

This workshop has been endorsed by the LSA Committee for Endangered Languages and their Preservation (CELP), the Endangered Language Fund (ELF), and the Dokumentation Bedrohter Sprachen Project (DoBeS).

Contact information

Course organizers
Jeff Good
Chair, OLAC Working Group on Outreach
University of Pittsburgh
2816 Cathedral of Learning
Pittsburgh, PA 15260
jcgood@pitt.edu

Heidi Johnson
Project Manager, The Archive of the Indigenous Languages of Latin America
Dept. of Anthropology
1 University Station C3200
The University of Texas at Austin
Austin, TX 78712
hjohnson@mail.utexas.edu

Workshop organizers
Helen Aristar-Dry & Anthony Aristar
Principal Investigators, E-MELD & Moderators, The LINGUIST List
Address: Aristar-Dry
c/o Dept. of English Language and Literature
Eastern Michigan University
Ypsilanti, MI 48197
hdry@linguistlist.org