International Conference "Contrastive studies" and the MONDILEX Project Open Workshop *„Representing Semantics in Digital Lexicography"*, 2009 June 29th – July 1st, Pałac Staszica PAN, Warszawa.

Interactive Discovery of Ontological Knowledge for Modelling Language Resources

André WLODARCZYK
CELTA - Centre de Linguistique Théorique et Appliquée
andre.wlodarczyk@paris-sorbonne.fr

Abstract

There is a constantly growing demand for a 'deeper' semantic description of natural languages. Indeed, in order to properly differentiate various linguistic units from each other, it is necessary to define these units with more *specific* (fine-grained) sets of high (*viz.* adequate and consistent) *quality* feature structures. Both above problems (granularity and quality of descriptions) are strictly interwoven.

The computer scientists who proposed many different approaches (algorithms and data structures) creating the Natural Language Processing framework have adopted most linguistic notions (or even complete theories) without paying due attention to the need for their logical reconstruction. For this reason, in order to remedy for this and develop new lexicons, we propose the approach which follows the discovery procedure from "raw" data to structures.

Following some logicians (MacCarty J., Barwise J. & Perry J., Wolenski B.) and those computer scientists who are involved in modelling of the semantic web and its ontological foundations, we claim that linguistic signs inherit their properties from multiple *ontologies*. Some of them specifically concern language itself (ex. parts of speech, genders, etc.), the others refer to the world. For example, verbs inherit their properties at the same time from phonemic structures, valence schemas, roles, situation frames, etc. It is therefore necessary to build a number of local meta-ontological (universal) mono- and multi-base hierarchies of concepts which underlie particular language-specific cases.

Because knowledge acquisition using the Knowledge Discovery in Databases (KDD) technology is situated halfway between Database Management and Automated Discovery, we claim that it is computationally possible to reveal, from a very simple tabular representation of gathered *atomic* data, usually "invisible" ("hidden") remarkably compound relations. KDD technology makes it namely possible (a) to transform tabular representations (or charts) into lattices (which are more powerful than trees because they allow multi-base inheritance), (b) to apply approximation techniques allowing to reason with uncertain data and (c) to provide hierarchical analyses reflecting the mutual dependencies of data in the system.

To start, let us formally define the linguistic **sign** as a structure with ***Usages*** as a set of objects (mophemes), ***Descriptions*** as a set of propositional formulae and ***Assignment*** as an assignment function from descriptors to usages:

Sign = <***Usage, Description, Assignment***>.

On the other hand, semion will be defined as a formal concept (a pair of a subset of usages (***M*** ⊆ ***Usages***) and subset of descriptions (***Δ*** ⊆ ***Descriptions***):

Semion = < ***M***, ***Δ*** >.

Lexicons and dictionaries were, in the history of mankind, the first attempts at using language resources for annotation and translation purposes. Among them, thesauri are the most structured collections of words. However, due to the intrinsic polysemy of signs, thesauri cannot but very approximately capture relationships between signs. For this reason, *dynamic* semantic maps and semantic lattices among others will be useful both as well during the description research and development stage as in the future utilisation of computerised dictionaries.

- **Semantic Map (S-Map)** - a set of similar signs (with attributes arranged by *opposition* relationships).
- **Semantic Lattice (S-Lattice)** - a set of signs (with attributes arranged by *entailment* relationships).

Thus, the meaning conveyed by natural languages is defined as a function from signs (in fact, from their schematic or partial semantic representations) into the individualized ontologies. We will keep in mind therefore that any description of a natural language semantic domain and the representation of local domain ontologies must match.

The SEMANA software consists of two sorts of operations : (1) creation and dynamic maintenance of the database and (2) KDD proper algorithms for both symbolical and statistical data analyses.

(1) **Data Base Builder** : database construction environment with facilities for dynamic restructuring of data - Editor of Records, Tree Builder Assistant and Attribute Editor.

(2) **SEMANA Editor** : This is the monitor of SEMANA in which it is possible to open a file, create a file, edit a file as well as to discover similarities and analogies useful for building semantic fields etc.

a) Symbolical Data Analysers

- Formal Concept Analyser - FCA (cf. R. Wille 1982, 1997; B. Ganter and R. Wille 1999)
- Rough Set Analyser - RSA (cf. Pawlak Z. 1982)
- Formal Rough Concept Analyser – FRCA (cf. J. Saquer and J. S. Deogun 1999)
- Rough Decision Logic Analyser – RDLA (cf. Bolc, Cytowski and Stacewicz 1996)

b) Statistical Data Analysers STA 3

- Factor Correspondence Analysis (J.-P. Benzécri)
- Ascending Cluster Analysis (J.-P. Benzécri)

As a sample solution, let us first state that morphemes are *opposed* by pairs of similarity and distinction (see definition of semion above). Structural linguists proposed 3 kinds of oppositions: *privative* (binary), *equipollent* (multi-value) and *gradual* (degree-value). The interactive research in the KDD framework allowed us to discover two special types of linguistic binary oppositions: a **double converse opposition** (±A ⇄ ∓B) and **a double binary opposition** (+A → -A and +B → -B).

Of course, in both cases, there are only two morphemes in question. In the double converse opposition the morphemes are *infomorphic* (a special kind of isomorphism proposed by Barwise J. & Seligman J.). The capitals A and B represent binary attributes which are converse of each other (*viz.* +A = -B and +B = -A) in the double converse opposition, and they represent two different attributes (*viz.* +A ≠ -B and +B ≠ -A) which belong to the same hierarchical domain in the a double binary opposition.

One interesting and original goal of the interactive research in linguistic semantics is building data banks of both ontological and linguistic knowledge structures. Such structures could be accessed by definitions composed in natural languages using parsing mechanisms enhanced with powerful approximation functions.