Toward a database of classification systems
Semantic descriptions and analyses

Christophe Coupé
Laboratoire Dynamique du Langage
CNRS – University of Lyon

Email: ccoupe@ish-lyon.cnrs.fr
Address:
Christophe Coupé
Laboratoire Dynamique du Langage
Institut des Sciences de l’Homme
14 Avenue Berthelot
69363 Lyon cedex 07
FRANCE

Classification, which could be defined as “a mental operation that causes an object or a multitude of objects to fall under a concept X” (Seiler, 1986), is linguistically expressed in many languages of the world, from Mandarin, Japanese or Burmese to languages of the Bantu or Amerindian families. It also appears in writing systems, for example in the ancient cuneiform, Mayan or hieroglyphic scripts. Although scholars disagree on the exact scope of this phenomenon, and the terms that should be used to refer to its various expressions – classifiers, semantic determinatives etc., these disagreements are partly due to the separate histories of since long established scholastic traditions, while “cross-field” investigations clearly point to the existence of common underlying cognitive and linguistic strategies.

In order to get a better grasp on the previous strategies, and to propose a better unified approach to the notion of classification, the European COST Action A31 gathers experts from various fields and linguistic domains. Among other targets, the development of a multi-language database is a joint effort to integrate and analyze the numerous data that can be provided by participants. To this end, a generic framework has to be agreed upon before data can be entered in an efficient and purpose-driven way. Among other constraints, the structure of the database should:

- both preserve the fine details that characterize each instance of classification and reduce the data to allow cross-comparisons and extraction of general principles - this is achieved by a dual system of description, with both textfields and more constrained data structures;
- pay precise attention to phonetic and written forms, with possibility to enter various scripts, variable forms etc.;
- describe the syntactic contexts of the instances of classification;
- describe the semantic properties of the operations of classification.

This last requirement is of special significance since it underlies possible cross-comparisons and the depth of the analyses that could be performed in and between various classificatory systems. Given the issues raised by the idea of universal ontologies, the most significant problem is the definition of a set of shared entities that can fit the cultural and linguistic specificities of each system. How large would/should this set be? Should it be built incrementally, following the principles of recent emerging and self-organizing typologies (cf. the AUTOTYP project), or should it be derived from existing ontologies like SUMO-MILO or Wordnet? Which adaptations would then be necessary? How does one solution or the other constraint possible analyses by easing some and preventing others?

In this talk, after describing the ambitions and general structure of the database, we will focus on several alternatives for semantic description, from the direct or refined use of Wordnet to emerging systems. In each case, we will provide examples from various sources and estimate the consequences in terms of further large-scale analyses. All in all, we hope to stress the richness and pitfalls of such a database project, and more generally the use of computational tools in (cognitive) linguistics.

(494 words)