The E-MELD Project

Anthony Aristar & Helen Aristar-Dry

1. An Overview of the E-MELD Project

The E-MELD project was begun as part of a general consensus amongst the investigators of the project that linguistics had reached something of a crisis point. Language data is central to the research of a large social sciences community, including not only linguists, but also anthropologists, archaeologists, historians, sociologists, and political scientists interested in the culture of indigenous peoples. Yet in the face of the rapid diminution of the number of languages in the world, there is no common standard for the digitization of linguistic data. There have been a number of attempts to set up standards for such data, for example by the ISO organization, but these are largely being 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 instead have thus tended to be designed on the basis of well-known western languages, and are not adequate for the wide range of linguistic diversity that exists.

It was felt that if digital archives of language data and documentation are to offer the widest possible access and to provide information in a maximally useful form, then, consensus must be reached about certain aspects of archive infrastructure. The LINGUIST List <http://www.linguistlist.org> therefore undertook 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. This project was funded by the National Science Foundation of the USA in mid-2001. One outcome of the project will be a LINGUIST List digital archive housing data from 10 endangered languages. 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; it will also disseminate 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, which LINGUIST will set up as part of the OLAC project, using OLAC harvesting techniques; this server will eventually organize information 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.

As a first step in the accomplishment of these goals, a workshop was held, at which linguists, archivists and software engineers could come together, and arrive at a consensus on language engineering standards. The first such conference was held in Santa Barbara, California, on June 21-24, 2001. This paper is a brief progress report of the conference, and emphasizes the issue of language codes, since that is a critical part of the design of a linguistic database.

2. The Recommendations

The workshop was divided into three working groups: the language codes group, the metadata group and the linguistic markup group.

2.1. REPORT ON LANGUAGE CODES WORKGROUP RECOMMENDATIONS

The mandate of this workgroup was to discuss the issue of language codes, given that linguists need a system of codes which makes the special distinctions which they need, but also has the stability and lack of ambiguity required by computational systems. To posit an extreme case: a language may be classified, or even named, differently in different archives (e.g., *Waikurean* vs. *Guaicuruan*, *Lappish* vs. *Sami*). There needs to be a single way of indicating that these are the same language. The group was asked to consider a number of difficult questions, including:

- How far should variant views of whether something is a language or not be considered? For instance, how do we treat dialects?

- How should we handle subgroups as opposed to languages? By the standard historical method, any subgroup is simply a set of languages, all of which are derived from an earlier proto-language. Thus the node which defines a subgroup can also be seen as simply an earlier, extinct language. Vulgar Latin is in this sense equivalent to Proto-Romance. Should we therefore treat nodes in a family tree simply as languages?

- How far should a classification system go? Should a classification system be one that generates an environment which to some degree "knows" the place in a family tree that a language belongs to? We should certainly not have to give an entire tree every time we mention a language; yet we do want to be able to extract all material which belongs to a particular subgroup. How would you implement such a system?

- How can we best handle variant subgrouping ? Whatever coding system we use, it should be able to represent variant trees in a family, rather than imposing one view on the entire community.

- How do we build a system of coding which can handle changing views of groupings? What happens when we add a subgroup, delete one, join families into macro-groupings?

It was the general consensus of the group that present standards, specifically ISO 639, were inadequate for linguistic work, and that there was a need for a comprehensive and officially accepted set of language codes. The proposals that follow were made by the working group.

2.1.1. GENERAL RECOMMENDATION: UNIVERSAL LANGUAGE CODE CONSORTIUM (ULCC)

The most sweeping proposal made at the conference was that an international consortium of linguistics-related groups and individuals be formed as a body which would be responsible for sanctioning (though not necessarily for producing) an inventory of language codes and "standard" views of subgrouping. In the absence of a previously existing or better designation, the group proposed to refer to the set of language tags and subgrouping definitions as the "universal language code" (ULC), and the proposed group as the "Universal Language Code Consortium (ULCC).

They recommended in addition:

- That this consortium be as international and linguistically diverse as possible. To begin with, they presupposed that LINGUIST List, OLAC and the SIL should be part of this consortium. In addition they proposed that representation be invited from the major national and international linguistics societies and standards organizations.

The question of membership by representatives of for-profit corporations was raised, but not discussed in any detail.

- That the experience and practice of existing standards-related consortia be consulted for examples to be followed and pitfalls to be avoided.

The group did not attempt to define precisely what the functions of the ULCC would be, or the processes by which it would operate. The general consensus, however, was that such a group was needed as a means of sanctioning the coding efforts which were already in existence. No coding effort, no matter how complete or well-conceived, would be accepted by the community of linguists unless that community had a stake in the process. This consortium was the means by which that could be achieved.

2.1.2. INDIVIDUAL LANGUAGE CODES

The workgroup agreed with the principle, expressed by Constable & Simons, that "language" be operationally defined by lack of mutual intelligibility with any other speech variety. Although the determination of mutual intelligibility is not always a trivial task "on the ground," for purposes of linguistic research no other basis for classification makes sense. Care must therefore be taken in code documentation to indicate how this determination was made. (Other criteria of course, such as nationality, script, ethnicity, etc., could be used to define other sets of language codes.) In this context, the committee suggested the following:

- That the ULC be based upon those of the SIL's *Ethnologue: languages of the world, 14th ed.* (ed. Barbara Grimes, 2000; Dallas: SIL International), as the most complete set of language codes based upon the intelligibility principle. SIL International has generously offered to make the *Ethnologue* codes available to the linguistic community.

- That the Ethnologue language codes should be supplemented by codes, to be created by LINGUIST, in conjunction with advisors, for those languages not treated by Ethnologue, for example attested ancient and constructed languages.

- That the general principles of Ethnologue be followed in the assigning of codes, though with those modifications made necessary by the peculiar nature of the material from which the evidence for ancient languages is derived. Most importantly:

  - For persistence of reference, once a code is assigned it should not be reused with a new reference, even if it is dropped (through a merger or split) from the officially supported list.

2.1.3. EXTENSION OF CURRENT CODES

The Ethnologue system, as we all know, is a very complex one, and with the addition of ancient language codes, it would become more complex still. It is interesting in this context, therefore, to note that the working group, which was composed largely of field linguists, felt that the system was not full enough for their purposes. Specifically, they felt that as field linguists it was necessary to be able to differentiate not just between languages, but between dialects. They also felt that some central repository of information on family relationships was essential.

Though they did not make any recommendations about how this should be implemented, they made the following suggestions:

- Since it is impractical for a central authority to assign codes for every possible dialect in a language at a central site, these variety/dialect designations would not be directly sanctioned by the consortium. However a mechanism for registration of variety/dialect codes proposed by individual investigators should be arranged on central servers, for example, the OLAC metadata servers, in such a way that material linked to the same distinct variety or dialect of a single language can be related across sites.

- Language grouping information should be kept at a central site, showing the potential relationships between languages.

2.1.4. LANGUAGE GROUPS

In the past, there has been no central repository of information on genetic relationships between languages. Ethnologue, as we all know, does mark such information, but has formalized no system for the indication of such relationships, and makes no attempt to indicate situations where multiple potential groupings exist. Although there is frequently general agreement, for a given language, about some of the larger family sub-groupings (e.g., Romance, Semitic), and often a higher ranked "super-family" (e.g., Indo-European, Afroasiatic), more detailed family-tree structure beyond that rapidly involves conflicting historical scenarios and views about the nature of language change. Thus it seems futile to attempt to define a single set of trees for language relationships. As an interim measures the committee recommended the following:

- The IULC servers should set up a system whereby conflicting information about the subgrouping of a language can be collected in a central place, using a flexible coding system from which multiple relationship trees can be generated if needed. Initially the LINGUIST family coding system (http://saussure.linguistlist.org/cfdocs/pub/find-a-language-or-family.cfm) would be used until one is decided upon by general consensus.

- All subgrouping information provided by the servers, whether about languages or subgroups, should be labeled as to its provenance. It should always be possible to discover which scholars who have proposed it, and who disagrees with it.

- No attempt should be made to present an "approved" system of classification, though information about the greater or lesser acceptance of a particular view should be provided.

This system obviously entails that some centralized organization, however defined, should be mandated to oversee the subgrouping efforts, sanctioned by the ULCC. It also obviously entails a centralized site which would serve as the central store for all the different kinds of information which the system would need to function.

2.2. METADATA ISSUES

2.2.1. MAJOR ISSUES

The metadata group included representatives not just of OLAC, but also of the Isle Metadata Initiative (IMDI). One of the most important issues dealt with by the working group was, therefore, the relationship between the two groups. One major point was made:

- There is a pressing need to harmonize the controlled vocabulary of these two schemata for metadata, especially in the field of content descriptions.

It was pointed out that there was a major difference between OLAC and IMDI, in that:

- IMDI has a focus on storage of primary data rather than just on metadata.

- IMDI evidently offers much finer analysis of the resources it describes than OLAC intends to do through its metadata.

IMDI is thus a complement to OLAC, and offers OLAC a source for material which otherwise would be unavailable.

The working group recommended that OLAC and IMDO should do everything possible to achieve inter-operability between OLAC and IMDI. In particular:

- IMDI might define wrappers for sets of resources, with pointers to OLAC entries.

- Peter Wittenberg of IMDI declared that, where possible, IMDI would carry over the same terminology used in OLAC, subject to the proviso that this remain an open standard.

In return, OLAC records could bear an IMDI icon, showing that they were susceptible of more detailed (IMDI) search.

In particular, the working group felt that specific, technical discussions on how the two formats would interact should be organized, and that these might be suitable for funding within the EMELD grant.

3. Markup Issues

Since Terry Langendoen will be talking in his paper at this conference about the work of the latter group, I will say little more of it here, except to note that the markup group recommended that:

- The EUROTYP markup be used as a basis for linguistic markup, and that a large-scale ontology of linguistic markup be designed.

4. Conclusion

This paper is essentially a report on the recommendations of the Santa Barbara working group, but it is worth taking note of what has been accomplished in the EMELD project. We would like to note that some progress has been made in the months since then. In particular, the EMELD project has instantiated a full database of language codes on its site, and has defined a set of approximately 200 ancient languages, as well as some 20 constructed languages, all of which have been assigned codes and brief descriptions. Since Ethnologue has also generously provided its codes to LINGUIST and EMELD, it is now possible to search both sets simultaneously through a single facility, at the URL http://saussure.linguistlist.org/cfdocs/pub/.

Since this search facility allows for fuzzy matching on strings, it is a very effective way of finding language names and the codes with which they are associated. An initial implementation of a subgrouping system (distinct to some degree from the Ethnologue system) has also been implemented, and is searchable at the same site. Within the LINGUIST site, all data is now categorized by the Ethnologue language codes for modern languages, and by the LINGUIST codes for ancient and constructed languages. Parts of this system are already beginning to appear publicly. For example, it is now possible to look up persons, and find out what languages they have been working on -- or on which they wrote their dissertations -- by their Ethnologue/LINGUIST code, e.g. at the URL: http://saussure.linguistlist.org/cfdocs/pub/LL-WorkingDirs/people/person-lookup-1.cfm. More will soon appear.