Some Higher-Level Design Features of SSWEL
(Syntactic Structures of the World's Languages).
Chris Collins, NYU

In this short talk, I will add some detail to the proposal that Richie and I sent out earlier. These remarks are the results of discussions between Richie and I over the last few weeks. My goal is to approach things from the top-down, outlining the highest level design features of the database. At the very end of the paper, I sketch some ideas on the steps of implementation that might help to clarify higher level design features. Some of the following points address issues that came up in the commentary, but I have made no attempt to be systematic.

1. Data

The database will be used to do comparative syntax, in the broadest sense. The database will not contain information on phonetics or phonology. The basic data will be (a) property-value pairs, and (b) sentences that are glossed and translated into English, (c) references to other work. A toy example of a property value pair is:

(1) Property: Adpositions
Values: Prepositions,
Postpositions

This property says that there is a syntactic category of "adposition" and a language either has prepositions as in English ("in the house") or

postpositions as in Japanese. Given the existence of these properties, queries such as the following should be possible:

(2) a. Find all languages that have postpositions and serial verb constructions.

b. Find all languages that have postpositions which have agreement suffixes.

WALS ("Word Atlas of Language Structures") has shown the usefulness of an interactive database containing property-value pairs encoding grammatical information.

For each property-value pair characterizing a specific language, there will be sentences from that language that illustrate the property. There are many reasons why the database should contain glossed sentences in addition to property-value pairs. First, the sentences will allow a user to verify claims made in the property-value pairs, thus increasing the reliability of the database. Second, these sentences, and their glosses, will also be searchable. A good model for this part of the proposal is ODIN (Online Database of Interlinear Text), where searches of glosses are possible. Third, the presence of these glossed sentences will allow our database to serve an archival function, in particular with regard to less accessible and endangered languages.

Given the presence of glossed sentences, queries such as the following will be possible:

(3) a. Find all sentences from a romance language that contain “all” in the gloss.

b. Find all West African languages for which there is a sentence that contains PASS (“passive”) in the gloss.

Searches involving combinations of property-value pairs, and elements of the glossed sentences should also be possible.

Lastly, an important kind of data is a set of references and links to other work (including papers, grammars, personal web pages). Some examples of other systems that we could directly link up to include Ethnologue, the OLAC Catalogue (“Open Language Archives Community”), and ODIN (Online Database of Interlinear Text). Given the ease with which it is possible to link to other systems (see section 5 “Interoperability”), there is no need for our database to be a repository of all information about all languages.

An important question, which this workshop should address, is what other kinds of information are absolutely necessary for us to include in the database.

2. Open-Endedness

Perhaps the most fundamental feature of the database, which clearly distinguishes it from other similar projects, is that it will be completely open-ended, in the sense that data can be added at anytime by anybody (see the section 3 “Users”). The kinds of data that can be added will include the following:

(4) a. Values of existing properties for particular languages
(where the languages can be ones that already exist in the database or completely new ones)

b. Glossed example sentences

c. References and links to outside work

d. New properties

e. Commentary on all of the above (see section 4 “Forum for Interaction”)

Since the database will be open-ended, and it will be possible to add limitless amounts of data to it, the issue of reliability comes up sharply. Suppose I enter some data, in the form of a set of property-value pairs and example sentences for Ewe. How is a database user to know how reliable this data is? Part of the answer to this question is that all the data I enter will be clearly tagged as having been entered by me (Chris Collins), on November 9, 2007. I will also indicate that I obtained the data during fieldwork in Togo in Agbanon with an informant on such and such a date.

A core part of the open-endedness of the database will the ability to add new properties. To take a concrete example, in working on Khoisan languages such as N|uu, =Hoan and Ju|’hoansi, I have investigated a morpheme (dubbed the “linker”) that appears preceding various post-verbal constituents. I should be able to formulate properties describing the linker, and set the values of those properties for the Khoisan languages. Subsequently, the new properties will become visible to all users of the database, who can then add information for other languages (if relevant). The ability to add new properties to the database raises the question of how a user will know whether their property duplicates an existing property.

Another issue is whether it will be possible for the created properties to have a uniform format, so that they employ the same terminology, concepts and definitions of the pre-existing properties and the same definitional style. For example, if I label one of my properties “Presence of Linker” how can I be sure that I am using the term “linker” in a way that is consistent with other uses in the database. The notion of an “ontology” should help to resolve this issue.

An important principle of design is that data should be frozen once added (“Freezing Principle”). For example, if person X adds some sentences on logophoric pronouns in Ewe, it should be impossible for person Y to delete these sentences. Furthermore, if person X then returns to his/her old data, and decides that there were some problems with it, even person X would not be able to delete the data. Rather, person X could enter a new and improved version of the data (which would co-exist with the original version). Since the data will be cumulative, all data must be clearly marked as to when it was entered.

The open-ended and fundamentally dynamic nature of the database will make it impossible for the data entered to be refereed in any standard sense. Therefore, if somebody adds data to the database, this data will probably not count toward tenure and promotion decisions in universities. Furthermore, it is important to emphasize that the database will have no final publishable form. Rather, it will simply keep growing as long as there is somebody to maintain it.

However, since all the data will be carefully tagged for who entered it, and what the ultimate source is, it should be possible to respect the intellectual property rights of the users.

3. Users

Every person on earth will be able to access the database on the internet to do searches. There will be no constraints on who can type in the URL, and start accessing data.

Furthermore, after a process of registration, users will be able to add data. There will be no constraints on who can register to add data. One goal of the workshop should be to articulate this process of registration.

An important question is who the users will be. We think it is important to aim at the broadest possible audience in order to increase the total amount of data in the database. Therefore, we propose that the database be designed so that it can be used by “amateurs”. By an amateur, we mean somebody who has a serious interest in language, but who does not have a Ph.D. in linguistics or a related field. The decision to aim the database at amateurs has many specific consequences. Mostly importantly, it forces us to think about how to design the data-entry interface to be as simple as possible. One possibility is that the interface would be designed in a way similar to “tax-preparation software”. For a particular domain, there would be a set of explicit questions that are relatively easy to answer.

Similarly, the concept of “questionnaire” could be useful. For example, in work on Northern Italian dialects, there could be a questionnaire in the form of a series of sentences in standard Italian that the amateur could translate, on the model of the questionnaire used by 'Syntactic Atlas of Northern Italy'. These sentences would automatically go into the database.

4. Forum for Interaction

We have already seen two mechanisms that will help to ensure reliability. First, users who want to enter data will be registered. Second, every piece of data in the database will be tagged for certain information (who entered it, at what time, what was the source, etc.).

Since the database is open-ended, another way to ensure quality is to allow users to comment on each other's data. Consider the property of "Order of Adpositions" in Ewe. Suppose that I set the value of the property to "preposition", and another person disagrees with the claim that Ewe has prepositions. Then they could easily register this disagreement, and explain why they disagree. Such disagreements would be immediately accessible to anybody looking at the "Order of Adpositions" property for Ewe.

Similarly, it should be possible for other people to indicate whether they agree or disagree with the sentences that I have added. If enough people register their agreement with the data that I add, a user should be relatively comfortable in using it in their own research. If I disagree with a sentence, or a property-value pair, I could easily add a whole paradigm of sentences to prove my point.

The notion that the database is a forum for interaction should pervade all aspects of design, and all types of information. This will have the consequence that users of the database must be comfortable with disagreements and with the dynamic ever changing nature of the information in the database.

5. Interoperability

The database will be designed to be maximally interoperable with other databases and projects that exist on the internet. This feature will allow users to import data into (and export data out of) the database in an efficient manner. It will also allow the database to link with other projects efficiently (see the end of section 1 above). For all the data in the database, we will adhere to standards in the field. One example (suggested by a number of participants) is to use the ISO 639-3 codes for languages. Furthermore, we will attempt to follow the recommendations outlined in Bird and Simons (2003) and the EMELD web page ("Electronic Metastructure for Endangered Languages Data") to the greatest extent possible. Some examples of these recommendations include:

(5) a. "Encode characters with Unicode"
(Bird and Simons 2003: 575)

b. "Prefer XML … over other schemes of descriptive markup"
(Bird and Simons 2003: 575)

c. "Follow OLAC (Open Language Archives Community) recommendations on best practice for describing language resources" (Bird and Simons 2003: 576).
(http://www.language-archives.org/)

d. "Map terminology and abbreviations used in descriptions to a common ontology of linguistic terms" (Bird and Simmons 2003: 574).

Adopting these standards leads to many questions. For example, the database will allow arbitrarily many dialects of a given language to be described. This raises the question of how the name of dialect for which there is no pre-existing language code will be entered into the database.

One question which the workshop should address is what kinds of standards are out there that we can incorporate to guarantee interoperability of our database with other projects.

6. Role of Linguistic Theories

Every generalization made about language, and every gloss that is given to a sentence is a theoretical statement. So in some sense, the database could never be “theory neutral” or “theory free”. However, the database will be “neutral between theories” in that it will not be oriented specifically toward any existing theoretical framework, e.g., Minimalism, HPSG, Arc-Pair Grammar, LFG, etc.

Everybody will be able to search the database. Furthermore, people who have registered will be able to enter data (sentences, references, properties, property values, comments, etc.). Nobody will be prohibited from entering data or creating properties on the basis of their linguistic framework.

The property of neutrality has the direct consequence that a particular phenomenon might be classified in several different ways on the basis of different theoretical orientations.

7. Order of Implementation

It may help to conceptualize the implementation of the database in three steps, each of which brings up its own problems. Considering the problems brought up in this hypothetical series of steps might clarify some of the higher level design features.

The first step would be to select a fairly large set of languages (e.g., 30), and a fairly large set of properties (e.g., 100), and to create an on-line database that includes data on these languages and these properties. Volunteers would be needed to complete this step. This preliminary version of the database would be open to everybody to use (to do searches), but it would not be possible for everybody to add data (it would not be fully open-ended).

The second step would be to make the database open-ended in the sense that people could set property values for new languages that are not already in the database. However, the database would not be fully open-ended, since it would not be possible to create new properties. At this step, it should also be possible for people to add glossed data, and to comment on data.

The last step would be to open the database up completely so that people could add new properties. This last step brings up issues such as how the properties are organized internal to the database, and the consistency of new properties with properties that already exist, etc.

References

Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. Language 79.3, pgs. 557-582.

Haspelmath, Martin, Matthew S. Dryer, David Gil, Bernard Comrie. 2005. The World Atlas of Language Structures. Oxford: Oxford University Press.

ODIN (Online Database of Interlinear Text)
http://www.csufresno.edu/odin/

Ethnologue
http://www.ethnologue.com/

OLAC (Open Language Archives Community)
http://www.language-archives.org/

Syntactic Atlas of Northern Italy
http://asis-cnr.unipd.it/db.en.html

E-MELD (Electronic Metastructure for Endangered Languages Data)
http://www.emeld.org/index.cfm