Journal : 日本語科学 **(Japanese Linguistics)**, **Volume :** 21, edited by 国立国語研究所 **(The National Institute for Japanese Language),** April 2007, Tokyo

Université Paris Sorbonne (Paris 4)
CELTA - Centre de Linguistique Théorique et Appliquée

- CASK (Computer-aided Acquisition of Semantic Knowledge) Project -

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

Keywords:

(1) Theory of Language (language modelling, sign, semantic field), (2) KDD : Knowledge Discovery in Databases (Decision Logic, Formal Concept Analysis, Rough-Set Theory, Cluster Analysis, Factor Analysis), (3) DBMS : Database Management Systems (software engineering), e-dictionary

1. Introduction

The initiative of Computer-aided Acquisition of Semantic Knowledge **(**CASK) is a research program of the Centre for Theoretical and Applied Linguistics (CELTA), an institution of Paris-Sorbonne University http://www.celta.paris-sorbonne.fr (with 66 active researchers and as many doctorate students) working on 10 European languages and Japanese. CELTA was established in 2000 under its first label "Forme-Discours-Cognition" by Hélène Wlodarczyk, professor of Slavonic linguistics. The centre has been officially recognized by the French Ministry of Education and Research in 2002. CELTA is divided into six research teams. CASK is the research program of the 3rd team : *Semantic Analyses* under the guidance of the author.

Computer-aided Acquisition of Semantic Knowledge is aimed at describing a number of semantic fields of a few European languages using techniques of KDD (Knowledge Discovery in Databases). Hence, the project is interdisciplinary by its nature and assumes scientific cooperation of experts and specialists in the fields of Theoretical Linguistics and Information Technology. The task of linguists consists in an interactive discovery of ontology-based hierarchical definitions of Feature Structures using SEMANA (Semantic Analyser) software designed especially for this project in order to build linguistic databases of semantic knowledge. SEMANA, presented as a platform putting together several KDD algorithms, is being designed and implemented in *Transcript*[1] by Georges Sauvet and André Wlodarczyk.

The CASK Project is the first attempt of applying computational (including approximation-based) methods in order to determine the relevance and the relative importance of descriptions. Only very detailed formal descriptions of different languages gathered in databases can lead to the verified (i.e. experimentally tested) and comparable cross-language definitions of semantic Concepts.

2. Linguistic and Ontological Backgrounds

If we want to reach better results in the field of semantic analysis of linguistic phenomena certain foundational concepts (notions) currently in use must be formally reconstructed. From the linguistic (more generally semiotic) point of view, semantic concepts (contents) must not be considered in separation from signs (units defined originally as pairs of Form and Content). Hence, the present approach is based on the assumption that meaning of man-made signs, as such, being inaccessible for inspection, the only reasonable solution for semantic research is modelling.

In addition, signs are ontology-based semantic objects. Ontologies are seen as motivations (hierarchically structured foundations) of semantic properties of signs. Semantics of human languages is application-domain specific (i.e.: it can capture most of all local domains). All the more, linguistic units (signs) inherit their properties from multiple *ontologies*. For example, a verb can inherit its properties at the same time from phonemic structures, valence schemas, roles, situation frames etc. Nevertheless, it seems possible to build meta-ontological (universal) hierarchies of concepts motivating particular semantic solutions.

Lexicons and dictionaries are, 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 capture inter-sign relationships but very approximately, For this reason, dynamic semantic maps and lattices (as defined below) among others should reveal useful both during the description phase of research and for the future utilisation of computerised dictionaries.

- **Semantic Map (S-Map)** - a set of similar signs (with respect to their usage types) with descriptors arranged by similarity relationships.
- **Semantic Lattice (S-Lattice)** - a set of similar signs (with respect to their usage types) with descriptors arranged by implication relationships.

Signs and their *systems* as defined above can be represented either using *formal concept contexts*[2], tables with single-valued

[1] *Transcript*, is an object-oriented programming language based on Apple Hypertalk.

[2] Wille R. (1982, 2001), Ganter B. & Wille R. (1999)

Boolean attributes or *information systems*[3], tables with multi-valued attributes.

3. Computational Tools - SEMANA Platform

Nowadays, computers make it more and more possible to view linguistics as an **experimental science**. Collecting numerous samples of usages (in databases), describing and analysing these data with symbolic and statistical KDD methods is clearly opposed to the Linguistics which emphasizes the **hypothetico-deductive power** of its metodology which presupposes only a rather poor set of examples as illustrations.

However, it must be stressed that semantic data input constitutes a hard task. At the stage of collecting and annotating linguistic data intuition of linguists (based on their own speaker's competence enhanced by their academic knowledge of a given language) cannot be avoided. But, due to the dynamic character of SEMANA, interaction between man and machine consists in creating and using lists of explicitly defined attributes which can be easily modified. This can prevent from the subjectivity and variability of human appreciation of the meaning of expressions as used in different contexts.

On the other hand, the difficulty of data input takes also its origin partly in the fact that linguistic expressions in context have also implicit meaning and entail as well presupposed as inferred knowledge. Namely, it is difficult to establish which part of the presupposed or inferred knowledge has to be taken into account in the description: very often, the part of implicit knowledge that has to be made explicit depends on the language which serves as contrastive reference. Contrasting one language with more than one is supposed to yield a more detailed description of semantic contents of their respective expression units.

The principles of knowledge discovery in databases techniques which are often enumerated in the object literature are quoted below :

- **Tasks** (visualization, classification, clustering, regression etc.)
- **Structure of the Model** adapted to data (it determines the limits of what will be compared or revealed)
- **Evaluation function** (adequacy / correspondence and generalization problems)
- **Search or Optimalization Methods** (heart of data exploration algorithms)
- **Data Management Techniques** (tools for data accumulation and indexation).

SEMANA (acronym for "Semantic Analyser") software contains a dynamic database builder and a platform which has been designed for computer-aided inquiry into the domain of ontology for sake of research on linguistic semantics. Linguists are well aware of the overwhelming complexity of their object of study. It should be stressed however that data structures must not have a complex view in order to reflect complexity of relations. The figures below show that using a lattice representation (which is even more powerful than tree representation) it is computationally possible to reveal rather compound relations which may seem invisible ("hidden") in a collection of descriptions using a very simple tabular representation.

[3] Pawlak Z. (1981), Orłowska E. & Pawlak Z. (1984)

	displace	inAir	mvCloseTo	mvFarFrom	onFoot	onSurface	withWings
se-déplacer	X						
courir	X				X	X	
voler	X	X					X
s'approcher	X		X				
s'envoler	X	X		X			X
accourir	X		X		X	X	
s'éloigner	X			X			

Table 1: Tabular view on the assignment of attributes

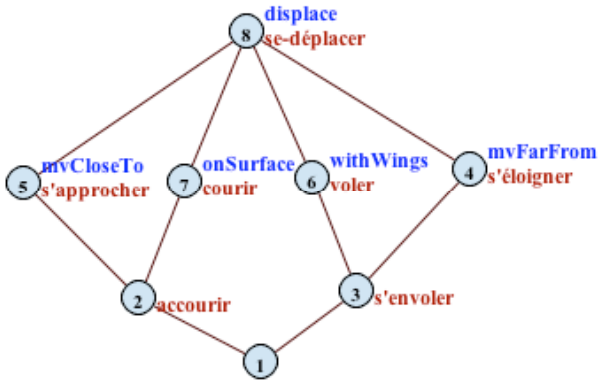

Fig. 1: S-Lattice (lattice view on the implications between attributes)

Note that the table above (tab. 1) represents a semantic description (using only 7 attributes) of 7 French verbs but the possibility of changing the view on data as in S-Lattice (fig. 1) is especially useful when a big number of rows and lines is under consideration.

3.1. Dynamic Database Management System with KDD tools

In order to conduct research on such heterogeneous objects as signs, we must collect data in a very flexible system environment. Our "db Builder" (acronym of Database Management System) has been designed especially for the purpose of research on linguistic data with little *a priori* structured knowledge. This system is suited to the semantic knowledge acquisition and experimentation. "Db Builder" makes it possible to

(1) collect samples of utterances containing a sample of the sign in question with its contextual environment, translation into other languages and free format observations in natural language

(2) describe meaning using attributes with their values (parameterized features). Sets of attributes used in collections of usages of signs are variable.

However, the number of attributes describing a category is supposed to be finite. The linguist's task is to stabilize configurations of attributes with respect to the given semantic domain ('field'). All the attributes must have their definitions in form of ontological hierarchies which constitute what is well known as 'Feature Structures'.

The proposed discovery procedure is as follows :

- collect a number of usages of one linguistic sign (or expression) and build an information system adding (when necessary) ontology-based attribute-value descriptions
- split automatically the database into as many information systems as necessary
- reduce and stabilize knowledge contained in each of the information systems
- merge fixed information systems into one huge formal concept context

The structure obtained is a semantic structural description of the linguistic unit. Among the variety of specialised KDD functions making it possible to experiment with descriptions within the attribute spaces, two particularly useful tasks consist in establishing relations between signs (as mentioned above). The automation of procedures concerning construction of S-Maps from a collection of formal concept contexts (e-dictionary) is still under study.

3.2. General architecture of SEMANA Platform

The SEMANA platform consists of two sorts of operations : (1) creation and dynamic maintenance of the database and (2) SEMANA 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
- 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)

Databases of SEMANA Vers. 2 are formatted using XML.

4. Research on European Languages with Japanese as "contrastive pivot language"

At CELTA, in the framework of the CASK project, the SEMANA platform is currently used for research on European languages. Linguists, members of the CASK project, are experts in the fields that were chosen for the first phase of research (aspect, modality and motion), as authors of monographs, papers and doctoral theses on these subjects.

The category of Aspect crosses the boundaries between grammar and lexicon. It may concern as well *semantic types of situations*[4] as expressed by verbs as those which can be expressed by a variety of grammatical and lexical devices. At present, research on Aspect in Slavonic Languages[5] (Polish and Russian) is carried in contrast with French, English and German: this allows us to compare grammatical and lexical means of expression of Aspect in two different types of languages: languages with a fairly complex grammatical verbal aspect (but more implicit noun determination system) like all the Slavonic languages, and languages with a less complex verbal Aspect but more complex noun determination system (languages with articles such as French, English and German).

At the present stage of our research, our collection of feature structures describing different usages of Aspect in utterances is not exhaustive as yet. However, while testing with SEMANA the coherence of our descriptions, we could already improve our theory of Aspect.

SEMANA databases with information concerning many different languages will make it possible to generalize the definition of Aspect : data coming from more languages may lead to add new attributes to the actual tree of Aspect ontology structure and possibly to reorganize them. This will constitute a corroboration of our actual meta-ontology of Aspect and allow for its homogenous description concerning different languages. In this respect, acquisition of new data concerning the usage of Aspect in Japanese would be welcome. The first results contributed already to the formalisation of some traditional, yet not clearly defined, notions of Aspect (called *Aktionsart*) which have been used with different meanings in descriptions of different European languages. Similar research is being carried on Modality in Polish, French and Russian languages. A first sketch of the ontological tree of Modality is already worked out but still needs to be confirmed (or even modified) by a larger amount of data.

The CASK project is based on the assumption that multilingual contrastive approach can help deepening the semantic descriptions of one language by adding and modifying features through the comparison with other languages. We also claim that contrastive approach is a good way towards the construction of an ontology that would come out from real linguistic data. The usefulness of the contrastive description is already significant for different types of European languages but the impact of this method may reveal much more important while putting all these languages into contrast with a typologically more distant language such as Japanese. Data on the Japanese language, some of them are already available in various Japanese research institutions, will be used as "contrastive pivot" for the European language. Especially, we are going to use available Japanese electronic dictionaries. In this respect, research carried by professor Ikehara's laboratory at Tottori University is a good example of successful contrastive approach: the contrast between Japanese and English led to a deeper and more varied descriptions of Japanese lexemes.

5. International Cooperation and Perspectives

The CASK research is conducted in international context on the basis of bilateral scientific contacts. A French-Polish bilateral project is currently running. The Polish team's members belong to Warsaw University (Warszawa), the Jaguellonian University (Krakow) and Śląski University while the French team's members belong to Paris-Sorbonne University (Paris 4), Charles de Gaulle University (Lille 3) and Aix-en-Provence University. The cooperation was prepared thanks to two preliminary meetings in Poland in 2004 and 2005. Officially, this CASK bilateral project started in January 2006

[4] Wlodarczyk A. (2003)

[5] Wlodarczyk A. & Wlodarczyk H. (2003, 2006)

and will be conducted until December 2007. two meetings were held in 2006, the first in May in Krakow, the second in September in Paris. On December 22nd 2006, the 3rd CASK Initiative Workshop was held at Waseda University (hosted by the Institute for DECODE and co-hosted by the Language and Speech Science Research Laboratories) in collaboration with the National Institute for Japanese Language in Tokyo. This CASK workshop was convened by Professor Harada Yasunari. Next meetings will take place in April '07 in Krakow and in September '07 in Paris. Further European development of CASK will integrate joint international research on Spanish, German and Russian languages.

The main idea of CASK initiative is to build a common bank of semantic feature structures which would be based on the ontological inquiry into a few most salient linguistic semantic fields of European Languages in contrast with the Japanese language. In Japan, some preliminary contacts have been established between CELTA's representatives and the following researchers in Japan : Aizawa Masao (senior researcher at The National Institute for Japanese Language), Arakawa Naoya (researcher at Justsystem Corporation), Harada Yasunari (professor at Waseda University), Ikehara Satoru (professor at Tottori University), Isahara Hitoshi (Research Group Leader at Japanese Institute of Communication Technology – NICT), Kashino Wakako (researcher at The National Institute for Japanese Language), Kanzaki Kyoko (researcher at The Japanese Institute of Communication Technology – NICT), Kuroda Kô (senior researcher at The Japanese Institute of Communication Technology – NICT), Murakami Yuko (associate professor at National Institute of Informatics), Antonio Ruiz Tinoco (professor at Sophia University), Yokoi Toshio (professor at Tokyo University of Technology),Yokoyama Shoichi (professor at Yamagata University of Technology) and Yoshimoto Kei (professor at Tôhoku University).

6. References (elements)

(A) Language Theory

Hartshorne Ch. and Weiss P. (eds.) (1934) *Collected Papers of Charles Sanders Peirce*, *Volume 5 : Pragmatism and Pragmaticism*, Cambridge, MA. : Harvard University Press.

Pogonowski, J. (1993) *Linguistic Oppositions*, Wyd. Naukowe UAM, Seria Językozawstwo Nr 17, Poznań, 1-136.

Putnam, H. (1975) The Meaning of 'Meaning'. Gunderson, K (ed.) *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, vol. 1, Minneapolis: University of Minnesota Press, 358-398.

Wlodarczyk, A. & Wlodarczyk, H. (2003) Les paramètres aspectuels des situations sémantiques. *Etudes Cognitives / Studia Kognitywne* V, Warszawa: SOW Publishing House, 11-34.

Wlodarczyk, A. & Wlodarczyk, H. (2006) Semantic Structures of Aspect (A Cognitive Approach). *Od Fonemu do Tekstu, in honour of Roman Laskowski*, Krakow : Lexis Pub. Co., 389-408.

Wlodarczyk, A. (2003) Les Cadres des situations sémantiques. *Études Cognitives / Studia Kognitywne* V, Warszawa: SOW Publishing House, 35-51.

(B) Knowledge Discovery in Databases (KDD)

Bolc, L., Cytowski, J. & Stacewicz, P. (1996) *O Logice i Wnioskowaniu Przybliżonym (On Logic and Rough Reasoning).* Institute of Computer Science, Polish Academy of Sciences, ICS PAS Report 822 (in Polish), 1-54.

Ganter, B. & Wille, R. (1999) *Formal concept analysis: Mathematical foundations*, Berlin: Springer.

Orłowska, E. & Pawlak, Z. (1984) Logical Foundations of Knowledge Representation. IPI-PAN, ICS PAS Report 537, Warszawa, 1-106.

Pawlak, Z. (1982) Rough Sets. *International Journal of Information and Computer Sciences*, Vol. 11, No. 5, 341-356.

Pawlak, Z. (1992) *Rough Sets: Theoretical Aspects of Reasoning About Data.* Dordrecht: Kluwer Academic Publications.

Saquer, J. & Deogun, J. S. (1999) Formal Rough Concept Analysis. Zhong, N., Skowron, A., & Ohsuga, S (eds.) *Lecture Notes in Computer Science*, Berlin/ Heidelberg : Springer-Verlag, 91-99.

Wille, R. (1982) Restructuring Lattice Theory: an Approach based on hierarchies of concepts. I. Rival (ed.) *Ordered Sets*, Dordrecht-Boston: D. Reidel, 445-470.

Wille, R. (2001) Why Can Concept Lattices Support Knowledge Discovery in Databases ? Mephu, E. N. et al. (eds.) *ICCS 2001 International Workshop on Concept Lattice-based Theory, Methods and Tools for Knowledge Discovery in Databases*. Palo Alto, CA: Stanford University, 7-20.