INDUCING TYPOLOGICAL GENERALISATIONS IN A CROSSLINGUISTIC DATABASE

Simon Musgrave

Spinoza Project: Lexicon and Syntax, University of Leiden Centre for Linguistics
Postbus 9515, 2300RA Leiden
The Netherlands
S.Musgrave@let.leidenuniv.nl

Abstract

The Spinoza Typological Database (STDB) is being developed as a tool for the comparison of the syntactic, morphological and lexical properties of a sizable sample of languages. It includes primary data for each language, in the form of a text sample of at least 50 clauses and a vocabulary list of around 200 items. Every portion of the data (specifically the text sample) will be analyzed to the extent that is relevant and quantitative typological generalizations will be produced by the application on the basis of the stored data. These statements then provide the basis for fine-grained cross-linguistic comparison. The resulting architecture means that STDB is an innovative and powerful tool, useful for cross-linguistic research but also with great potential for use by field linguists.

1. Introduction

The Spinoza Project: Lexicon and Syntax (SPLS) is a research project directed by Prof Pieter Muysken (Catholic University of Nijmegen). The project aims to investigate basic properties of human language through a detailed study of the phenomena observed in situations of intensive language contact, especially where the languages involved are genealogically unrelated. Four geographical areas were selected for close study: the Balkans, Bolivia/Rondonia, Eastern Indonesia, and Suriname/Benin/Ghana. Data on approximately eighty languages from these four regions is available to the project, or is being collected as a part of the project. In order to assist comparison over such a large body of data, the project is developing a database to contain both primary and analytic data, the Spinoza Typological Database (STDB). In addition to the data generated by the areal studies of the project, the database will also contain parallel data from a balanced sample of the world's languages (Rijkhoff, Bakker, Hengeveld & Kahrel 1993). This data will serve as controlled comparative data for the areal sample, and will be assembled under the supervision of Prof Kees Hengeveld (University of Amsterdam).

2. Types of data

The STDB contains six types of data for each language included.

i. General background data - This data includes information about the geographical location of the speech community which uses the language, the size of that community, the status of the language in the community (i.e. whether it is the only language, the primary language or a secondary language), whether it is used as a written language, and alternative names for the language.

ii. Data on sources - This data enables the end-user of the database to trace any piece of a primary linguistic data to a specific source. The possible types of source include published works, field notes, recordings, native speaker judgments etc.

iii. Data on analysts - This data enables the end-user to trace any piece of primary linguistic data to a specific analyst.

iv. Texts - This data is the core of the database. For each language, we aim to have a text of at least 50 clauses. Several representations of each clause are stored: orthographical, a Roman transliteration where required, phonemic, morphological analysis, morphological gloss, partial syntactic analysis and a free translation. Information about borrowed items (a very important matter in the context of the overall project) is also stored. Isolated sentences and paragraphs will also be included here, where they are necessary to clarify analytic points.

v. Vocabulary list - For each language, a basic vocabulary list will be collected. This list consists of the 200 word Swadesh list plus a small number of additional items from the Natural Semantic

Metalanguage list of semantic primes (Wierzbicka 1996). The research team for each of the four target regions can also nominate additional items to be included for that region alone.

vi. Typological analysis – Data on a range of typologically interesting variables is stored for each language. The majority of this data consists of classical word order information, but information about word class systems and processes of derivational morphology is also included. As far as possible, this data is collected as generalisations over the analysis of individual units, rather than as higher-level analytic statements inputted directly by an analyst (see below for further discussion).

3. Interrelations in STDB

All data pertaining to any particular language is interrelated by the use of a unique identifying number for each language. Each item of primary data is tied to a specific analyst and a specific source. Where applicable, an item in the vocabulary list is cross-referenced to occurrences in the texts for that language. Finally, typological statements about a language are linked to a set of examples in the primary data for that language. Such sets of examples include units of language of varying size, from a single morpheme up to a text line. The strategy used to keep track of these units is discussed below.

4. Text data and typological generalisation

The STDB was always intended to treat primary data, in particular text data, as being of great importance. However, in the process of developing the application, this orientation has assumed greater importance. The original intention was that the text data would be available to illustrate the typological analysis which the analyst provided, that is the data would have the kind of top-down structure commonly used in typological databases. Higher-level generalisations are entered directly, and a greater or lesser amount of supporting evidence is provided as the analyst sees fit. The current architecture of the STDB is rather different. In this scheme, the analyst identifies units in the primary data, and the application then responds by asking the analytic questions relevant to that linguistic unit. The typological statements which can be made about any particular language in the database are then summations of the individual analyses which have been entered.

This approach was initially used for morphological analysis. One feature that was considered highly desirable in the database was for morphological analysis and glossing to be represented in aligned interlinear text (as in the SIL Shoebox application). To do this, it is necessary to make the morpheme the basic unit of stored text data; this move in turn opened up many other possibilities. For example, information about derivational morphology is naturally gathered during the process of inputting the morphological analysis. Whenever a morpheme is identified as an affix (simplifying slightly - there are other possibilities), this triggers a form which first asks whether the morpheme is derivational in effect, and then goes on to gather additional data about the derivational process, if relevant.

The further possibilities are dependent on a significant difference between STDB and Shoebox. Interlinear text is generated in Shoebox and then saved as a part of the data structure. The information on which the interlinear is based, the sequence of morphemes in a text line, is not saved. STDB, on the other hand, saves the sequence of morphemes and generates interlinear text on each occasion that it is needed. Thus, one representation of a line of text in this system is as an ordered list of references to dictionary entries. But from this point of view, all linguistic units above the morpheme can be represented identically. Therefore it is straightforward to define such units (NPs, clauses etc.) in the same way, as ordered lists of dictionary references, and to tie analytic statements to the units so defined. Once the analyst has broken a text into morphemes, they can then be asked to identify relevant units within each text line, and appropriate analytic questions can be asked about those units. The application can be constructed to repeat this process until individual words or morphemes are reached. In fact, the Spinoza application will not be exhaustive in this sense, but the architecture allows the possibility.

The application then automatically generates quantitative typological information on the basis of the stored analysis. For any language, a summary of the distribution of the analyzed features is available. For example, if the relative order of noun head and numeral modifier is of interest, the summary will tell how many such examples occur in the text sample, how many examples have the head first and how many have the numeral first (of course, one figure may well be zero). The full

data set exemplifying each summary number is available with one mouse click. The final output of the STDB is therefore rather different from that of many databases of typological information. There are no categorical statements of the type: language X has numerals preceding noun heads, but more detailed and more nuanced information is provided. This has several advantages:

i. it reflects more truly the facts of natural language use
ii. it serves the purposes of the SPLS, which by its very nature will be looking at cases in which the typology of a language may be changing under the influence of one or more contact languages
iii. it allows the user to assess the data for herself: the primary data supporting each statement are easily accessible, especially the examples which are counter to a common pattern.

5. Potential uses

Although STDB has been designed with the needs of the SPLS in mind, it has great potential as a tool for field linguists. The project's data will be inputted by individual analysts working on one language at a time; comparison across languages will be essentially a feature of the application in output mode. Therefore the application can be used to analyze data from a single language for morphological and syntactic features. The morphological analysis section is intended to cover the full range of possibilities in human language – this is necessary given the size of the sample which is to be included. The lexical information collected in the current implementation is minimal, but it would be a straightforward task to enrich this component. The design of the syntactic analysis module is influenced by the fact that the application will have many users, therefore the analytic schema and the terminology employed have been kept as theory-neutral as possible. These considerations have also meant that the analysis is not as exhaustive as might be considered desirable in some circumstances. However, the architecture of the application is such that a future user could modify these aspects and retain the crucial functionality which allows analytic statements to be linked to specified linguistic units.

6. Problems and solutions

6.1. Character sets

The adoption of the Unicode standard has simplified the problem of character sets for those working with computational tools for linguistics. The STDB is constructed as an Access database application, and therefore Lucida Sans Unicode was the natural choice as the basic font for the project, it being the Windows Unicode font with the most extensive character set. Two problems had to be solved nevertheless. Firstly, Access allows only very limited formatting in text fields, therefore a means of representing tone not dependent on such techniques was required. Fortunately, superscript integers 0-9 are separately coded in the Unicode standard, and STDB uses them for this purpose. Secondly, the only Windows Unicode font suitable for aligned interlinear text is Courier, but this is a limited character set without, for example, IPA characters. Instead, we use a font called Monospace, designed by George Williams (Williams 2000-1), which does include IPA characters.

6.2. Incomplete data

The initial design of the application treats all typological data as Boolean variables. During the development process, it has become clear that this will not be satisfactory. There are inevitably cases where the data does not allow for a definite answer to an analytic question, and where additional data is hard or impossible to obtain. The only honest response in these situations is "don't know", and the general approach described above requires that this is the response that the end-user should see. (Note that simply not responding in such a situation is not sufficient: then a NEG value is ambiguous between a real "no" and "don't know".) This issue does not arise in some parts of the analysis, where the questions to be answered are dependent on the units identified by the analyst (if a numeral and a noun head are identified, they must have a relative order), but in other places the question does arise. There are therefore two problems to solve here. First, we must consider carefully where the option "Don't know" should be allowed: to some extent this will not be fully resolved until testing with users begins. Secondly, there is the technical issue of how the information should be stored. Using Boolean variables has the advantage of giving a transparent data structure, an important consideration for long-term maintenance and for the possibility of the Spinoza database being part of an integrated group of typological databases (see contribution to this meeting by Monachesi et al). Nevertheless, it may be necessary to use multi-valued variables in some cases in order to cover the case of incomplete data.

7. References

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel (1993) A method of language sampling *Studies in Language* 17: 169-203

Wierzbicka, Anna (1996) *Semantics: Primes and Universals* Cambridge UK: Cambridge University Press

Williams, George (2000-1) Monospace font, Unicode and ISO 8859. Available at the following url: http://bibliofile.mc.duke.edu/gww/fonts/Monospace/index.html