Some thoughts on NYU's *Syntactic Structures of the World's Languages* (SSWEL)

MARTIN HASPELMATH
Max-Planck-Institut für evolutionäre Anthropologie, Leipzig

1. Realizing Gabelentz's (1894) Committee: "language-expert-based typology"

Georg von der Gabelentz. 1894. "Typologie der Sprachen, eine neue Aufgabe der Linguistik." *Indogermanische Forschungen* 4: 1-7.

Die Arbeit verlangt eine Kommission, und die Kommission verlangt ein bis ins Einzelnste ausgearbeitetes Programm, und dies Programm verlangt mehr selbstentsagenden Gehorsam, als man von der Mehrzahl der Gelehrten erwarten darf. Doch solche Schwierigkeiten sind zu überwinden. Unter dem Programme aber denke ich mir eine Art Fragebogen, der kategorienweise alle grammatischen Möglichkieten erschöpft, so dass jede Frage mit einem Ja oder Nein beantwortet ist. Eine solche Fragestellung ist schwierig für den Fragesteller selbst, manchmal auch für den Beantworter; aber Unmögliches wird keinem der Beteiligten zugemutet. (Gabelentz 1894:6)

"The work needs a committee, and the committee needs a programme attending to every single particular, and this programme needs more self-denying obedience than can be expected from the majority of scholars. But such difficulties have to be overcome. I conceive of the programme as a kind of questionnaire which covers all grammatical possibilities category by category, with each question answerable by a yes or no. Framing such questions is difficult for those asking them and probably also those answering them; but nothing impossible will be asked of them."

Two projects of this kind at the Max Planck Institute for Evolutionary Anthropology:

***Loanword Typology**: Lexical Borrowing in the world's languages*
(Haspelmath & Tadmor (eds.), 2003-2008)
– 40 experts contribute lexical data with historical information (ca. 1500 words)

***APiCS** (Atlas of Pidgin and Creole Language Structures)*
(Michaelis, Maurer, Huber & Haspelmath (eds.), 2006-2010)
– 60-70 experts on pidgins and creoles contribute data on 120 structural properties

2. Problem # 1: Contributor motivation

Why would anyone contribute to SSWEL? To share their knowledge with the world! This works with Wikipedia, but will it work in science as well? Does it work with professionals?

Glottopedia (http://www.glottopedia.org) works on this principle, but so far it does not work particularly well (too few contributions).

Hypothesis: Wikipedia works because lay experts have no good alternative audiences; but academics can always decide to spend their time for another paper or conference presentation, where they get more tangible recognition.

The World Atlas of Language Structures (WALS):
Every contributor got a conventional publication (even though a short one: 4 pages), and a copy of the book. In addition, many of the *WALS* authors were invited to *WALS* workshops in Leipzig, and a *WALS* workshop was held in Santa Barbara in 2001.

Similar approach for *Loanword Typology* and *APiCS*

3. Suggested solution: Recognized publication

The datasets contributed by the SSWEL authors should be conceived of as regular publications, with proper refereeing and editorial decisions. This means that
– an editorial structure has to be put in place (editor-in-chief, editorial board, referees);
– only relatively coherent, relatively sizable contributions are accepted;
– each dataset should have a clear "address" in the overall system, and each data point should be easily relatable to the dataset it belongs to.

Three necessary (and jointly sufficient) conditions for regular, recognized publication:
– peer recognition (note: "intellectual property" is not an issue for academic linguists)
– permanent availability
– fixed content

SSWEL would then be a kind of "journal" for language-specific datasets prompted by the SSWEL property set.

> Chris Collins: "*The open-ended and fundamentally dynamic nature of the database will make it impossible for the data to be refereed in any standard sense.*"

• Why? All journals are open-ended, and science is by nature dynamic. It is true that with such a database, the contributions are not naturally individuated, but an artificial delimitation is always possible.

> Chris Collins: "*it is important to aim at the broadest possible audience in order to increase the total amount of data in the database... be designed so that it can be used by amateurs*"

• But amateurs do not have a reputation to lose, and may find it hard to look objectively at "their language".

• Aiming at amateurs would not only mean that "the data-entry interface has to be simple", but also that sophisticated syntactic properties have to be left aside. Note that even linguists often find it difficult to understand other linguists (about 15% of all answers to the first version of the *APiCS* questionnaire were wrong because the property and value descriptions were misunderstood).

4. Problem # 2 (the biggest problem): Cross-linguistic identification of properties and values

Chris Collins: "*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 problem is more serious:
• How can I be sure that the concept "linker" even makes sense in other languages?

Syntactic categories are defined by language-specific properties, so they cannot be carried over from one language to another (Boas 1911, Saussure 1915, Fries 1952, Dryer 1997, Croft 2001, Lazard 2005, Haspelmath 2007)

> Example from *APiCS*:
> Property 63: Order of Adjective and Noun (adopted from *WALS*)
>
> 1. Adjective precedes noun
> 2. Adjective follows noun
> 3. Adjective occurs in internally-headed relative clause
>
> Nigerian Pidgin:
> "Adjectives do not exist. Instead stative verbs are used to convey meanings equivalent to those conveyed by adjectives in languages such as English."
>
> Should we add this as a further property value? But how do we define "adjective" across languages?

• Chris Collins: "*The notion of an "ontology" should help to resolve this issue.*"

But ontologies can help only where the concepts are in principle clear. However, terms for grammatical categories are extremely language-dependent, or at least extremely theory-dependent.
(Generative grammarians often assume, but have rarely argued, that language-specific categories can be identified with categories in other languages. In practice, the difficulties are enormous.)

• Grammatical concepts are often defined by a "battery of diagnostics/tests", which ideally yield the same results:

Hornstein 1999, for distinguishing	**PRO** from (control)	**pro**
non-sentence-internal antecedent	no	yes
antecedent choice subject to MDP	yes	no
non-c-commanding antecedent	no	yes
strict reading under ellipsis allowed	no	yes
split antecedent allowed	no	yes
object antecedent in adjuncts	no	yes
alternates with an overt DP	no?	yes

In published papers, these test batteries behave uniformly, and many theorists expect them to behave uniformly in all languages. However, a system such as SSWEL can hardly be based on the correctness of this conjecture.

5. Suggested solution: **Acknowledge that the cross-linguistic properties are distinct from language-specific categories, and that their definition is somewhat arbitrary**

e.g. *adjective* = property concept word (denoting age, dimension, value, colour); adjectives are Adjectives in English, Stative Verbs in Nigerian Pidgin, Nouns in Quechua, etc.
Haspelmath (in prep.): ***descriptive categories*** vs. ***comparative concepts***
("universal/cross-linguistic categories" play no role in typology, pace Newmeyer 2007)

comparative concepts:
concepts that can be based on meaning (or even better, non-linguistic stimuli) plus (possibly) on general, nongrammatical concepts such as 'A precedes/follows B', 'A is identical to/different from B', 'A is overt/covert'

e.g. *ergative case* = a case that marks the agent of a typical transitive verb (such as 'kill', 'break', 'throw'), when the patient is marked in the same way as the intransitive single argument.
Ergative case is Relative Case in Inuktitut (ergative + possessive), Instrumental Case (ergative + instrumental) in Kalkatungu, etc.

"Composite properties" such as "control" have to be taken apart; only their more elementary components should be properties of the database.

Even concepts such as "c-commanding antecedent" or "ellipsis" are very problematic, because they can be defined only on the basis of highly specific assumptions and claims.

6. Problem #3: Languages have far more properties than can be captured by pre-formulated property checklists

Checklists like the *WALS* feature set or the *APiCS* questionnaire only ask about a few selected properties that are salient in the eyes of the current generation of linguists.

e.g. Comitative-Instrumental Identity vs. Differentiation (*WALS* ch. 52)

But many other things could be said about comitative markers and instrumental markers (they could express yet other meanings, they can govern specific cases, they can occur in specific stranding constructions, they may show different kinds of agreement, they may occur with certain verbs which are lexically specified for them, they may show peculiar allomorphy, syncretism, etc.)

7. Solutions: **(1) Accept the shallowness of cross-linguistic work; (2) regard the database as a kind of "index"**

(1) Shallowness of comparative work

While descriptive grammars need to go into all the details of language structures, comparativists have to focus on those aspects of language structure that are readily comparable. These are relatively circumscribed aspects of language structure (though perhaps intuitively "the core"), and one could say that comparativists only "scratch the surface" of the particular languages they include in their surveys.

This may be unsatisfactory from the point of view of the Parametric Program ("reduce ALL core grammatical properties to a few dozen parameters"), but it is a necessary step on the way to carrying out the Program if it is realistic after all.

(2) Database as "index"

The property-value pairs are considered as only one salient aspect of the properties of the language, and as an opportunity for "comments", which can be short, longer, or even lengthy prose statements.

Some people would say that *these prose statements in the comments sections are the real value of such a database*, because they lend themselves more readily to reinterpretation. The property-value pairs would then not be much more than an index to the prose statements and to the example sentences.

8. Problem #4: Often linguists are more interested in the properties of different constructions than in the properties of different languages

e.g. *Variation in Control Structures* questionnaire:
"List all the control structures of your language; for each structure, answer the following questions..."

Berlin-Utrecht Reciprocals Database (BURS) questionnaire:

"List all the reciprocal markers of your language; for each marker, answer the following questions..."

Leipzig Ditransitive Constructions questionnaire (Malchukov, Haspelmath, Comrie): "List all the ditransitive constructions of your language; for each construction, answer the following questions..."

This is a real problem only if the database is conceived of as rigidly as a set of property-value pairs for entire languages. But the database can just as easily be conceived of as a **database of constructions**.

9. On the units/subdatabases of the overall SSWEL database

The SSWEL data should be integrated and interoperable, so that one can search for data across languages and across properties. Minimally, the units are ***language*** and ***property***.

But other units are possible, and probably necessary. Linguists are more likely to contribute information if they know that their name will be associated with a recognizable part of the database (and ideally these parts should count as separate publications).

Property subsets

A set of properties may relate to the same overall theme (such as "Control", "Anaphora", "Ditransitives"), and it may require a lengthy introduction (cf. the *Variation in Control Structures* and the *Afranaph* questionnaires). The introduction and the formulation of the questionnaire should count as a publication in its own right. And of course such questionnaires will often be theory-specific, but if they are clearly labeled as belonging to a particular property subset, this is not a problem.

Language datasets for property subsets

Some linguists may be willing to contribute information on one aspect of their language that they happen to have data on (e.g. Control, Ditransitives), without being able to contribute data on all the SSWEL properties. In such cases, it would be good if their names could be associated (again, ideally through publication) with a dataset for a well-delimited property subset.

10. Plan for a new database journal ("*World Language Data Journal*"), to be published by the Max Planck Digital Library

We want to create an open-access online journal that will publish cross-linguistic databases in a standard peer-reviewed fashion. These additional databases will be integrated with the *WALS* data and will allow all the working tools (searching,

geographical display, exporting, etc.) that the *WALS* Tool allows. The database journal addresses a clear need in the field of language typology: Unlike books and papers, cross-linguistic databases cannot at the moment be published in a regular, accepted way (many of them just sit on the author's website and are accessible just in whatever way the author happens to allow, and many others just sit on the author's hard disk).

Bibliographical reference (hypothetical):

> Bondarchuk, Vera N. 2011. "Multiplicative numerals in the world's languages." *World Language Data Journal* 3/12. (URN:nbn:de:0008-5-357)

If SSWEL also becomes a database journal, then the question arises how SSWEL and the WLDJ would relate to each other. How could they be searched simultaneously?

Two ways of dealing with this come to mind:

(i) **A meta search engine (the "union catalogue" model)**. The Typological Database System (TDS) was apparently conceived of in roughly such a way. The idea was originally that it would allow users to simultaneously search different databases that exist in different places. This is analogous to "union catalogues" of several libraries that can be searched simultaneously.

(ii) **A desktop database system using common standards** (the **"EndNote" model**). Suppose we had something like EndNote (call it "CrossLang"), that would allow one to store one's cross-linguistic data in a standardized format, without having to worry about interoperability. Endnote provides for standardized data structures and export formats for bibliographical reference information, and if a resource uses a standard format, Endnote can easily download information from that resource. CrossLang would do the same for typologists: It allows a typologist to download available data from elsewhere (e.g. data on languages, such as names, references, etc.), to enter one's own data, and to share them with others. Basically it would have all the functionality of *WALS*'s CD-ROM, plus the ability to import and export data much more freely. If we had such a programme that was freely available to typologists, then all we would have to worry about would be that published typological databases should have export functions that are compatible with CrossLang.

References

Boas, Franz. 1911. Introduction to *The Handbook of American Indian Languages*.
Croft, William. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.
Dryer, Matthew. 1997. "Are grammatical relations universal?" In: Bybee, Joan & Haiman, John & Thompson, Sandra A. (eds.) *Essays on language function and language type*. Amsterdam: Benjamins, 115-143.
Fries, Charles C. 1952. *The structure of English: an introduction to the construction of English sentences*. Longman.

Gabelentz, Georg von der. 1894. "Typologie der Sprachen, eine neue Aufgabe der Linguistik." *Indogermanische Forschungen* 4: 1-7.

Haspelmath, Martin. 2007. Pre-established categories don't exist—consequences for language description and typology. *Linguistic Typology* 11:119-132.

Haspelmath, Martin. In prep. "Descriptive categories and comparative concepts in cross-linguistic comparison."

Haspelmath, Martin. & Matthew S. Dryer & David Gil & Bernard Comrie, (eds.) *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Lazard, Gilbert. 2005. "What are we typologists doing?" In: Frajzyngier, Zygmunt & Hodges, Adam & Rood, David S. (eds.) *Linguistic diversity and language theories.* Amsterdam: Benjamins, 1-23.

Newmeyer, Frederick J. 2007. Linguistic typology requires crosslinguistic formal categories. *Linguistic Typology* 11:133-157.

Saussure, Ferdinand de. 1915. *Cours de linguistique générale*.