E-MELD: Electronic Metastructure for Endangered Languages Data

Anthony Aristar
Wayne State U.
aristar@linguistlist.org

Helen Aristar Dry
Eastern Michigan U.
hdry@linguistlist.org

Paper presented at the workshop on
Web-Based Language Documentation and Description
12-15 December 2000, Philadelphia, USA.

Abstract:

The LINGUIST List <http://www.linguistlist.org> is organizing a collaborative project with a dual objective: (1) to preserve endangered languages (EL) data and documentation and (2) to aid in the development of infrastructure for linguistic archives. The 5-year project, tentatively entitled E-MELD (Electronic Metastructure for Endangered Languages Data), is a 6-institution collaboration involving individuals from The Linguistic Data Consortium, The Endangered Languages Fund, and The Summer Institute of Linguistics, as well as the University of Arizona, Eastern Michigan University, and Wayne State University. Funding is currently being sought; if support is secured, the project will begin in the fall of 2001. One outcome of E-MELD will be a LINGUIST List digital archive housing data from 10 endangered languages. But the focus on infrastructure will produce other results as well. The LINGUIST archive will be designed to function, not only as a repository, but also as a "showroom of best practice." The archive will offer EL data marked up and catalogued according to community consensus about best practice, as well as reference material delineating best practice and software tools supporting it. A second outcome will be the establishment on the LINGUIST List site of a central metadata server for the discipline. This facility will collect and distribute metadata on all the language-related resources residing at distributed sites, not just information on EL data alone. And a third outcome—perhaps the most important—will be the involvement of a large segment of the linguistics community in the various enterprises underlying the archive and server. Although the data collection efforts will focus initially on endangered languages, the metadata server, the recommendations for best practice, and the distribution of supporting software will contribute to the development of infrastructure necessary to the digitization of all types of language documentation.

1. INTRODUCTION

Language data is central to the research of a large social sciences community, including not only linguists, but also anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous peoples. Members of this research community are currently faced with two urgent situations: the number of languages in the world is rapidly diminishing while the number of initiatives

to create digital archives of language data is rapidly multiplying as a result of the increasing availability and sophistication of web technology. The latter might seem to be an unalloyed good in the face of the former, but there are two ways things may go wrong without adequate collaboration among archivists, linguists, and language engineers. First, a common standard for the digitization of linguistic data may never be agreed upon. And the resulting variation in archiving practices and language representation would seriously inhibit data access, searching, and scientific investigation. Second, standards may be implemented without guidance from the people who best know the range of structural possibilities in human language—descriptive linguists who have documented hundreds of little-known languages. Guidelines which are designed on the basis of well-known western languages will not be adequate to the urgent task of archiving as much linguistic data as possible in the face of widespread language attrition and loss.

If digital archives of language data and documentation are to offer the widest possible access and to provide information in a maximally useful form, consensus must be reached about certain aspects of archive infrastructure. The LINGUIST List <http://www.linguistlist.org> is attempting to organize a collaborative project with a dual objective: (1) to preserve EL data and documentation and (2) to aid in the development of infrastructure for linguistic archives. One outcome of the project will be a LINGUIST List digital archive housing data from 10 endangered languages (ELs). But the focus on infrastructure will produce other, equally important results. In the first place, The LINGUIST archive will function, not only as a repository, but also as a "showroom of best practice." The archive will offer EL data marked up and catalogued according to community consensus about best practice; furthermore, the archive will disseminate reference material delineating best practice and software tools supporting it. Another outcome of the project will be the establishment on the LINGUIST List site of a central metadata server for the discipline; this metadata server will eventually organize information on the multitude of language-related resources residing at distributed sites, not just information on EL data alone. And a third outcome—perhaps the most important—will be the involvement of a large segment of the linguistics community in the various enterprises underlying the archive and server.

Although the data collection efforts will focus initially on endangered languages, the metadata server, the recommendations for best practice, and the distribution of supporting software will have a significant impact on all empirical research in linguistics.

2. THE PROBLEM

The death of a language or dialect represents not just an academic loss, but a significant loss in knowledge and culture. For this reason, organizations across the world are addressing the problem of language endangerment; and a number of digital archives of EL data are currently being planned or developed. Your handout includes a partial list.

The establishment of multiple archives is to be welcomed, since the magnitude of the task requires distributed effort. No one institution can archive all the important data on all the currently endangered languages—certainly not within the time limits imposed

by impending language attrition and by the ongoing deterioration of the existing documentation. Digital archiving at distributed sites offers the best hope for preserving this valuable linguistic material. But developing all the infrastructure necessary for a digital archive of language data (including delivery mechanism, formatting guidelines, and supporting software) is a huge task that is beyond the capacity of any single institution to accomplish on its own (Simons, 2000:1) And once multiple institutions have set up online archives, resorting to different strategies for designing infrastructure, it will be more difficult to implement any general solution.

Without such a common infrastructure, the individual linguist will find it very difficult to identify all the resources pertinent to a given language. To posit an extreme case: the language in question may be classified, or even named, differently in different archives (e.g., Waikurean vs. Guaicuruan, Lappish vs. Sami). The language data may be marked up using different sets of structural tags (e.g., possessive vs. genitive). The texts may have different organizations (e.g., chronological organization vs. frequency organization of the senses in a dictionary entry). And the files may have different formats because they have been created with incompatible software tools. In this situation, even a linguist with access to resources might not be able to compare them well enough to make reliable linguistic judgments. But—what is perhaps even more disturbing—locating all the relevant material in the first place will be a formidable task. It is unlikely that all the sound and video recordings, texts, grammars, dictionaries, and cultural information pertinent to a given language will ever reside on a single site. And if various archives develop different ways of describing and indexing their resources, no central meta-index can easily be developed: the amount of data will defeat a human librarian, and the different formats will defeat a machine.

2.1. The Scope of the Problem

All of the problems enumerated above arise in the context of archiving any electronic language data, not EL data alone. It is the impending disappearance of so many endangered languages that leads us to focus first on this aspect of the more general language data problem. However, this focus has a distinct—although paradoxical—benefit: the challenging nature of the data set. Many, if not most, ELs have structures which diverge so widely from each other and from those of Western European languages that metadata and markup guidelines adequate for these languages will almost certainly be adequate for other language data as well. Thus an attempt to define standards for the digitization of ELs is, in fact, also an attempt to define standards for the digitization of languages in general.

3. TOWARD A SOLUTION: E-MELD

3.1. Necessary Elements

Any attempt to address the language archiving problem must have at least 3 components.

1) **Community Involvement.** All the different stakeholders in the EL archiving enterprise must be kept fully informed and continually consulted: we must (a) enhance communication between descriptive linguists and computational linguists, (b) ensure that computational solutions and supporting rationales are promulgated effectively throughout the linguistics community, and (c) involve indigenous communities in archive design.

2) **Flexibility.** Any proposed solution must (a) have the capacity to handle legacy data in various formats and (b) allow for some continuing variation in individual practice.

3) **Collaboration.** Organizations must pool their resources in light of: (a) the volume of work and the range of expertise needed for a unified solution and (b) the danger that partial, uncoordinated "solutions" will only exacerbate the problem.

The E-MELD project was structured with these 3 requirements in mind. It implements part of a distributed solution proposed in Simons (2000a), which recommends a coordination of effort among the Linguistic Data Consortium, the Summer Institute of Linguistics, and The Linguist List: The Linguistic Data Consortium will function as a central repository of standards and software (which may be developed elsewhere); the Ethnologue will constitute the standard reference for language classification; and The LINGUIST List will serve as a central repository of metadata, as well as an institutionalized conduit of information between language engineering projects and the linguistics community.

3.2. Project Components

The E-MELD project has been proposed to funding agencies. As yet, we do not know the outcome, but the full proposal is available at http://linguistlist.org/proposals/NSF-infra-proposal.rtf. In its general outlines the E-MELD project involves:

1. Formulation and promulgation of best practice in:
    a. linguistic markup, i.e., of texts and lexicons
    b. the creation of metadata for language resources
2. Establishment of a metadata server on the LINGUIST List site:
    a. Database configuration
    b. Design of web interfaces for input, query, and display
    c. Collection of metadata on existing language resources (not just EL resources)
    d. Conversion of metadata in foreign formats into the best practice format
    e. In addition to metadata describing the language resources and their availability, the database will offer:
        i. Typological information collected via questionnaire
        ii. Genetic and ethnographic information provided via an interface to the Ethnologue (http://www.sil.org/ethnologue)
3. Data conversion and software development
    a. Conversion of texts and lexicons from 10 endangered languages into the best practice format

    b. Development of markup conversion software
    c. Development of software for field linguists facilitating the use of the recommended markup
4. Establishment of an EL archive which will also function as a "showroom of best practice," making available:
    a. Data from 10 ELs representing a wide geographical and genetic spread
    b. Software tools (described above)
    c. Reference material (e.g., files & hyperlinks) delineating recommended standards
    d. A Query Room, where questions may be addressed to native speakers and additional data provided upon request
5. Organized communication with the research community, involving:
    a. 3 Workshops: 2-3 day sessions with limited agendas intended to promote communication among field linguists, archivists, and computational linguists. Funding for the first of these workshops has been awarded by the NSF Linguistics Division.
    b. 3 "Digital Institutes": 1-week institutes for 10-15 field linguists designed to:
        i. Introduce proposed recommendations of best practice in markup and metadata
        ii. Distribute the field software and provide training in its use
        iii. Encourage the participants to test both the software and the proposed markup on substantial samples of their own data
    c. Liaisons with professional associations, e.g. CELP, ELF, FEL, ICHEL
    d. Regular email bulletins distributed via The LINGUIST List
    e. An E-MELD homepage on the LINGUIST List site offering:
        i. Project descriptions and progress reports
        ii. Web questionnaires soliciting feedback on proposed guidelines
        iii. Vitae and contact information on key project participants
        iv. Organized sets of links to related sites
        v. Detailed reports on Workshops and "Digital Institutes" (see above).

The need for metadata and markup standards is already apparent to this audience; and in fact they are the foci of their own sessions at this workshop. So here we will emphasize aspects of E-MELD that will not be discussed elsewhere, that is, (2), (4), and (5) above: the establishment of a central metadata repository, the "showroom of best practice," and the need for organized involvement of the linguistic community in the development of archive infrastructure.

3.3. A Central Metadata Server

One of the most important parts of the E-MELD project is the initiative to collect metadata on language resources at a central site. Though we will focus initially on EL resources, the facilities created can and will be extended as soon as possible to catalogue linguistics-related resources of all types. Such a catalogue will not only allow extant

material to be identified and retrieved; but it will also enable distributed data to be pieced together. Given a markup standard and a metadata server, it will not matter if a dictionary of a language appears at one site and a grammar of the same language appears at another. They can be linked through their metadata, and used in conjunction with one another. But in order to establish such a central index, it will be necessary to adopt and promulgate recommendations of best practice in the creation of metadata for language resources, as well as to collect existing metadata and convert it into this format, and to institute user-friendly systems for input and query of the information.

The LINGUIST project will create user-friendly web interfaces for metadata input; and the PIs will contact cooperating archivists to elicit the exchange of metadata. In addition, LINGUIST intends to implement an innovative procedure to identify other sites on the Internet which store language data but may not yet have participated in the project. This will involve using a spider to index other linguistics-related sites and configuring search software to search the index using a keyword list. In this way potential sources of metadata may be identified. The site owners will then be approached and invited to contribute to the database. The specifics of this idea will be presented in the panel discussion tomorrow.

We believe that LINGUIST can play a role in involving the linguistics community in standards-setting, as well as in disseminating the results. For instance, we intend to hold a workshop where field linguists and language engineers can meet and discuss potential problems. The workshop is scheduled for June, 2001 in Santa Barbara, just before the LSA institute; and key participants in this Workshop will be attending. We expect field linguists to be particularly interested in markup and in two potentials of metadata: to implement the definition of equivalences and to include typological information.

3.4. Markup

The establishment of standards of best practice in linguistic markup is crucial for the understanding of *any* language data, but especially for the complex structures regularly found in ELs. This is true for two reasons. First, without compatible markup, no two bodies of data are comparable. The linguistic similarities and differences will be difficult to see even by human inspection. Computationally they are essentially undiscoverable, since no search-engine can be expected to "know" that differently named entities are equivalent. Second, a lack of standardization makes data difficult to interpret in and of itself, because a linguist must first learn the nature of the data markup before he or she is able to understand a new body of data.

A start was made on the difficult issue of linguistic markup as part of the Text Encoding Initiative (TEI) (http://www.uic.edu/orgs/tei/), an international project whose main aim was to develop guidelines for the preparation and interchange of electronic texts for scholarly research. Unfortunately, the work on linguistic markup within the TEI was incomplete when the project ended. Moreover, the markup structures recommended by TEI have generally been judged too unwieldy, and the recommendations themselves too complex, to have gained much acceptance in the linguistics. On the other hand, the

simpler markup standards currently in existence— such as those developed by the Expert Advisory Group on Language Engineering (EAGLES) (http://www.ilc.pi.cnr.it/EAGLES/annotate/annotate.html)—are designed to handle only Western European languages.[1] As a result there still exists no set of guidelines adequate for the markup of EL data.

In modifying and extending existing specifications for linguistic markup in light of the particular needs of the EL community, the project will focus on the markup of dictionaries and of glossed text, since successful development of recommendations in these two areas will deal with some of the most pressing problems currently faced by the community. One is lack of standardization in electronic texts representing language data,[2] and a corresponding lack of interoperability in corpus-handling software. Electronic EL texts, in particular, have been developed using a wide variety of different standards, and, while they are usually consistent in themselves, little thought is typically given to making them conform to a more widely acceptable format.[3] Another is the inability to find and compare instances of specific grammatical forms. A linguist, for example, might well need to extract from a set of glossed texts every example of a first-person plural object marker occurring in the same word as a third-person singular subject marker. Or, to take another, more concrete example, a linguist might with to query a lexical database of Yoeme material to find the base forms for every verb that has a distinct variant when the subject is plural (a matter of considerable interest for doing comparative Uto-Aztecan research), or all the reduplicated forms together with their base forms. To do this simply by looking through the printed text would be prohibitively time-consuming.

Hence success in standardizing electronic dictionary and glossed text markup will offer the linguistics community tools which have significant research advantages over what is currently available. And it will make a start toward providing a *catalogue raisonnée* of the morphological distinctions found in the languages of the world, with guidelines for their encoding. (For example, when should one use “genitive”, when “possessive”, and when doesn’t it matter?) One of the problems, particularly in doing comparative linguistic research and dialect variation studies using glossed text markup, is understanding how and when the same technical term is understood differently in different texts, and conversely how different terms may be understood to denote the same

---

[1] The EAGLES group is continuing its work as part of the International Standard for Language Engineering group (ISLE) (http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm), in conjunction with the Linguistic Annotation project at the Linguistic Data Consortium, but there has as yet been no specific markup standard developed, although general guidelines have been widely accepted. The LINGUIST project has advisors closely associated with both ISLE and the LDC, so we will not duplicate work underway in either workgroup.

[2] This problem is far less characteristic of electronic dictionaries, since prestigious organizations, e.g. the Oxford Text Archive and the OED, have already taken the lead in designing and exemplifying best practice in entry formatting. However, the problem of linguistic markup within entries still remains. Since these electronic dictionaries treat primarily western European languages, we still lack adequate markup to represent the morphological and syntactic structures of Els.

[3] For example, the Australian National University, the Summer Institute for Linguistics, and the U. of California at Berkeley have developed formats for EL texts which are followed consistently within each publishing domain; but the 3 formats are not consistent with each other.

thing. Attempts have been made to systematize grammatical descriptions (e.g. Comrie 1987) but these still need to be translated effectively into a computational environment.

3.5. Publicizing and enabling best practice: The "Best Practice Showroom"

If the project is funded, we will undertake numerous initiatives to ensure that the linguistics community is kept fully informed of the progress made and that all the data and software produced is made freely available. In addition to general publicity initiated via The LINGUIST email list and website, we will hold 2 summer workshops and 2 "Digital Institutes" in order to make sure that an influential core of linguists has the opportunity to learn about and provide feedback on the metadata and markup recommendations. A the weekend workshops for archivists & field linguists, we will solicit input on markup and metadata standards, and test and publicize the resulting recommendations. At the "Digital Institutes" field workers will be provided with software to facilitate best practice, trained in its use, and asked to provide feedback regarding its helpfulness with their data

Furthermore, our work on formats and tools will be disseminated in a practical fashion, as a `best practice showroom.' The 'Best Practice Showroom' will include data from 10 endangered languages, marked up as recommended, and user-friendly web interfaces for input, query, and flexible data display. Providing EL data is important, not only because it will preserve information about the languages in question, but because it is not possible to promulgate standards in the abstract. Rather, data must be provided, both as examples and as lures. Data will not only attract visitors to the site but also induce them to learn the recommendations for best practice in the course of using them to manipulate the data.

The LDC, the ELF and SIL have each undertaken to provide data from two languages and convert them into the recommended format. LINGUIST will add data from two more languages. And data from two others will be provided by post-doctoral associates chosen for their ability to contribute to the project. Thus we will be able to open the "Showroom Of Best Practice" with data from 10 diverse languages. These languages offer a challenging range of linguistic features on which to test both the completeness of the markup tagset and the utility of the data conversion software. They include: Biao Min and Mocovi (data prepared by LINGUIST), Ega and Cambap (data prepared by the scientific Linguistic Data Consortium), Tofa and Lakota (data prepared by the Endangered Languages Fund) , and an Austronesian and a Papua New Guinean language as yet to be determined (data prepared by SIL).

4. Benefits

The creation of well-designed EL archives will have numerous benefits for scientific research. For example, future scholars will have direct access to enormous numbers of fieldwork notes and recordings; they will not have to rely on data as presented in third party discussions. Cross-linguistic hypotheses can be pursued. Linguistic features can be statistically analyzed, studied in context, and plotted on maps. Perhaps most importantly, data will be preserved for whatever use future scholars may

wish to make of it. There are also important educational benefits. Members of small communities who are losing their ancestral language, or indeed have already lost it, can use such an archive for purposes of study or revitalization efforts. However, to ensure that we reap these benefits from electronic archives of endangered languages, many guidelines for best practice must be developed, publicized, and adhered to. If it is implemented, the E-MELD project should constitute a significant step toward reaching disciplinary consensus about aspects of archive infrastructure.