Syntactic Structures of the World's Languages

The purpose of the NYU workshop is to investigate the feasibility of creating a database of the syntactic structures of the world's languages. The main purpose of the database will be to provide a tool for syntacticians, morphologists and semanticists doing cross-linguistic work which will allow them to explore the connections between various properties of the world's languages.

1. Description of Database

Linguists are working toward understanding what all human languages have in common and, simultaneously, toward understanding the ways in which human languages differ from one another, and what the limits on those differences are (see Chomsky 1981, Greenberg 1966).

The database will focus on those aspects of human language that fall under the rubric of syntax (grammar). It will not include the subpart of linguistics called phonology that studies the sound systems of human languages. There will be substantial ties to questions of morphology (having to do with the structure of words) and to questions of semantics.

In doing their work, syntacticians take into account data about the properties of many individual languages. The number of languages taken into account has been increasing substantially (see Baker 1996, Julien 2002, Kayne 1994, Cinque 1999, Dryer 1992, Haspelmath et. al. 2005). So much so, in fact, that it has become increasingly difficult to keep track of them, to integrate the data, the descriptions, the theoretical implications that this ever larger number of languages is feeding into the field. Technological advances have helped. The use of computers allows searches to be done far more quickly than in the past. At the same time, the field has not yet made significant use of the internet, or at least not to the extent that it should. The aim of this project is to develop a readily usable web-based database that will allow researchers access to the properties of a far greater number of languages than would otherwise be possible.

Simply using the web is not sufficient, though. The kind of database we have in mind would take inspiration from open-ended systems such as Wikipedia. It would be constructed in such a way as to allow linguists from anywhere in the world to add new languages to it, or to add new data or new generalizations concerning some language already in it. The number of languages in the proposed database would be constantly increasing, as more and more languages from around the world are added. Some of these languages would be relatively well-known languages that have not previously received much attention. Others would be lesser-known languages and endangered languages that linguists from a new generation would have found the means to study in detail. Still others would be what are often called dialects, but deserve to be studied as separate languages, often with interesting and important syntactic differences relative to their better-known cousins. For example, in addition to information on Standard American English, there would be information on AAVE (African American English), and Appalachian English, as well as many others.

Since dialects can often profitably be divided into (syntactically distinct) subdialects, it is clear that by having the database open to new dialect distinctions, as well to the entry of previously little-studied languages from all over the world, the number of

languages/dialects that the database will contain will ultimately be orders of magnitude greater than the number 6000-7000 (see Ethnologue 2005) often cited as the number of languages currently spoken.

The database we have in mind will also aim to take into account a far greater number of syntactic properties than has ever been done before. In part, this will simply reflect the knowledge already accumulated by syntacticians, especially over the past 50 years. In part, it will reflect the open-source character of the database. Although we plan to start the database with a given set properties, we very explicitly intend to allow for the addition to the database of new properties discovered in the future (or currently known to some, but overlooked in the original set).

Just as the set of languages to be incorporated in the database will be finer-grained (by virtue of including large numbers of dialects) than in any previous work, so will be the set of syntactic properties. One way in which our understanding of syntax has progressed over the decades is in paying ever greater attention to what might in earlier stages of the field have been called very small differences across languages, which have often turned out to be of considerable importance to the development of an adequate theory of syntax. For example, it has long been understood that languages differ with respect to the relative order of adpositions and their objects. Adpositions in English (e.g. 'to', 'at', 'by', 'with', 'of') are called prepositions because they typically precede their object ('to the city', not '*the city to'). Comparable elements in Japanese are called postpositions because they follow their object, reversing the English order. It has also been known for a long time (see Greenberg 1966) that whether a language has prepositions or postpositions correlates with the relative order of verb and object. Languages that exclusively have postpositions invariably have the verb following its objects, in the general case.

A 'smaller' property having to do with adpositions involves agreement. In some languages adpositions agree with their object, in some languages they don't. In building up a sense of which languages fall into which group, syntacticians have discovered that languages whose adpositions agree with their object never have subject-verb-object word order (but only subject- object-verb or verb-subject-object order). A still smaller property, one that has hardly been studied at all yet, but which our database will include, and will stimulate and facilitate the study of, concerns what could be called the morphology of adpositional agreement. In some languages, the morpheme that corresponds to agreement (in person and/or number and/or gender) follows the adposition in question, whereas in others the agreement morpheme precedes the adposition. Whether this cross-linguistic difference correlates with others, and why, is something that having a database such as the one we envision will make it possible to investigate.

Other properties in the database will have to do with various other aspects of syntax: passives, causatives, reciprocal suffixes, ellipsis (including sluicing, gapping, pseudo-gapping, etc.), case systems (ergative, absolutive, split), the presence or absence of certain grammatically important lexical items (the word “have”), strategies for question formation (wh-movement versus wh-in-situ), properties of relative clauses (head internal versus head external, relative pronouns, pied-piping), morphological properties of noun phrases (noun class prefixes/suffixes, gender prefixes/suffixes, plurality), referential properties of quantifiers and noun phrases (the presence of “every”, “each”, “no”, definites, indefinites, question words, etc.), morphological features of pronouns

(singular, plural, dual, inclusive, exclusive, masculine, feminine, etc.), referential properties of pronouns (e.g., possibility of bound variable anaphora, weak crossover effects), strategies of negation (double negation, negative concord, negative polarity items), lexical category information (nouns, verbs, adjectives, prepositions), and argument structure (double object verbs, various locative alternations), etc.

As the field of syntax continues to expand, other properties will be thought of that are of interest and importance. Our database will be constructed so as to allow them to be added, without limit.

The only project that is directly related to our proposed database is the "The World Atlas of Language Structures" written by Haspelmath et. al. (2005) (abbreviated WALS). WALS allows users to search a database of properties on a CD, and to correlate those properties. For example, it is possible to search for the languages that have the basic word order SOV, and to see how that set of languages corresponds to the set of postpositional languages (where the adposition follows the noun phrase).

However, our project differs greatly from WALS. The primary difference is that we foresee the internet database to be completely open, such that linguistic researchers can continually add new information. WALS is closed in the sense that new information can only be added by the authors of the system. This property of being open-ended will mean that the amount of information available on the internet database will be astronomically larger than what is given in WALS. The kind of information that researchers will be able to add will be of two kinds: First, it will be possible to describe new languages in terms of the properties already in the database. Second, it will be possible add properties.

There will be many smaller differences between our database and WALS as well. For example, every property for every language will be exemplified with a number of example sentences. As a consequence, our database will contain detailed grammatical descriptions of each language. By contrast, WALS has very little actual linguistic data in it (only a very few properties are actually exemplified).

2. Description of Workshop

The workshop will call together a group of scholars who have expertise related to the project. Some of the questions that will be discussed at the workshop include the following:

Linguistic Considerations:

What properties should be on the initial list of properties in the database?
What sorts of research questions would people use the database to investigate?
How is it possible to compare languages that are not related at all, or not closely related, and that are quite different syntactically? What does it mean to say that morpheme X in one language is the same as morpheme Y in another language?

Open-Endedness:

What kinds of mechanisms can be put in place to ensure high quality data? How will new data be tagged so as to increase its reliability (author, source, etc.)? How will users register (especially data providers)? What happens in case of conflict? For example, suppose two experts on language X disagree on the facts concerning adjectival agreement, or quantifier interpretation, how will these differences be registered and/or resolved. What is the best way to manage the addition of new properties? Should anybody be able to add a new property? Will there be some kind of regulatory system in operation (e.g., editors, discussion groups, rotating committees, etc.)?

Implementation:

What kinds of computer software, and hardware will be needed to implement such a project? What kinds of skills will the programmer who creates the system need? What precise steps will be needed to create the database? How long will it take to put together? Are there other projects similar to our own on the internet right from which we could learn lessons about how things should or should not be done? What kind of standards are out there for the representation of linguistic data on the internet?

References

Baker, Mark. 1996. The Polysynthesis Parameter. Oxford University Press, New York.

Chomsky, Noam. 1982. Lectures on Government and Binding. Dordrecht: Foris Publications.

Cinque, Guglielmo. 1999. Adverbs and Functional Heads. Oxford University Press, Oxford.

Dryer, Matthew S. 1992 "The Greenbergian Word Order Correlations." *Language* 68: 81-138.

Gordon, Raymond G. 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International.

Greenberg, Joseph H. 1966. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg (ed.), Universals of Language, pp. 73-113. MIT Press, Cambridge, MA.

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

Julien, Marit. 2002. Syntactic Heads and Word Formation. Oxford University Press, Oxford.

Kayne, Richard. 1994. The Antisymmetry of Syntax. MIT Press, Cambridg