The corpus as a database: Towards a multifactorial typology of clause linkage

Volker Gast, Balthasar Bickel

Friedrich-Schiller-Universität Jena, Universität Zürich

31.08.2011
Introduction

- Typological databases contain information about languages in terms of a specific typological dimension
- Examples: Word order (Dryer/WALS), agreement, syncretism, suppletion (Surrey Morphology Group), intensifiers and reflexives (FU Berlin/Jena), etc.
- Descriptive databases as ‘grammar fragments’ (or ‘chapters’) (cf. Gast 2009)
The LTRC initiative

- LTRC-programme (*Language Typology Resource Centre*), sponsored by the EU (6th Framework) and coordinated by M. Everaert (Utrecht)

- Establishment of a network of typological databases (cf. the more recent Typological Database System/TDS, http://languagelink.let.uu.nl/tds/)

- *Typological Database of Intensifiers and Reflexives*: Published in 2003 (version 1.0)


- Objective: Make data gathered in a typological research project available to the public
The data in TDIR

- Data is stored in a MySQL database, accessed by PHP-pages
- Two types of entities, **LANGUAGES** and **EXAMPLES**
- Intensifiers/reflexives as properties of languages
- Consequence: Properties of intensifiers are (technically) properties of languages
- Sentences are linked to languages
Data contained in TDIR

- Information on 101 languages, with 689 glossed examples
- Some basic information about the languages (areal, genealogical)
- A list of glosses and references
- A lot of prose explanations (descriptive focus)
The extensible linguistic database system (XLD)

- XLD: *Extensible Linguistic Database* system, developed by A. Dimitriadis and F. van Vugt (Utrecht)
- Emerged from a joint research project on reciprocity carried out at the universities of Utrecht and Berlin (E. König, M. Everaert, A. Dimitriadis, V. Gast)
- Flexible/dynamic/extensible typological database system that can be adapted during data input (cf. the AUTOTYP method of Bickel, Nichols)
The Berlin-Utrecht Reciprocals Survey

- Based on the XLD system
- [http://languagelink.let.uu.nl/burs/db-internal/login.php](http://languagelink.let.uu.nl/burs/db-internal/login.php)
- Three entities: **ANSWERSETS**, **MARKERS** and **SENTENCES**
- Information on sources, varieties etc. (attributes of answersets)
- Some usefuly features, e.g. user management, value control, input of glossed examples, etc.
- Dynamic system: new attributes (questions) and values (answers) can be added (cf. the AUTOTYP method)
- Database mirrors questionnaire (actually is a questionnaire in the input mode)
- Rich search interface
The data in BURS

- There is a public and a private version (some datasets are not fit for publication yet; public access is a property of an answerset)

- Private version:
  - 168 answersets
  - 374 markers
  - 2278 glossed examples

- Public version
  - 110 answersets
  - 216 reciprocal markers
  - 1554 glossed examples
Major types of subordinate clauses

- Four major types of traditional grammar

<table>
<thead>
<tr>
<th></th>
<th>nominal projection</th>
<th>verbal projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjunction</td>
<td>relative clause</td>
<td>adverbial clause</td>
</tr>
<tr>
<td>complementation</td>
<td>nominal complement clause</td>
<td>verbal complement clause</td>
</tr>
</tbody>
</table>
The typology of clause linkage

- Major parameters of variation:
  - Type of dependency (adjunct, complement)
  - Attachment site (verbal, nominal)

  - Autonomy/integration (hierarchical downgrading, syntactic level)
  - Expansion vs. reduction (desententialization of subordinate clause, grammaticalization of main verb)
  - Isolation vs. linkage (interlacing, explicitness of linking)
Towards a probabilistic typology

- New challenge in typology: Identify probabilistic typological patterns (e.g. Bickel 2007, Bickel et al. 2009)
- Cooperative project: A multifactorial typology of clause linkage (Zurich/Jena)
- Quantitative/corpus-based typology
- Determine correlations between (ideally atomic) variables, ‘distil’ types of constructions by applying (multifactorial) statistical methods (bottom-up).
Dimensions of typologizing

- Holistic vs. **parametric**
- Aprioristic vs. **emergent**
- Theory-driven vs. **data-driven**
- Categorical vs. **probabilistic**
The corpus as a database

- The corpus only provides information about exemplars.
- Exemplars can be described in terms of typological parameters of variation.
- Emerging types can be determined on the basis of statistical distributions.
- Prerequisites
  - Richly annotated corpora
  - Methods of extracting types from corpus data
Typologizing connectives

- Properties of the connectives
  - Semantics: Range of interpretation
  - Syntax: Position relative to subordinate clause
  - Morphology: Internal make-up
  - Morphosyntax: Inflection, agreement
  - Pragmatics: Information structural properties, e.g. definiteness

- Properties of the contexts in which connectives occur
  - Tense, aspect, mood
  - Finiteness properties of the subordinate clause
  - etc.

- A corpus-based approach allows for a systematic investigation of the distributional properties of subordinators
The adverbial subordinators of Tzotzil (Maya)

- Most frequent items
  - *k’alal*: temporal
  - *yu’un*: causal
  - *sventa*: purposive
  - *yo’*: purposive
  - *mi*: conditional
Important features of Tzotzil subordinators

- Subordinators may be combined (e.g. temporal and conditional operators).
- Subordinators may take a ‘subjunctive’ suffix (‘emotive inflection’).
- Subordinate clauses may be definite or indefinite (property of the subordinator or of the clause?)
Combinations of subordinators: *k’alal mi*

- Subordinate clauses may be both temporal and conditional

(1) *K’alal mu to cham-em-uk li hka’-e,*
when NEG still/yet die-PERF-SUBJ DET my.horse-CL
...

‘When my horse was still alive, . . .’

(2) *K’alal mi ch-cham chkiltik ti htottik-e,*
When Q.Pol he.dies we.see DET our.father-CL
‘When/if we see that our father (the sun) dies, . . .’
Emotive inflection

Subordinators may take (subjunctive) suffixes that are otherwise found on verbal and adjectival predicates (‘emotive inflection’)

(3) $K'u$-uk cha’al-uk mi yantik x-hel  
Q.Wh-SUBJ how-SUBJ Q.Pol piece.by.piece ICP-change  
ti h-vo’ne h-kostumpre-tik-e  
DET 1POSS-old 1POSS-tradition-PL-CL  
‘Even though our old traditions are changing . . . ’
A tree structure

\[
S_1
\]

\[
\text{INDIF.EMOT}
\]

\[
S_2
\]

\[
\text{INDIF} \quad \text{SUBJ}
\]

\[
S_3
\]

\[
\text{SUBORD} \quad \text{mi}
\]

\[
yantik \ xhel \ ti \ hvo’ne \ kostumpretike
\]
Definiteness of subordinators: \textit{ti k’alal/manchuk/mi}

- Most subordinators are used with as well as without definite marker.

(4) \textit{Ti k’alal chlok’ x-ch’ulel li hbankil-e...}
DET when goes.out 3POSS-soul DET my.brother-CL
‘In the moment when my brother died, . . . ’

(5) \textit{Ti manchuk yakub-em-ot-e...}
DET COND.CTF drink-PERF-2ABS-CL
‘If you had not drunk, you would be ok now.’
How to capture such behaviour in a database

- All of these properties pose non-trivial problems for a type-based database
- Where and how are ‘composite subordinators’ stored/described?
- Are ‘inflected subordinators’ entries in their own right or is it a property of (specific) subordinators that they may take subjunctive inflection?
- How are definiteness properties captured?
- More economic and flexible (in terms of data structure): Annotating sentences directly and extracting generalizations on a quantitative basis (generate databases on the fly).
Definiteness and temporal clauses

- Temporal clauses with a definiteness marker (invariably?) refer to specific moments that are (invariably?) located in the past.
- Correlations with specific context features are expected, e.g. tense/aspect, person categories, information structure (ordering of clauses), etc.

(6) *Ti k’alal chlok’ xch’ulel li hbankile, . . .*  DET when exit his.soul DET my.brother
    ‘(At the moment) when my brother died, . . .’

(7) *K’alal chlok’ xch’ulel li hbankile, . . .*  when exit his.soul DET my.brother, . . .
    'While my brother was dying, . . .’
Subordinators and emotive inflection

- Emotive inflection is hard to capture descriptively, but it is expected to correlate with specific context features, e.g. person features (1st person?), lexical items (empathy), etc.

(8) *K’alal chive’-e, chinop*
when 1.eat-CL 1.get.full
‘When I eat, I satisfy my hunger.’

(9) *K’alal-uk chive’-e, lah hti’ kok’.*
when-SUBJ 1.eat-CL CP 1.bit my.tongue
‘When I ate, I bit my tongue.’
Towards typological generalizations

- Annotated corpora allow for language-specific distributional analyses
- From a crosslinguistic point of view, they allow for estimating typological patterns such as:
  - (dis)similarities between types of subordinators or subordinate clauses;
  - associations between linguistic variables, or between linguistic variables and families or areas, etc.
- Example of a generalization: Subordinators with similar meanings are expected to be associated with similar distributional properties
- Can the readings of underspecified subordinators (e.g. Lat. *cum*) be predicted in this way?
Literature