Workshop on the Feasibility of a Web-based Database of the Syntactic Structures of the World's Languages

Peter Cole — Max Planck Institute for Evolutionary Anthropology
November 9, 2007 — University of Delaware

I. General Remarks on What We are Trying to Do

- Primary Purpose of Database
  - To make our swiftly increasing factual knowledge about the structures found in languages easily available to the community of linguists
- Possible Approaches
  - Annotate the data itself for structure (as in Treebanks), thereby allowing searches on structure
    - Problems:
      - Not everything can be found in trees
        - e.g. How do you search for relative clauses when they might consist of a variety of structures?
        - Typology often deals with distribution of functional information across languages
      - Structure is controversial: whose structures should be used?
      - How abstract should structures be?
    - Despite these problems, it would be very useful to have a large quantity of parsed data about a large number of languages.
  - Database of answers to questions someone has asked
    - What questions are asked depend on theoretical assumptions (questions can get stale fast)
      - 1965: What transformations occur in language X and how are they ordered?
      - 2007: Does internal merge predict the facts of Language L?
      - etc.
    - The answers to different questions need differently structured databases.
  - What level of abstraction is desirable in databases of this kind?
    - We may (will?) want less abstraction in our databases than in our theories.
    - Over the years we have added new sets of "facts" that need to be explained by each theory we come up with.
    - These sets of facts do evolve, but much more slowly than do our theories to explain the facts.
    - We need to aim at a level of abstraction that allows us to separate the facts from the theories so we can test the theories against the facts.
    - It is hard to give a formal definition of what that level of abstraction is, but there may be more agreement in identifying it than in defining it.

- I.e. there may be more art to what we are tring to accomplish than science.

II. Questionnaires

- A step toward a solution
  - Base databases on explicit, published questionnaires, each dealing with a different problem or based on different assumptions
  - The database is a set of structured answers to the questions in the database
  - Attempts can be made to eliminate **unnecessary** structural and definitional dissimilarities among databases, to allow databases to be combined, searches across databases etc.
  - Means of communication among developers of databases/questionnaires need to be formalized to encourage cross-database similarities
  - Need to aim at appropriate (intermediate) level of abstraction

- Questionnaires are critical to creating databases
  - The simplest database
    - prose questions
    - prose answers
  - These can be surprisingly effective as a research tool if they are electronically searchable
  - Examples: The Comrie/Smith Lingua questionnaire and the series of grammars derived from it, a large and very useful dfatabase
- What questionnaires exist and how can linguists access them?
- What **answers** can be found and how can they be accessed/searched?
  - We have started making the questionnaires accessible, but there is less progress in making the answers electronically available.
- As far as I know, the only **collection** of typologically oriented questionnaires is on a website maintained by MPI EVA in Leipzig: **Typological Tools for Field Linguistics** (formerly Software Tools) http://lingweb.eva.mpg.de/fieldtools/tools.htm

- Website edited by
  - Peter Cole
  - Jeff Good
  - Claudia Schmidt (day-to-day responsibility for website)
- Structure of Website
  - Questionnaires
  - Stimulis Kits for Data Elicitation
  - Books and Articles to Guide Data Collection and Grammar Writing
  - Glossing Rules
  - Information on other websites
- There are many more questionnaires than other types of material
  - Questionnaires were originally presented in random list
  - There are now too many for this to work

- We have revised the random list to a structured list and are in the process of implementing this
- Eventually, we may need to have a database of questionnaires if the list continues to grow

III. Statement of Purpose of website

This website contains tools for use in field linguistics and language description. Most of the items on the website are questionnaires designed to assist in eliciting data in such a fashion that the data will be comparable across languages. The idea behind this website is that field linguistics should be typologically informed and that the results of field work should be of typological interest. The questionnaires and other tools presented here help the field linguist understand what questions might be of typological (and theoretical) interest and guide the linguist in both eliciting data and extracting information from naturalistic texts. In addition, we would like to include "elicitation kits", allowing the researcher to present movie clips and similar language stimuli to native speaker consultants to see how they would describe the event shown. The number of elicitation kits available is still very limited, but we hope that more will be available in the future. We also include information on books and articles that provide down-to-earth guidance to the field linguist.

The emphasis in this website is on tools for language description, as opposed to tools for language documentation, suggestions for choice of hardware etc. Links to websites that provide information on topics beyond the purview of our website can be found at the bottom of the page. By a "typologically informed field linguistics" we mean field work guided by the goal of examining the range of possible variation in human language. This goal can be approached in a very concrete fashion that examines surface variation, or it can be approached from the perspective of a highly abstract model in which surface variation is of less interest than variation in abstract patterns or distributions. Our intent is to be resolutely agnostic about such questions, and to include "tools" designed for both purposes. We see questionnaires, elicitation kits etc. as a way for experts in a particular area of interest to pass on their practical knowledge to field workers whose training or primary interest may be in some other area of linguistics. We also see the website as a useful tool for classes in linguistic field methods. The materials on our website can help guide the students toward a productive use of their time with native speaker consultants.

- Although the website was conceived of as a research tool for the field linguist, its typological orientation makes it a tool for any sort of typological or cross-linguistic study
    - This would be much more the case, if the website featured answer sets, not just questionnaires

IV. Structured List of Questionnaires as of 10/26/2007

What questionnaires are available? This is what is on the website right now.

1) General Questionnaires

The Lingua Descriptive Studies Questionnaire (Comrie & Smith 1977)

2) Syntax/Morpho-Syntax/Semantics

a. Tense/Aspect

1. Dahl, Östen, 1985

b. Anaphora

1. Utrecht Anaphora Questionnaire (Alexis Dimitriadis and Martin Everaert)
2. Rutgers Questionnaire on Anaphora in African Languages (Ken Safir)

c. Control
1. Questionnaire for Control Verbs (Barbara Stiebels)

d. Information structure
1. Berlin Questionnaire on Typology of Information Structure (Stavros Skopeteas, Ines Fiedler, Samantha Hellmuth, Anne Schwarz, Ruben Stoel1, Gisbert Fanselow, Caroline Féry, Manfred Krifka)

e. Imperatives
1. A Typological Questionnaire on Imperative Constructions (Viktor S. Xrakovskij)

f. Valence
1. Questionnaire for Transitivizing/Detransitivizing verb systems (Johanna Nichols)
2. Valency Questionnaire (developed by Marian Klamer for the 2000 East Nusantara Linguistics)
3. Questionnaire for a diachronic typological study of valency-changing categories (Leonid Kulikov)

g. Negation
1. Questionnaire for description of negation systems (René van den Berg and Pater Kahre)

h. Ditransitive Constructions
Questionnaire on Ditransitive Constructions. (Bernard Comrie, Martin Haspelmath & Andrej Malchukov)

i. Word Formation
1. Word-formation questionniare (Pavol Stekauer)

j. Motion Events
1. Checklist for the description of Motion Events (Bernhard Wälchli)
2. Questionnaire on Motion in Australian Languages (David Wilkins, David Nash and Jane Simpson)

k. Grammatical Domains
1. Questionnaires relating to phonological and grammatical domains in languages (developed by the Word Domains project, University of Leipzig)

3) Phonology

a. General Phonology
Phonology questionnaire (Dan Everett)

b. Stress
1. Questionnaire on Stress Typology (Harry van der Hulst & Rob Goedemans)

i. Phonological Domains
1. Questionnaires relating to phonological and grammatical domains in languages (developed by the Word Domains project, University of Leipzig)

4) Lexographic

a. STEDT Questionnaires: the linked questionnaires were developed by James A. Matisoff and collaborators as a means of obtaining data for the Sino-Tibetan Etymological Dictionary and Thesaurus.

b. Typological Aspects of Figurative Language (David Gil and Yeshayahu Shen)
c. SIL Comparative African Wordlist (SILCAWL) (James Roberts and Keith Snider)

5) Languoid Specific Questionnaires
a. East Nusantara Questionnaires for various years)
b. Outline/questionnaire intended to assist researchers in writing grammatical sketches of languages related to Persian (John Roberts)
c. Outline for a grammar of a Papuan language (John Roberts)
d. SIL Comparative African Wordlist (SILCAWL) (James Roberts and Keith Snider)

6) Miscellaneous
a. Language Contact
Language Contact Questionnaire (developed by John Bowden for the 2000 East Nusantara)
b. Oral Traditions
Oral Traditions Questionnaire (developed by Margaret Florey for the 2000 East Nusantara)

V. Possible activities with a short-term to medium-term payoff

- Expand number of publicly available questionnaires
- Provide repositories for **both the questionnaires/stimulus kits and the responses** (different groups of researchers can provide questionnaires per their specific interests)
- **Link** these various sites and make use of a cover site (like the Leipzig website), where anyone looking for this sort of typological data can go **first** for immediate help.

VI. Some Miscellaneous Issues

- Plusses and minuses of **unstructured text** databases (no fields)
  - \+ easier to include **analysis**, making the data more meaningful and less open to misinterpretation
  - \- easier to obscure data by the analysis, making it hard to separate the facts from the theory
  - \+ easier to view the work as traditional scholarship, and, therefore, to reward in a academic context
  - \- much harder to extract information than from a structured database
- A suggestion: provide different products based on a single questionnaire
  - a structured database consisting only of "facts" (to the extent possible)
  - accompany the database by a fully searchable prose work based on the same questionnaire, in which the author provides an analysis for the data
  - referee the whole process and suggest on the website what a contribution is "worth" (article, monograph etc.) in terms of traditional scholarly categories
- An example:
  - Utrecht Anaphora Questionnaire (Alexis Dimitriadis and Martin Everaert)

- The Anaphora Typology Database (http://languagelink.let.uu.nl/anatyp/), based on that questionnaire as applied to four languages, Hungarian (collected by Judit Gervain), Korean (collected by Na-Rae Han), the Peranakan Javanese of Semarang, Indonesia (collected by Peter Cole, Gabriella Hermon, Yassir Tjung, Chonghyuck Kim, Chang-Yong Sim, Yaping Tsai), and Sakha (Yakut), collected by Nadya Vinokurova.
- a short monograph (154 pages), providing **analysis** for the data in the database (and other data relevant to and providing context for the analysis)
  Title: Anaphoric Expressions in the Peranakan Javanese of Semarang
  Series Title: LINCOM Studies in Asian Linguistics 72
  Published: 2007
  Publisher: Lincom GmbH
  http://www.lincom.eu
  Authors: Peter Cole, Gabriella Hermon, Yassir Tjung, Chang-Yong Sim and Chonghyuck Kim
  Paperback: ISBN: 9783895860409 Price: € 58.00 (sic!!!)