Toward a database of classification systems.
Semantic description and analyses

PALC 2007 (Lodz)
April 21, 2006

Christophe Coupé (CNRS)
Laboratoire Dynamique du Langage (CNRS – University of Lyon)
Overview

1. Why build a database of classification systems?
2. Structure of the data
3. “Efficient” approaches to semantic description
4. Examples
Overview

1. Why build a database of classification systems?
Classification: a window to the mind

- Definition: a mental operation that causes an object or a multitude of objects to fall under a concept X (Seiler, 1986)

- Numeral classifiers…
  - e.g. in Japanese: 三匹ねこ (3 - Cl for animals - cats)

- … but also “graphemic classifiers” in written scripts…
  - eg. in Ancient Egyptian

- … or in some Chinese texts like the Leishu(s) (1 heading specifying the topic(s) of a set of texts)…

- Classification helps to understand how different cultures & languages divide the world into categories
  - “Language as a knife”
Methodological aspects

- **Concrete questions:**
  - how is “turtle” conceptualized in different cultures with respect to other animals or concepts?
    - Along with mammals? Fishes? Something else?
  - In a language, how do a set of classifiers divide a semantic domain, overlap, leave empty regions, cope with new entities, evolve etc.?
  - How do languages resemble and differ?

- **To answer such questions:**
  - Need to efficiently browse a large number of instances to extract recurring features
  - Get access to data from different languages the expert in Chinese / Maya / Korean etc. is not familiar with
Relying on computer software

- Build a computerized database of classification systems
  - Take advantage on automatic analyses and fast searching in large amounts of data
  - Put already existing “digital knowledge structures” into use (e.g. “digital” ontologies)

- A “bridging” perspective to address general and therefore difficult (otherwise too difficult?) questions

- One of the aims of the European COST A31 Action on classification systems (Head: T. Wiebusch)
Overview

1. Why build a database of classification systems?

2. Structure of the data
The database in a glance

- **Software: Filemaker 8.5**
  - **Pro:** tools to design layouts, scripts, security, web, multi-user & remote access
  - **Con:** less flexible than other options (PHP, SQL etc.)

- **A classical dilemma:**
  - a. Need of detailed description vs.
  - b. generalization & simplification for the sake of automatic analyses & queries
  - a dual “textfield + structured data” approach

- **Guideline:** “goal-orientedness”
  - Think to cross-linguistic comparisons & automatic analyses
  - Develop “fast” procedures to enter data
The "Combination" layout

Tabbed panels to "compress" the display

Reusable and parameterized scripts

Protection against unwanted modifications
The issue of semantic descriptions

- To address the building of concepts in different languages through classification:
  - Need to provide semantic descriptions of the CIs, CEs and Combinations
  - Solve the opposite constraints of:
    - Cultural relativism
    - Cross-linguistic comparisons
  - Keep in mind “user-friendliness” & automated analyses

→ Some considerations and choices
Overview

1. Why build a database of classification systems?

2. Structure of the data

3. “Efficient” approaches to semantic description
“Digital” ontologies

- Designing (universal) ontologies to “sort out the world”
  - A since long chased dream… with mixed success

- Digital ontologies:
  - Efforts to encode (large or specific) sets of (abstract or concrete) entities and their relationships in a fashion prone to automatic treatment

- Main idea:
  - The previous entities may be used to describe the main components of our database = they may be used as “SEMANTIC DESCRIPTORS”
  - The relationships between them may be relied on during analyses
Dove in Akkadian

| Dove | Divinity/divine | Evil |

Part of a relevant ontology

In the Wordnet ontology:

Bird
  - Beak
  - Feather
  - Wing

Vertebrate
  - Tail
  - Digit

Columbiform bird
  - Pigeon
  - Dove

Gallinaceous bird
  - Comb

Combiformes order
  - Australian turtledove, turtledove
  - Turtledove
  - Morning dove, Zenaidura
## Comparing digital ontologies

<table>
<thead>
<tr>
<th></th>
<th>Wordnet</th>
<th>SUMO/MILO</th>
<th>Framenet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very large (&gt;117,000)</td>
<td>Middle size (-)</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>Rather well adapted</td>
<td>SUMO: rather abstract entities with respect to</td>
<td>MILO: (slightly) too small</td>
<td>?</td>
</tr>
<tr>
<td>to entities</td>
<td>classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>met in preliminary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>investigations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexically-based (English)</td>
<td>Language-independent (?) (+)</td>
<td>Lexical</td>
<td>?</td>
</tr>
<tr>
<td>Simple to import in</td>
<td>Requires parsing the KIF language (-)</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>Filemaker (+)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many ontological</td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>relationships (+), but</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>many are missing too</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can be expanded with</td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>additional relationships (+)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Self-organizing emergent” ontologies

- General ontologies will never be perfectly suitable to our specific purposes
  - “Why need 10,000 or 100,000 semantic descriptors if one focuses on the domain of birds?”
    - hard to navigate, to select the right entity (too many or not a single one, redundancy) etc.
  - Which Wordnet/SUMO entities to choose to describe a CI, CE or Combination in the database?

- Build a set of semantic descriptors from the data themselves
  - Would be “perfectly” adapted to the data
  - Relevant since the extracted knowledge relies on the raw data
Self-organizing typologies
(B. Bickel’s AUTOTYP)

- **Procedure for a given linguistic aspect**
  - Study 2 languages, and describe them with a minimal set of “typological descriptors” (1 only)
  - Add a 3rd language; if the existing typology doesn’t account for it, refine it
  - Consider a 4th language etc.
  - Usually stabilizes after a few dozen languages

- **Is it applicable to classification systems?**
  - The classified domains are *highly dimensional*, contrary to the previous case → extremely difficult to describe in an emergent fashion
  - Self-organization required *within* and *between* languages

  - The amount of collaborative work implied seems much too high
  - Pre-select entities to ease the study of a specific conceptual domain?
Beyond simple sets of semantic descriptors (1)

A set of Wordnet entities can be supplemented by an additional “syntax”
Beyond simple sets of semantic descriptors (2)

- Full logical formula (with binary connectors)
  - CI: (bird AND (NOT pigeon) AND flying) OR bat OR (fish AND flying)
  - A trade-off
    - Much more precise, and perhaps necessary in some complex cases?
    - But also much more difficult to analyze (esp. within Filemaker)

- Unary modifiers:
  - “cultural” versus “universal”
  - A degree of confidence (eg. “-” vs. “supposed to be”)
  - Extension / scope (eg. “-” vs. “similar to”)
  - “Not”
  - “Part of”, “Has”, etc.

- Some investigations suggest
  - to rely on formula only for CI = final description after the study of occurrences
  - that formula without OR are (logically) enough for CEs & Combinations
Beyond simple sets of semantic descriptors (3)

- An artificial but possible example

<table>
<thead>
<tr>
<th>Turtle in Ancient Egyptian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turtle</td>
</tr>
<tr>
<td>(Has a) [Carapace AND Black]</td>
</tr>
<tr>
<td>(culture-specific) Divinity</td>
</tr>
<tr>
<td>(supposed to be – culture-specific) Big</td>
</tr>
<tr>
<td>(similar to – culture-specific) Fish</td>
</tr>
</tbody>
</table>

The main concept

Redundant with knowledge already in Wordnet, but not really a problem

Q: Give me all the CE similar to Fish

Neutral or favorable

Such a description opens the door for many further semantic analyses, e.g. all CI classifying according to size
Overview

1. Why build a database of classification systems?

2. Structure of the data

3. “Efficient” approaches to semantic description

4. Examples
This is always right now that bugs occur…
Perspectives & Conclusion

- We are now completing a tool to:
  - Compile, search and analyze large amounts of data in a reliable and efficient way
  - Investigate the classification in a cross-linguistic fashion

- Some choices made may not be the best theoretical ones, but practicality imposes some limits

- Next steps:
  - Enter data from various languages and scripts
  - Try to focus on and reasonably cover a few fields like a subset of “fringe” animals or tools
  - Perform analyses and develop additional tools to this end
Thank you for your attention,

Comments welcome 😊

- Acknowledgment: special thanks to N. Allon, O. Goldwasser, G. Selz, K. Wagensonnner, T. Wiebusch