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.