A Proposal for a Database of the Syntactic Structures of the World's Languages
Chris Collins and Richard Kayne, NYU

Abstract: On November 9-10, 2007, a conference on creating a database of the syntactic structures of the world's languages was held at NYU. This document contains the original proposal for the database (Chris Collins and Richard Kayne), the paper presented by Chris Collins and the paper presented by Richard Kayne. The program of the conference can be found at: http://linguistics.as.nyu.edu/object/linguistics.events.database_workshop

Keywords: adposition, agreement, comparative syntax, database, dialect, glossing, primitives, questionnaire, replicability, Wikipedia

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

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

Some Requirements for the Database

Richard S. Kayne
New York University

Workshop on the Feasibility of a Web-based Database
of the Syntactic Structures of the World's Languages
November 9-10, 2007

Like other scientists, syntacticians (of all stripes) are in constant need of more data (sometimes without knowing it). This constant need for more data parallels a constant need (again, sometimes without knowing it) for more thinking (sometimes called theorizing).

Data comes in many different colors. One bedrock type (thinking of Chomsky's 'observational adequacy') is the acceptability status of particular sentences in particular languages. (Distinguishing among possible interpretations is sometimes essential.) Another type (thinking of Chomsky's 'descriptive adequacy') involves generalizations of various sorts.

These generalizations can be made relative to just one language. For example: English particles come after the verb (*They picked it up* vs. **They uppicked it*). Even for one language it is challenging to formulate things in exactly the right way (thinking of *They uploaded it*).

Comparative syntax (of various stripes) is a subpart of syntax that focusses on questions of parametric variation (whatever the ultimate form of parameters turns out to be). These questions interact constantly and fruitfully with questions about what in the language faculty is not subject to variation.

In comparative syntax, observational adequacy involves discovering syntactic differences. For example, English *They see us often* is unacceptable in French. This kind of statement is of course shorthand for something like: The English sentence *They see us often*, if transposed into French by substituting for each English morpheme the corresponding French one (without altering order), yields an unacceptable French sentence (**Ils voient nous souvent*). A related and equally correct observation is that French *Ils nous voient souvent* is unacceptable in English (**They us see often*).

In trying to pin down the parameter (or parameters) that underlies this English-French difference, one is led to formulate generalizations heading in various directions. For example: In Romance languages, direct object pronouns in simple sentences (with a finite verb) must precede the verb. This amounts to saying that the fact noted for French is part of something broader. Actually at least two somethings: The initial observation for French was stated for the pronouns *nous (=us)*, but it holds for all simple pronouns. Furthermore it holds for all Romance languages. (Well, it almost does, thinking of the Borgomanerese dialect spoken a bit west of Milan.)

Some comparative syntax generalizations are of the 'linked' sort, as in the work of Joseph Greenberg. One such example (that seems to be correct) is: If a language has canonical verb-object order, then its declarative complementizer (if it has one) is never sentence-final (cf. Dryer 1992; for a proposal on why, cf. Kayne (2000, chap. 15)).

These linked generalizations are also found in finer-grained cases. One that seems to be correct is: If in a Romance language the counterpart of **They don't know if to leave* is acceptable (as it is in Italian - *Loro non sanno se partire*), then that language has its object clitic pronouns following the infinitive (rather than preceding the infinitive). (For a proposal on why this should hold and on how it tells us something important about the syntactic status of the unpronounced subjects of infinitives, cf. Kayne (1991).)

The additional data that syntacticians are in constant need of falls into all the above subtypes, ranging from the acceptability status of individual sentences to generalizations of various degrees of abstraction, both internal to one language and across two or many or all languages.

The database that Chris Collins and I have in mind is intended to take advantage of current (web-based) technology to meet, to a much greater extent than ever before possible, this need that syntacticians have and will always have.

To maximize the advantage to be had, we would like the database to be maximally open (as Chris will be discussing in more detail in his presentation).

The number of languages that syntacticians (as a whole) have taken into account in their work is extremely impressive. The total amount of data taken into account, over all these languages, is equally impressive. To get a feel for the amount of data in question, you can take a look at the descriptive grammar of English (thoroughly informed by generative syntax work) edited by Huddleston and Pullum (2002), about 2000 larger-than-average pages, densely printed and densely written. A comparable grammar of Spanish (over 5000 pages equally dense pages) has been edited by Bosque and Demonte (1999).

One needs to keep in mind, of course, the fact that these grammars are very, very, very far from being exhaustive (the tip of an unbounded iceberg, in effect). The French linguist Maurice Gross (1975, 18) once estimated the number of well-formed French sentences of 20 words or less to be on the order of 10^{50}. In this light, although the amount of data/knowledge accumulated by syntacticians over the years is extremely impressive (and underestimated by many non-syntacticians), the amount of data that syntacticians have not yet taken into account is even more impressive, orders of magnitude more impressive, and always will be.

Similarly, although the number of languages taken into account by syntacticians over the years is extremely impressive, the number of languages that have not been taken into account is even more impressive, orders of magnitude more impressive, I would say. This is fairly clear just by thinking of all the languages that were spoken in the past and which we have no access to (and similarly for the future). Even for the present, there is reason to think that the number of distinct languages/grammars is at least as great as the number of (non-infant) human beings currently alive (Kayne (2000, chap. 1)). In addition, on a reasonable calculation based on the likely number of (binary-valued, independent) parameters, it is perfectly plausible (I think) that the number of possible human languages is (at least) on the order of 10^{30}.

As in other sciences, the fact that we will never get to the end of things, that we will never be able to study all languages, that we will never be able to know everything

about even one language, is perfectly compatible with the idea that we can (and therefore must try to) make substantial progress in understanding the language faculty.

Our database will aim to accumulate as much data, as many generalizations, over as many languages as possible. In so doing, it will increase the likelihood of our coming up with a correct theory/understanding of the language faculty.

These data and generalizations must be as accurate, as solid as possible. Replicability is of fundamental importance.

It is hoped that the wikipedia-like character of the database will enhance the replicability of all the data and generalizations that find their way into the database.

This will be particularly important in the case of lesser-studied languages. For languages like English, Japanese, Mandarin, etc. the number of syntacticians (and others) contributing data/generalizations is large enough (and will get even larger) to ensure a high degree of confidence in a huge number of cases. (The Huddleston & Pullum grammar I mentioned is extremely solid, as far as I can see.)

In cases where for a particular language there is disagreement on certain facts, the database will be designed to facilitate figuring out whether what's at issue is a(n irreducible) dialect difference or something else (sometimes the problem is that the initial question was poorly formulated).

Maximizing replicability for acceptability judgments in a particular language is a relatively simpler task than doing the same for cross-linguistic generalizations. In part, this is because comparative syntax is intrinsically more difficult than work on one language (which it subsumes).

In part it is because such cross-linguistic generalizations depend (even more than work on a single language, in all likelihood) on a proper understanding of what the primitives of syntax are. Comparing English and French (or Germanic languages and Romance languages) seems relatively straightforward, if only because it is/seems easy to find in one language the morpheme that corresponds to a specific morpheme in the next language. (In fact the task is more challenging than it looks, even for such relatively similar languages.) If one wants to bring Japanese or Mandarin or both, etc., into the comparison, the problem of ascertaining morpheme correspondences is more difficult. (No matter what the degree of language difference, the question of silent morphemes enhances the challenge.)

Ascertaining morpheme correspondences plays a key role in the apparently banal task of glossing sentences, which will be basic to the database. The non-trivial character of morpheme correspondences (cf. Kayne (2005) on the question whether French *peu* corresponds to English *little* or to English *bit*, and on the probable absence in French of any (overt) correspondent to English *every*) means, I think, that even glossing must necessarily be taken to be a theoretical enterprise, with particular choices of glosses being subject in effect to future disconfirmation. The database will have to allow for that.

Formulating generalizations across more than one language (ranging from two languages to all possible languages), in addition to depending on getting morpheme correspondences right, depends on using appropriate categories. For example, the Greenbergian generalization mentioned above involves the notion 'declarative

complementizer', yet there's some reason to think that the language faculty contains no such primitive category (i.e. English *that* is (still) a demonstrative, etc. - cf. Kayne (2007)). A correct account of such generalizations is (much) more likely to be found if the notions that make up the generalization are the right ones. The database will have to allow for disagreements about what the proper primitives are, not in the sense that everybody has the right to their own primitives, of course, but with the understanding (on my part, at least) that today's disagreements must ultimately give way to (relative) winners and losers.

A parallel point to complementizers can be made about the category of adpositions (prepositions + postpositions), which, quite apart from the word order facts, is hardly likely to constitute a unified category. The Greenbergian generalization (cf. Dryer (1992)) that if a language is postpositional, then it is OV seems to admit a number of counterexamples, on the standard view of what a postposition is. But if adpositions actually break down into nouns (of a certain sort) and non-nouns, and if one reinterprets this Greenbergian generalization as saying that if a language is postpositional with respect to its non-noun 'adpositions' (cf. Kayne (1998)), then it is OV, then the counterexamples (seem to) disappear.

The set of possible syntactic properties and generalizations over them that one can work with is open-ended (though the set of primitive parameters is probably not). Some (but not all) of these will provide, via theoretical work, an important window onto the language faculty. Our database aims to increase the probability of reaching results of lasting significance.

References:

Bosque, I. and V. Demonte (1999) *Gramática Descriptiva de la Lengua Española* (three volumes), Espasa, Madrid.

Dryer, M. (1992) "The Greenbergian Word Order Correlations," *Language* 68, 81-138.

Gross, M. (1975) *Méthodes en syntaxe. Régime des constructions complétives*, Hermann, Paris.

Huddleston, R. and G.K. Pullum (2002) *The Cambridge Grammar of the English Language*, Cambridge University Press, Cambridge, UK.

Kayne, R.S. (1991) "Romance Clitics, Verb Movement and PRO," *Linguistic Inquiry*, 22, 647-686 (reprinted in Kayne (2000)).

Kayne, R.S. (1998b) "A Note on Prepositions and Complementizers," article posted on the Chomsky Internet Celebration, The MIT Press (also in Kayne (2000) as "A Note on Prepositions, Complementizers and Word Order Universals").

Kayne, R.S. (2000) *Parameters and Universals*, Oxford University Press, New York.

Kayne, R.S. (2005) "Some Notes on Comparative Syntax, with Special Reference to English and French" in G. Cinque and R. Kayne (eds.) *Handbook of Comparative Syntax*, Oxford University Press, New York, 3-69 (reprinted in Kayne (2005) *Movement and Silence*, Oxford University Press, New York).

Kayne, R.S. (2007) "Some thoughts on grammaticalization. The case of *that*", (handout of) talk presented at the XVIIIe Conférence internationale de linguistique historique, UQAM, Montreal.