SEARCHING AND LANGUAGE CLASSIFICATION

Anthony Aristar, Wayne State University; Michael Appleby, Eastern Michigan University; Gayathri Sriram, Eastern Michigan University.

1. INTRODUCTION

The LINGUIST List's E-MELD project (Electronic Metastructure for Endangered Languages Data) is an initiative to harvest and archive endangered language data and make it available and searchable on the web. Collection and organization of this data is of the highest urgency as the continued existence of some endangered languages becomes ever more uncertain. The success of this project depends on the mass support of researchers and field linguists, who will submit their data by various means, and on that of the user, who will retrieve it. Therefore, the resulting system must be user-friendly and cater to the accepted wisdom of both groups.

This paper will describe on-going work on language classification, being carried out by the Wayne State University and Eastern Michigan University contingent of E-MELD. First, in section 2, it will explain why a single and consistent language classification is important, what the requirements of such a system are, and why current standards are inadequate. Section 3 will propose a better system, one that is in line with the 'best practices' advocated by Bird and Simons (2002). Section 4 will discuss a related problem, of classifying (competing) language families and a solution in the form of a relational database design will be offered. Finally, section 5 will show preliminary user interfaces, to demonstrate what approaches are being taken in user-friendliness.

2. RELATING LANGUAGES TO DATA

When users type a language name into a search engine, it's often unclear whether what is returned will really be what they want. There are many reasons for this. For example, a language may have several different names (Welsh is 'Cymraeg' to speakers of the language), and the search will not bring up any resources that are classified by a name different to that given to the search engine. The opposite problem also exists: two different and unrelated languages may have the same name (for example, there is a Turkic language called Ainu spoken in Central Asia and another, entirely unrelated, Ainu spoken in northern Japan and the Kuril Islands); this would bring up irrelevant (i.e. too many) search results.

The standard way of dealing with this problem has been to assign to a language, and all its variant names, a single language code. A search engine (or rather, the database it searches) converts the language into the code and searches on the code instead.

The existing standards, ISO 639-1 and ISO 639-2, however, have a far too limited number of codes: somewhere in the order of 500. This is nowhere close to being sufficient for the purposes of the linguistic world, who need to distinguish around 7200 mutually unintelligible languages. The shortcomings of ISO 639 have to be accommodated by researchers somehow, either by using the nearest (but still incorrect) code and supplementing it with extra description, or by ignoring the ISO 639 standard altogether.

There is another shortcoming of the ISO system. There are no language family codes. There are indeed ISO codes which reference large groupings, but they were intended as grab-bags for minor languages which ISO did not want to assign codes to (e.g. the ISO 639-2 code NIC used to reference any Niger-Congo language not referenced with a more precise code). These codes then do not actually reference subgroups at all, but are rather what might be called "diffuse language codes".

3. E-MELD AND SIL/ETHNOLOGUE LANGUAGE CODES

The solution we have chosen is to base our language code system not on that of the ISO, but on that of the Ethnologue, produced by the SIL. The Ethnologue codes are an essentially complete classification of the currently used languages of the world. Its only shortfall is that it does not include codes for ancient languages which are no longer in use, nor for languages which have been extinct for a considerable time. It also includes no codes for constructed languages. Our first decision, then, was to supplement Ethnologue with codes, to be created by LINGUIST, for all attested ancient languages and all constructed languages.

The principles by which these codes are assigned, as well as links to the actual codes themselves, can be found at the URL:

http://www.language-archives.org/wg/language/

4. THE PROBLEM OF CLASSIFYING A LANGUAGE BY FAMILY

It is not sufficient to merely provide codes for languages. Linguists often work on subgroups of languages, and language families. Data must be classifiable in terms of these groupings, and retrievable by means of them. It must thus be possible to access linguistic data linked to the individual languages that make up a grouping, as well as data linked to the grouping itself. For example, a user searching under the term 'Indo-European' will want to know about resources directly linked to that term (a discussion of putative IE dialects, for example), what resources are linked to sub-families of IE (e.g. East Germanic), and what resources are linked to languages which are part of the Indo-European family (e.g. English, Russian, etc.). This means that (1) groupings must exist as entities in the database to which data can be linked, and (2) languages must "know" what subgroup they belong to.

These requirements constitute a considerable problem in themselves, for as new data are discovered and described, researchers will want to modify existing language family classification in accordance with their findings. The older classifications, however, will still have data linked to them, and must remain searchable. Therefore, more than one classification must be representable in the database, and there must be a facility to add new ones. Furthermore, different classifications are not all equal. For example, Greenberg's (1985) classification of American languages into just three families, Na-Dene, Eskimo-Aleut and Amerind, is highly controversial. Deprecated or unproven classifications must thus be marked as such. Classifications must also be given a provenance so that researchers can come to their own conclusions about a particular classification.

These needs constitute something of a quandary. It seems reasonable to use a fixed set of codes for languages, since though there may be differences between linguists on whether a particular variety is distinct enough to be called a language, the problems remain relatively minor. But subgrouping is so much more fluid and various that there is a serious question as to whether a coding system is the best solution. Yet at the same time to allow linguists to use the names of the subgroups is a recipe for chaos. Because of the varied views of different subgroups held by different linguists, a subgroup can really only be accurately defined in terms of a tree, or a code that in some way references a tree. The same subgroup, after all, can appear in very different trees.

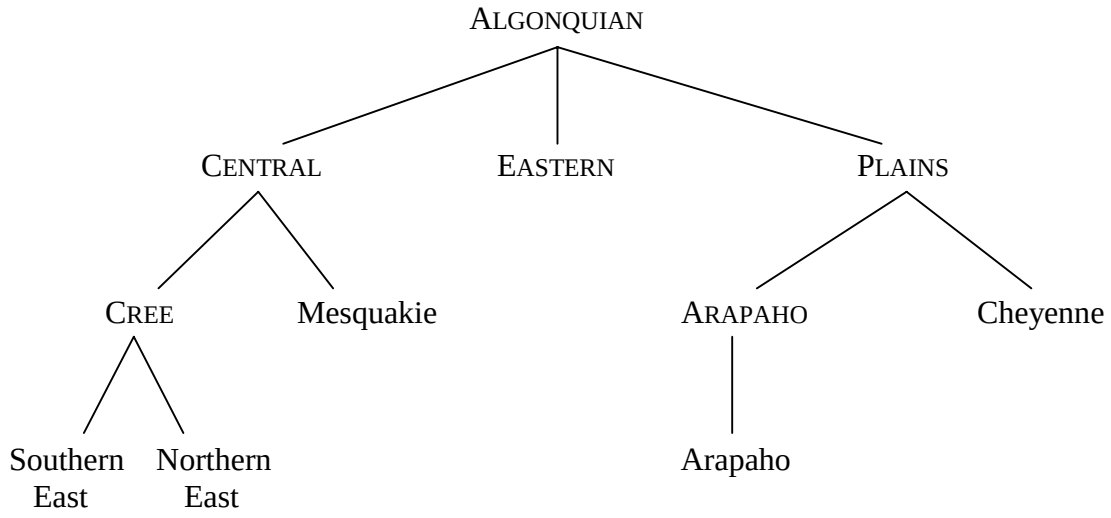
5. A PROPOSAL

The solution we have chosen in EMELD is twofold. First, we assign codes to nodes of trees, thus allowing them to be referenced from outside our database. These codes define unique trees. In order to ensure that they provide constant references to the right subgroup, they will never be retired, since even when a subgroup is no longer accepted by the community, the code

will still reference it accurately, and the languages which it included. Second, we allow both languages and subgroups to be members of more than one tree. There is no single tree for Indo-European, for example. Instead, there are multiple trees, each of which is annotated as to its acceptability. It is thus possible to present the tree of Indo-European which is most commonly accepted by linguists. But the less accepted subgroupings are there too, and retrievable as parts of other trees. This also means that languages — and subgroups — can have more than one subgrouping code, each of which defines a different tree.

This we have implemented in a rather simple relational database design, which can easily handle the requirements of multiple classifications. In the design, languages are considered to be conceptually different from language families or subgroups (which are merely labels for groups of languages) and so they are in separate tables. Therefore, Indo-European would be classed as a family, but Proto-Indo-European, a language. A language is related to one or more immediately dominating subgroup(s) and these relationships are recorded in a third table. A subgroup, in turn, is related to its immediately dominating node and this is recorded in a fourth table. In such a way, a family tree can be generated from the tables. It is easy to record provenance in these tables as just an extra field in the tables. A demonstration is below, with this simplified tree of some North American languages:

(1) A simplified tree of some North American languages:



Here, families are in upper case, and languages in lower case. This tree can be generated by the database structure, below:

(2) A corresponding database structure

Table: Language

LangID	LangName	Descrip	Prov
101	Sth. East. Cree		
102	Nth. East. Cree		
103	Mesquakie		
104	Arapaho		
105	Cheyenne		

Table: Lang_Family

LangID	FamID	Prov
101	1005	
102	1005	
103	1002	
104	1006	
105	1004	

Table: Fam_Fam

FamID	ParentFamID	Prov
1001	~~NULL~~	
1002	1001	
1003	1001	
1004	~~1001~~	
1005	1002	
1006	1004	

Table: Family

FamID	FamName	Prov
1001	Algonquian	
1002	Central	
1003	Eastern	
1004	Plains	
1005	Cree	
1006	Arapaho	

The Lang_Fam and Fam_Fam tables are the linking tables where the hierarchical information is stored.

It is a very simple process to add a newly discovered language; it only requires a new entry in the Language table and a family assigned to it in the Lang_Fam table.

Similarly, it is very simple to add an alternative analysis. Amerind can simply be added as a family to the Family table (with a FamilyID of 1007, say) and a new entry is added to the Fam_Fam table, associating Algonquian (FamID 1001) with having a ParentFamID of 1007. In such a way, Algonquian has two entries in this table representing both theories. There is no need to recreate in the database the part of the tree downwards from Algonquian. A reference to Greenberg's work would be added in the Provenance field to both the Family table and the Fam_Fam linking table, as a reference for the existence of the Amerind family and what the daughters of Amerind are.

It is similarly straightforward to define a new tree which lacks a subgroup. If a researcher wants to propose an analysis where there is no Arapaho family, this is equivalent to redefining the parent family, Plains Algonquian, so that it just has the two languages of Arapaho and Cheyenne as daughters. Therefore, a new Plains Algonquian is added to the Family table, with the same name but with a different FamID (say 1008, now). In the Lang_Fam linking table, then, the languages Arapaho and Cheyenne will each be given an extra association, that with Plains 1008.

Reassignment of one language to a different family will work in precisely the same way, with new mother nodes being created.

The structure proposed is very flexible and will allow more than just a genetic classification. Areal classifications can be incorporated very easily without any new tables. 'North American', for example, can be added as an entry to the Family table and it can be related to its daughter languages or language families in the linking tables as appropriate. Not shown in (2) is a required 'Type' field in the Family table, by which means would be differentiated the areal classifications from the genetic.

Also not shown in (2) is the 'Default' field in both the Lang_Fam and Fam_Fam linking tables. This would separate the 'default', or most usual classification from all the others. As will be explained in section 6, doing this is important so as to not swamp the user with too much information.

For the sake of exposition, alternative language names have not been described here. In truth, they are a simple matter, requiring no more than an extra table with the alternative names listed, and linked to the Language table. Treatment of alternative family names would work in the same manner.

We are now in process of putting together working groups of linguists, as part of the OLAC process, who will decide what subgrouping information is included in these tables. It is important to us that the linguistic community provide as much input as possible.

6. A USER INTERFACE

We are now making studied attempts to make the material in this database easily accessible to users. From a user's point of view, the amount of information that can potentially be stored in a database such as we have described is overwhelming. This is especially true in the case of the ordinary user doing preliminary research, who might just want to know what books have been written on the language Blackfoot, for example, and who does not care how that language is classified.

For those who are interested in the relationship a language has with others — someone researching cognates of borrowed words, for example — being presented with all possible trees at once is also daunting. The number of possible trees such a database can define is, in any case, potentially so large that it would not be computationally feasible to display all of them at once.

This is where default relationships are important; it is the default that will be displayed automatically. Other relationships can be hyperlinked from the page with the default tree, and the provenance information will be very useful in helping the user to decide which to click on.

Most of this has already been implemented on the LINGUIST and EMELD sites, and is available through the search facilities we have put in place there.

Our intention is that researchers who want to add a new language will submit information through a web form. They will be asked to fit the language somewhere in the genetic and areal trees. However, in doing so, they may prefer to propose a new genetic relationship between languages. In many cases, it might be simple change (as far as the database is concerned) and be a matter of reassigning some languages to a new or different node. For this, the user will be presented with a javascript interface showing the default tree, which will allow a simple point and click changes.

7. CONCLUSION

The language classification system we have described here is very much in the spirit of current approaches to language archiving. We have proposed what is very much a living system, one which is to a large extent user-maintained, and which is able to grow to accommodate the needs of the whole linguistic community. However, one thing is very clear: a system such as we have described here is only feasible if accessed through a central server through which codes can be accessed, with its data controlled by linguists acceptable as arbiters to the community. Just as

Ethnologue provides a central site for the assigning of language codes, so a central subgrouping server is also necessary. Where this server should be placed, we will leave up to the community of linguists to decide.

REFERENCES

Bird, Steven and Gary Simons. 2002. Seven Dimensions of Portability for Language Documentation and Description. http://ldc.upenn.edu/…

Greenberg, Joseph. 1985. Language in the Americas. Stanford: Stanford U.P.

Summer Institute of Linguistics. Ethnologue. http://www.ethnologue.com