The Typological Database System

Paola Monachesi
Utrecht University
Uil-OTS
Trans 10
3512 JK Utrecht

email: Paola.Monachesi@let.uu.nl WWW:

http://www-uilots.let.uu.nl/ Paola.Monachesi/
Project description

- The Typological Database System project is carried out at Utrecht University.
- It is partially funded by The Netherlands Organization for Scientific Research (NWO), as well as by the participating institutions which belong to the Netherlands Graduate School of Linguistics (LOT).
- Typological databases are a crucial tool to improve our understanding of the amount of variation natural language exhibits and to distinguish true universals from the spurious ones.
- The typological database will be part of a Language Typology Resource Centre, a web-accessible electronic archive for typological description, including powerful research tools such as extensive scientific grammars, typological databases and language-typological expert systems.
Project aims

• Collect relevant data and enter information for over a hundred languages (fifty from a genetic sample and fifty from various areal samples) into databases.

• Collect very detailed information for a limited set of specific grammatical properties, including anaphoric relations, scrambling, agglutinative morphology, inflectional paradigms, morphological marking of the (anti)causative alternation and quantification.

• Convert an open ended number of existing databases into the required format.

• Create a system that is able to link separate databases in such a way that we can ask questions over the whole set of databases: development of a metalanguage.
Participating Institutions: NL

• University of Amsterdam: databases on word order and agreement.
• University of Leiden: Spinoza project, Stresstyp.
• University of Nijmegen: database on intransitive predication, possession, comparison, temporal sequencing.
• Utrecht University: databases on anaphora, inflection and aspect.
Participating Institutions: International

- FU Berlin: database on intensifiers and reflexivity
- University of Delaware: database on wh-questions
- University of Konstanz: database on language universals
- Lancaster University: database on agreement
- MPI Leipzig
- University of Manchester
Spinoza project

- WWW: www.let.leidenuniv.nl/spls
- Site: Leiden
- Contact: P. Muyskens, M. Klamer, E. Mauden
- Description of the project: information about 100 languages will be encoded. Orthographic, phonological, morphological and syntactic information is foreseen. Glossed examples are provided.
- Software: MS Access
- Duration: Project started mid 99.
StressTyp project

- WWW: fonetiek-6.leidenuniv.nl/pil/stresstyp/stresstyp.html
- Site: Leiden
- Contact: R. Goedemans
- Description of the project: stress system of the languages of the world. Sources are grammars and articles that provide descriptions of the stress patterns and also theoretical work on stress. 500 languages are coded.
- Software: 4th dimension, runs on both PC and Mac
Eurotype project: word order

- Site: Amsterdam
- Contact: D. Bakker
- Description of the project: information about word order for 150 languages. Only European languages. Data collected through questionnaires. No examples, only use of variables.
- Software: own software
- Duration: project has ended.
Agreement

- Site: Amsterdam - Lancaster
- Contact: D. Bakker, A. Siewierska
- Description of the project: information about agreement collected for 400 languages. The languages constitute a balanced sample of all the languages of the world. The sources are grammars and articles. No examples, only use of variables.
- Software: own software
- Duration: database is still being developed.
- Example Agreement DB (Lancaster)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>V52</strong> – 1</td>
<td>type agreement</td>
</tr>
<tr>
<td><strong>V54 – 4b</strong></td>
<td>gender on possessive pronouns</td>
</tr>
<tr>
<td><strong>V54 – 5b</strong></td>
<td>incl-excl on poss pronouns</td>
</tr>
<tr>
<td><strong>V71 – 0</strong></td>
<td>agr Adj predicate</td>
</tr>
<tr>
<td><strong>V80 – 0</strong></td>
<td>presence of a passive</td>
</tr>
</tbody>
</table>
Typological Data Base Nijmegen

• Site: Nijmegen

• Contact: L. Stassen

• Description of the project: information on a variety of topics including: word order, intransitive predication, temporal sequencing, comparatives, possessive constructions. Number of languages vary from topic to topic, with a minimum of 140 for all topics, and a maximum of 410 for some topics. Sources are grammars, articles and native informants. There are no examples, only use of variables.

• Software: SPSS

• Duration: database is still being developed.
- Example DB Nijmegen

<table>
<thead>
<tr>
<th>V27</th>
<th>PRED ADJ AGR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicative adjs agree with the subj in nb and/or gender</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V106</th>
<th>ATTR ADJ AGR CASE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attributive adjs agree with their nominal heads in case</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V456</th>
<th>VERB FLEX SUBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Finite verbs agree with their subjects</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V469</th>
<th>FLEX ORDER = VERB-TMA-X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In a V the morpheme order is Stem-Tense/Mood/Aspect-Agr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V475</th>
<th>DEF ART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The definite article is obligatory</td>
</tr>
</tbody>
</table>
Metalanguage

- The ultimate goal of the project is not to construct a single database that contains all available linguistic information.
- Create a system that allows the user to query simultaneously several typological databases from a single gateway.
- In order to reach this goal, three different issues need to be solved:

1. metalinguistic (different terminology)
2. data-structure (primary/textual vs. secondary/analytic)
3. software (different software used: MS Access, SQLserver, 4th Dimension, etc)
Metalanguage

- Development of a linguistic metalanguage which should overcome the differences attested in the various databases at:
  1. linguistic level (i.e., different terminology)
  2. data-structure level (i.e. same data expressed in different formats such as texts, examples, glosses, variables).

- Development of a computational metalanguage to exchange information between the various databases and between the databases and the web: XML or JAVA (JDBC).
The Linguistic Metalanguage

- The linguistic metalanguage should contain a formalized model of the ‘world of linguistic typology’ such as general terminology and synonyms alternative definitions for central notions.
- Creation of a list of the relevant terminology (for the moment only in the area of agreement) in order to define a correspondence among possible synonymous uses of the same term in different research traditions.
- It is also necessary to find a way to relate different data structures since the same information is encoded in different ways in the various databases:
  1. Text translation supported by glosses (morphological, semantic, etc.)
  2. Information about the existence or the absence of phonological, syntactic, semantic relations in a given language.
• The system must be able to understand, retrieve and, possibly, combine the information belonging to different modules of the grammar and involving different structures.

• There are two possible ways to achieve this goal:

  1. The system gives a first answer containing general information about the relevant language. For more specific (linguistic, areal, statistic) information a series of links are provided to the databases containing the desired data.

  2. The different data structures can be related by creating a dictionary of concepts together with a link to the various data structures which might be employed to express these concepts.

| subj-verb agreement | variable X | SUBJ-AGR = VERB-AGR |
Evaluating the metalanguage approach

- Advantages:
  1. redundancy is avoided since the requested data is retrieved simultaneously from all the relevant databases.

- Problems:
  1. the user might not have full control over the data.
  2. the metalanguage will inevitably impose certain choices to the user including a simplification of the original definitions.
The OLAC initiative

- A new proposal has been launched during the ‘Linguistic Web-archive’ Workshop held in Philadelphia in December to create an Open Linguistic Archive Community (OLAC): www.language-archives.org
- Aim of the initiative is to create a web archive with different types of linguistic resources, such as grammars, databases, corpora, dictionaries, video and audio recordings.
- Metadata records will be developed to describe the various archives.
- An harvesting protocol will allow the user to retrieve information about where the data he is looking for can be found.
• In order to become members of OLAC, it is necessary to comply with two standards:

  1. Implement the OAI (Open Archive Initiative) metadata set
  2. Implement the OAI harvesting protocol (www.openarchives.org).

• A group of alpha testers will be identified who are committed to building a conformant archive during the first half of 2001.

• The Linguist List will act as service provider. A service provider uses the metadata harvesting protocol to collect metadata from participating archives and build a database which forms the basis of their service.
What is Metadata?

- The simplest definition of *metadata* is structured data about data.
- Library catalogs represent a well established variety of metadata that has served for decades as collection management and resource discovery tool.
- Dublin Core metadata is a standard specifically intended to support resource discovery.
- The elements identified represent the core set that is likely to be widely useful to support resource discovery.
- The Dublin Core Metadata elements are defined according to 15 attribute-value pairs.
The Dublin Core Metadata Element Set

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights
An alternative approach

- An architecture similar to that developed for OLAC can be envisaged for the ‘Typological Database’ project.
- The databases will not be linked together but metadata records should be developed to provide information about the status and the nature of the data they contain.
- Through the harvesting protocol the user will be able to retrieve information about where the data he is looking for can be found.
- Each database should be on-line.
Evaluating the metadata approach

- Advantages:
  1. The user will have full control over the data since he will explore each database individually.
  2. Participation in the OLAC initiative as alpha tester.

- Problems:
  1. This process might be redundant and time consuming.
Comparing and integrating the two approaches

- The difference between the two approaches is evident: while the former will provide the user with data, the latter will only provide metadata, that is information about how and where the requested data can be found.

- The two approaches can be combined, since the metadata approach can be viewed as first step towards the creation of a typological archive.

- Metadata can be developed for the description of typological archives. The *subject* field could become rather specific so that it might encode the same concepts developed for the metalanguage.