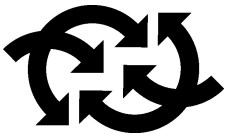


MONDILEX: Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources

Institute of Slavic Studies, Polish Academy of Sciences

Representing Semantics in Digital Lexicography

Innovative Solutions for Lexical Entry Content in Slavic Lexicography

MONDILEX Fourth Open Workshop
Warsaw, Poland, 29 June – 1 July, 2009

Proceedings

Violetta Koseska-Toszewa, **Ludmila Dimitrova**, **Roman Roszko** (Eds.)

The workshop is organized by the project
GA 211938 MONDILEX
Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources
supported by EU FP7 programme Capacities — Research Infrastructures Design Studies for Research Infrastructures in all S&T Fields

Warsaw 2009

Interactive Discovery of Ontological Knowledge for Modelling Language Resources — prolegomena —

André Wlodarczyk

CELTA — Centre de Linguistique Théorique et Appliquée
andre.wlodarczyk@paris4.sorbonne.fr

Abstract. Computer-aided Acquisition of Semantic Knowledge (CASK) is aimed at describing a number of semantic fields of a few European languages using data mining techniques elaborated within the framework of the new paradigm of computation known as Knowledge Discovery in Databases (KDD). CASK's motivation is to dig deeper in order to find building blocks which could be used in various sophisticated ways. The project is interdisciplinary involving scientific cooperation of experts in linguistics with information engineers. The task of linguists consists in an interactive (computer-aided) discovery of ontology-based definitions of feature structures using the SEMANA (Semantic Analyser) software which was designed especially in order to build linguistic databases with semantic knowledge.
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, (4) Automated Discovery.

1 Introduction

Natural languages are not like formal languages contrary to the famous statement (by Montague R., 1974): "English — as a Formal Language". All the more, formal languages imitate some functions of natural languages. Computational linguists are well aware that natural language processing needs sophisticated representation formalisms (data structures). Nevertheless, no matter how complex the representation be, the implemented system on a computer will not behave efficiently without properly designed foundations (axioms). Indeed, there is a constantly growing demand for a 'deeper' semantic description of natural languages. 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.

The computer scientists who proposed many different approaches (algorithms and data structures) creating the Natural Language Processing framework adopted most linguistic notions (or even complete theories) without paying due attention to the need for their logical reconstruction. In our approach, language is seen as a *massively distributed system* and therefore the foundations of language theory must be revisited[1] . 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 (McCarthy J. — 1989, Barwise J. & Perry J., Wolniewicz 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

[1] Nevertheless, the reconstruction of many meta-linguistic concepts must not neglect useful definitions and solutions which have been elaborated within the traditional framework of classical linguistics.

meta-ontological (universal) mono- and multi-base hierarchies of concepts which underlie particular language-specific cases.

Let us also mention that a few earlier endeavours to apply data mining technologies to language study date back to the late 1990th only (cf. Priss, U. (1998, 2000), Priss, U. & Old, L. J. (2004, 2006), Emelyanov, G. M. & Stepanova, N. A. (2005)).

2 Knowledge Discovery in Databases (KDD)

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 chart representation of gathered *atomic* data, usually "invisible" ("hidden") remarkably compound relations. KDD technology makes it namely possible (a) to transform 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.

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

(a) *tasks* (visualization, classification, clustering, regression etc.)
(b) *structure of the model* adapted to data (it determines the limits of what will be compared or revealed)
(c) *evaluation function* (adequacy / correspondence and generalization problems)
(d) *search or optimalisation methods* (heart of data exploration algorithms)
(e) *data management techniques* (tools for data accumulation and indexation).

Needless to say that in language studies, records with morphemes or expressions are seen as specimen ("raw" data) which must be described (transformed) by a fixed set of attributes. It will be easy to understand that the discovery procedure we adopted cannot be clearly split into two phases : one which is known in social sciences as *operationalisation* leading from the object (domain of interest) to its empiric model and the other known in computer science as *scaling* (mathematisation) leading from the empirical system to the representation system. Claiming nevertheless that the discovery procedure is a complex iterative process, we will elaborate[2] on this point in our paper on modelling principles (in collaboration with Stacewicz, P. — in preparation).

2.1 Computer-aided Acquisition of Semantic Knowledge (CASK)

We believe that only detailed formal descriptions of different languages gathered in databases can lead to experimentally tested and comparable cross-language definitions of semantic concepts. As the matter of fact, linguistic research using the CASK[3] The initiative of Computer-aided Acquisition of Semantic Knowledge is part of research program of the Centre for Theoretical and Applied Linguistics (CELTA) at Paris-Sorbonne University (`http://celta.paris-sorbonne.fr/`). method and its tools is perhaps one of the the earliest attempts of applying computational methods in order to determine the relevance and the relative importance of descriptions. It consists in the two following phases: **automated** exploration of texts (data extraction) and **semi-automated** (interactive) analysis of data (data mining). The basic idea of the CASK initiative is that today more fine-grained research on semantics is needed for designing new generation intelligent linguistic tools using resources with semantic properties. The methods presently selected allow to make very precise analyses using highly advanced technologies (and their combinations) such as algorithms

[2] See however the paragraph 4.1 below.
[3] 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 (Slavic, Roman and Anglo-Saxon) languages in contrast with the Japanese language.

of Decision Logic, Rough Set Theory and Formal Concept Analysis for *symbolic* data processing, on the one hand, and Cluster and Factor Analyses algorithms for *statistical* data processing, on the other hand. The above mentioned algorithms together with database building tools have been designed and implemented by Georges Sauvet and André Wlodarczyk in the SEMANA (acronym for "Semantic Analyser") software.

Thus, the CASK is a meta-theoretical framework and a software which enables experienced and trained linguists to define their own **semantic feature structures** for language semantic categories The KDD tools of the SEMANA software make it possible to verify theoretical hypotheses under the condition that a reasonably large amount of data is described. Moreover, SEMANA enables linguists to modify semantic features and their structure without loosing the accumulated data when they find it necessary (as a feedback of accumulating large databases).

2.2 Semiotic and Ontological Backgrounds

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, semantic situation frames etc. Nevertheless, it seems possible to build both abstract and concrete ontological hierarchies of concepts motivating particular semantic solutions.

In the Slavic domain, we must mention here the pioneering research by Bojar B. (1979, p. 215) in her cornerstone work on Polish motion verbs and the underlying ontological concepts: *"The elaboration of such a semantic code obviously is still to come, nevertheless it is undisputable that the main road to this kind of information language goes through the selection of its elements on the basis of as well meanings of lexical units of the natural language and expressions they make up as extra-linguistic situations because only then it will become possible to describe both contents of linguistic units of natural languages and extra-linguistic situations whose description using a natural language would be either not economical or not accurate enough."*

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 in classical linguistics). Hence, the present approach is based on the assumption that the physical meaning of signs, as such, being inaccessible for inspection, the only reasonable solution for semantic research is *modelling*.

3 Data Mining with the SEMANA software

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 Generative Grammar which emphasizes the **hypothetico-deductive** power of its methodology and 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.

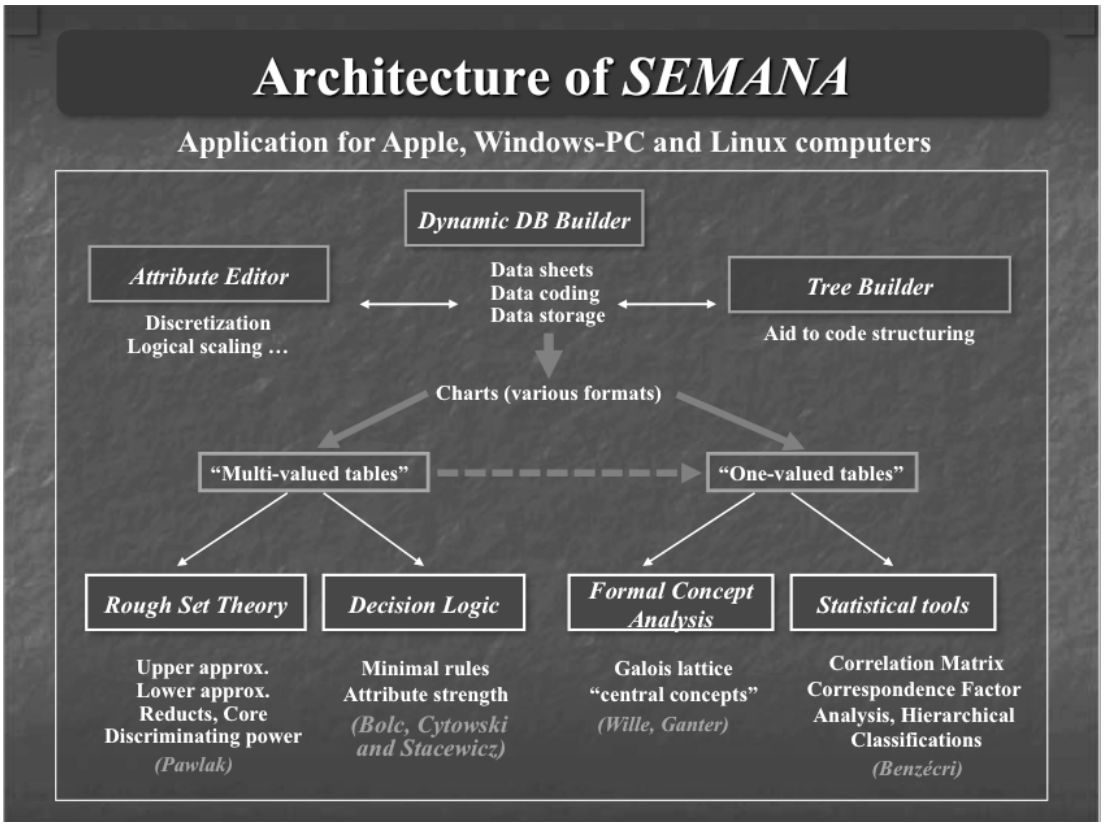
On the other hand, the difficulty of data input takes also its origin partly in the fact that linguistic expressions in context have also implicit meaning and entail as well presupposed as inferred knowledge. Namely, it is difficult to establish which part of the presupposed or inferred knowledge

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.

Using KDD algorithms gives spectacular results with adapted data. This is the case of KDD among others with rough decision algorithms implemented in the SEMANA software which contains a dynamic database builder and a software 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 chart representation.

General architecture of SEMANA Software

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



(1) **Data Base Builder** : database construction environment with facilities for dynamic restructuring of data (attribute edition assistance, tree-structure visualisation, conversion of multi-valued charts into one-valued ones, nominal and logical scaling functions, discretisation of quantitative variables, clarification of objects and attributes, co-occurrence tables (Burt's tables), etc.)

- 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.

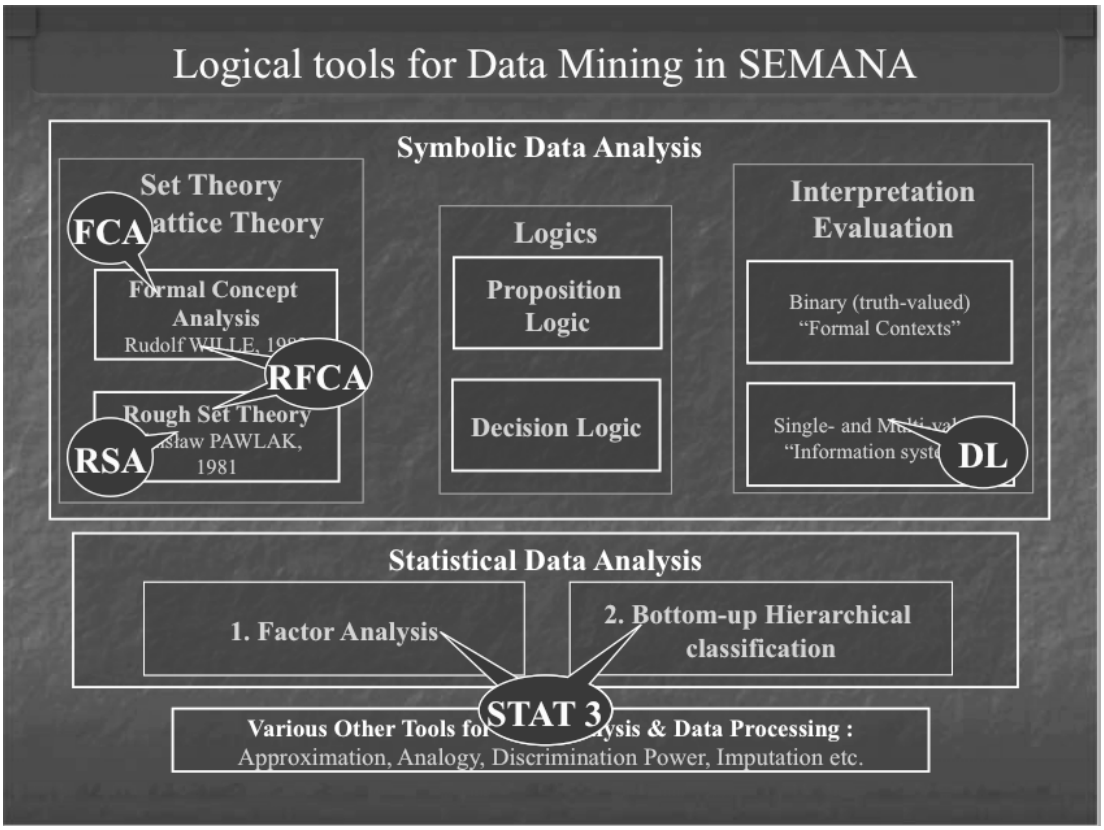
a) Symbolic Analysers

- Formal Concept Analyser — graphical representation of lattices, alpha-Galois lattices (cf. Wille R. 1982, 1997; Ganter B. and Wille R. 1999)
- Rough Set Analyser - lower and upper approximations, reducts, core, discrimination power (cf. Pawlak Z. 1982, Orłowska E. & Pawlak Z. 1984)
- Formal Rough Concept Analyser (cf. Saquer J. and Deogun J. S. 1999)
- Rough Decision Logic Analyser determination of minimal rules (for given subsets of conditional and decisional subsets of attributes), detection of inconsistenciesn(cf. Bolc L., Cytowski J. and Stacewicz P. 1996)

b) Statistical Analysers STA 3

- Factor Correspondence Analysis - including Correspondence Factor Analysis (Benzécri, J.-P. 1984)
- Ascending Cluster Analysis or Bottom-Up Hierarchical Classification (Jambu M. 1978)

and various classical statistics such as correlation matrices, similarity indices, feature matching, cluster coefficient, etc.



At CELTA, in the framework of the CASK project, the SEMANA software 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 actions), as authors of monographs, papers and doctoral theses on these subjects.

4 Interactive Linguistics

Defining consists in establishing an unambiguous *meaning* of a given concept. In another words, defining is an activity which aims at creating a formal language. The general structure of every definition is based on the equivalence relation (tautology) as established by the equality conjunction $=_{def}$ between two terms A and B. The formula $A =_{def} B$ is read as "A is the term which is being defined (*definiendum*) and B is its definition (*definiens*)".

From the syntactic point of view, the two following kinds of definitions can be distinguished: (1) *contextual*[4] definitions are composed of more than one term in their definienda; the complementary term being its 'context'**: A with respect to X, $A =_{def} B$** and (2) *direct* definitions have only one term in their definienda $A =_{def} B$. Obviously, for language studies, the contextual definitions are most likely to be attractive but, in our opinion, direct definitions reveal important as well.

4.1 Explanations guarantee accuracy

Explanation concerns the definiendum part of *definitions*. It has also two parts: ***explenans*** and ***explenandum***. In this pair of notions, it is the definiens of the definition that corresponds to the explenandum of an explanation. Explenans guarantees mutual dependencies between conjunctions of partial definitions. However, we consider that, in order to make the explenans play this function, it is necessary that the concepts which represent the explenans part of explanations be classified (ordered in such a way that they constitute tree-like structures). This remark will not astonish specialists in computer processing of natural languages since the data structures they manipulate are trees or, in better cases, DAGs (directed acyclic graphs).

Primarily, definitions are dichotomous attributes, but in most cases operationalisation is successful only while all the attributes are parameterized. The partiality (contextuality) is obtained by deduction under the closed world assumption. It is known as constraint in logic with *natural deduction* mechanisms (Gentzen). During the parameterization process attributes must be validated with respect to their belonging to the ontology of objects they represent. Attribute in parameters belong either to some unstructured clusters or to hierarchies with respect to which they must be validated; i.e.: selected from the hierarchy.

Importantly, the parameters whose attributes are coming from hierarchies always contain minus-valued (negative) attribute. Such attributes are the complements of all those which are hierarchically dependent. The next task consists in exploring the reasons (a) belonging to a tree structure or (b) being a set of attributes resulting from total combination of properties.

In order to conduct research on such heterogeneous objects as semiotic constructs, 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 (1) to 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 and (2) to describe the meaning of that sign using attributes with their values (parameterized features). Sets of attributes used in collections of usages of signs may be 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 be explained in form of ontological hierarchies which constitute what is well known as feature structures.

In the CASK framework, we propose a typical procedure for the semantic description of linguistic data.

1. initialize a set of uses of a linguistic sign (or expression) within its environment (context)

[4] From the point of view of the procedure, three kinds of contextual definitions are distinguished: (a) descriptive, (b) prospective (aiming at creating new concepts) and (c) normative. The three terms were coined by the author. They correspond to (a) reporting definitions, (b) projecting definitions and (c) regulative definitions of other athors. Cf. Pawłowski, T. (1978).

2. collect a number of uses of one linguistic sign (or expression) and build a database (when necessary add ontology-based explanation)
3. determine (step by step) the ontology of that sign (or expression) by creating attributes and establishing their constitutive (hierarchical when possible) structures
4. split automatically the database into as many information systems/contexts as necessary
5. add more uses (samples of utterances) and check the ontology quality (adequacy) with respect to the database
6. typify uses of the described signs (or expressions) using the Formal Concept Analysis.
7. check the consistency of the database
8. if possible, reduce and stabilize knowledge contained in each of the information systems using the Rough Set Analysis
9. for some purposes, merge fixed information systems into one formal concept context

The structure obtained is a semantic structural description of the linguistic unit. Let us also mention that, 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).

4.2 Logical Reconstruction of the Theory of Sign

Let us now see what are the theoretical foundations for interactive analyses of linguistic objects. From the computational point of view, following the new fuzzy and rough computing paradigm, it is easy to conclude that because signs are objects they also have granular structures. They can therefore be represented using Galois lattices. Let us then follow this viewpoint adopting the general assumption that signs (lexical and grammatical morphemes or expressions) can really be thought of in terms of granular structures. As it will be explained below, uses can be seen as granules of usages and sememes as granules of senses. Linguistic signs can therefore be described interactively using data mining technical tools such as *formal concept analysers*[5], *rough set information system analysers*[6], *ascending hierarchical classifiers and correspondence factor analysers*[7], etc.

In the sequel of this subsection, our purpose will be only to put together the notion of semiotic objects (as they are usually described in linguistic literature) and "formal contexts" as defined in computational Formal Concept Analysis in hopes that it will enable us to formalise the representational structure of signs and their uses in different contexts. As a matter of fact, in Semana, in collaboration with Sauvet G., we implemented two functions which compute centrality and priority of some formal concepts in a lattice. These functions suggest that lattices are suitable for representing linguistically motivated complex clusters of semantic structures. Indeed, below, we will endeavour to show that signs can be represented using lattices. And we hope that lattice representation of signs will reveal more adequate than DAGs of feature structures[8] . Although our research on this question is still in progress, we will sketch out the general idea we intend to develop.

Definition 1. Formally, the **Elemental Sign** is a structure with ***Uses*** U as a set of morphemes (or expressions), ***Semes***[9] S as a set of formulae or attributes or definitions and ***Assignment*** A as an assignment function from uses to semes (A: $U \rightarrow S$).

$$\textbf{Sign}=<U,\ S,\ A>$$

[5] Cf. Wille R. (1982, 2001), Ganter B. & Wille R. (1999).
[6] Cf. Pawlak Z. (1981), Orłowska E. & Pawlak Z. (1984).
[7] Cf. Benzécri, J.-P. (1984), Jambu, M. (1978) and Greenacre, M. (1983).
[8] Let us stress that features structures are explanations and what we need are definitions. Definitions can be automatically verified using data mining tools but explanations cannot.

Definition 2. We briefly introduce the **Concept**[10] defined as a pair of a subset of uses ($M \subseteq U$) and subset of semes ($\Sigma \subseteq S$). The concept must be *formal*, i.e.: it must be created by a *dual*[11] function from uses to semes and vice versa (Wille, R. – 1982).

$$\textbf{Concept} = \{M,\ \Sigma\ \} \text{ where } \{M : U \rightarrow S\} \text{ and } \{\Sigma : S \rightarrow U\}$$

Let S be a set of semes[12] $S = \{\alpha, \beta, \gamma \ldots\}$ and let Σ be a subset of semes in S ($\Sigma \subseteq S$). The **Usage** of a concept is defined as its extension.

$$[\Sigma]_{Sign} = \{m \in S : m \mid =_{Sign} \Sigma\ \}$$

Now, let M be a set of uses of a morpheme (or expression) $M = \{a,\ b,\ c...\}$ be a subset of uses U ($M \subseteq U$). The **Sense** of a concept is defined as its intension.

$$[M]_{Sign} = \{\sigma \in \Sigma : \sigma \mid =_{Sign} M\}$$

Informal definition 3. We will call ***semion***[13] the set of all the *realisations* of a given *concept*. Intuitively, while the concept is a *pair* of indiscernible usages (morphemes) and indiscernible senses (formulae), the semion is a *substructure* (substructure of the sign). Let us fix our terminology as follows (table #1):

	Form (extension)	Content (intension)
Concept	**Usage**	**Sense**
Semion	**Use**	**Sememe**

Table #1. Terminology of our theory of sign structure

Uses which have an intersection with all the items of the sense (intension) of a given concept constitute its object domain, sememes which have an intersection with all the usages (extension) of the same concept constitute its attribute co-domain. Obviously, both as well uses as sememes are distinguishable.

Thus, it is possible to consider concepts as *abstract representations* of semions. In other words, concepts should be seen as *types*[14] of semions. It should be clear therefore that our definition of semion only partially matches that of concept (in fact, defined as a formal concept) because the uses belonging to one usage are different from each other and so are the sememes belonging to one sense while in the concept the usages and the senses are indiscernible. A concept is only a pair of usage and sense whereas a semion is a substructure of a sign. In other words, due to variable contexts (*a fortiori* multiple semantic situations) of uses, linguistic signs usually contain more than one semion which are defined as a pair of usage and sense.

Moreover, both components (usage and sense) of a concept may be contained in more than one element (use and sememe respectively). Although, as we have said, the elements of every component are indiscernible within a concept, each of them may be further characterized by the

[10] In Formal Concept Analysis, the term used is Formal Concept (Wille R. — 1982, 2001).

[11] Our presentation of this problem is very succinct. The dual character of formal concepts lies at the basis of the algebric structure of lattice represenation (cf. the literature which is now very rich on FCA — Formal Concept Analysis).

[12] Note also that the original Wille's terminology significatly differs from ours because we limited our theory only to the semiotic objects. What we call semes, Wille calls attributes.

[13] Our definition of semion drastically differs from that of S. K. Saumjan. In Saumjan's Applicative Generative Grammar, the term *semion* refers to the smallest semiotic unit defined as an elementary object of the formal language designed to model the human language. The *two elementary semions* are the *name* and the *proposition* likewise in categorial grammars (Lesniewski, Ajdukiewicz). Saumjan, S. K. & Soboleva, P. A. (1973).

[14] Indeed, the idea of types as opposed to their realisations (instances) concerns semantic objects, too. In linguistics, the distinction of (formal) concept and semion is comparable to the distinction of phoneme and sound in phonology.

sememes not belonging to the usage of the concept. All the morphemes (or their homonyms) which belong to a concept are indistinguishable but each of them is different in the context of the semion.

Our model of **Elemental Sign** can be further elaborated in the two following ways: (a) internally by introducing a multi-dimensional vector space into its structure, we get then an improved differentiation of meanings (oppositions) and (b) externally by joining (formal) concepts of different elemental signs in associations; this gives rise to the definition of the **Relational Sign**.

Note also that semantic fields have the same granular structure as signs. The only difference is that the uses and usages are replaced by different words. In the case of signs, the morphemes cannot be but allomorphic. The lattices representing semions of semantic fields may contain "wholes", *viz.* concepts without name.

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 inter-sign relationships. For this reason, dynamic semantic maps and lattices we propose among others should reveal useful both as well during the research and development stage as for the future exploitation of computerised dictionaries.

- **Semantic Lattice (S-Lattice)** — a set of signs (with semes arranged by *implication* relationships).
- **Semantic Map (S-Map)** - a set of similar signs (with semes arranged by *similarity* relationships).

Thus, the meaning conveyed by natural languages is defined as a function from signs into[15] the individualized ontologies[16] . We will keep in mind therefore that any description of a natural language semantic field must match the representation of a local domain ontology. In other words, the language semantics (description) and the ontology (representation) are mutually bound. Obviously, the granularity (the scale or level of detail present in a set of data) of semantic descriptions and their ontological counterparts must match.

4.3 From "Raw" Data to Representations – Sample Solutions

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 special kinds of linguistic binary oppositions: a **double converse opposition** ($\pm$A $\rightleftarrows$ $\mp$B) and **a double** (or even **triple**) **binary opposition** (+A $\rightarrow -A$ and +B $\rightarrow -B$). Obviously, 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 within the framework of information flow by Barwise J. & Seligman J. – 1997). 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. They represent two different attributes (*viz.* +A $\neq -B$ and +B $\neq -A$) which belong to the same hierarchical domain in the a double (or triple) binary opposition.

Let us quote as examples some results obtained at CELTA (Université Paris-Sorbonne – Paris 4):

(a) the Japanese *wa* and *ga* particles have two converse binary senses each (Włodarczyk A. – 1998, 2005):

wa^{+} **Topic** / **ga**$^{-}$ **Subject**$_{old}$ $\rightleftarrows$ **ga**$^{+}$ **Focus** / **ga**$^{-}$ **Subject***new*

(b) the Polish verb past morphemes *–li* and *–ły* have two[17] senses each (Włodarczyk H. — 2009):

[15] As a matter of fact, this function from goes across the internal semantic representations.

[16] From our perspective, the semantic interpretation function of linguistic expressions should be characterised by both refinement and blending.

[17] If we consider that the neutre gender's meaning of nouns which refer to animate beings is $-A$dult, the number of binary oppositions, in this case, amounts to three (Wlodarczyk, H. – 2009).

(1) **–*li*$^+$ +Human / –*ły*$^-$ +Human**
(2) **–*ły*$^+$ +Feminin / –*li*$^-$ -Feminin**

The notation we used may be slightly misleading if one has the notion of markedness in mind. Note however that for a sign to be marked, it must not only bear a positively valued attribute. Additionally, it must be ambiguous with respect to another attribute (which presumably is situated higher in the hierarchy to which belongs the positive attribute under consideration).

5 Conclusion

At present, research on Polish aspect[18] is carried in contrast with French: this allows us to compare grammatical and lexical means of expression of aspect in two different types of languages.

The CASK method 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 or Hungarian. 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 Ikehara's laboratory (Ikehara S. – 1999) at Tottori University is a good example of successful ontology-based contrastive approach: the contrast-and-comparison of the Japanese language with English led to a deeper and more varied descriptions of Japanese lexemes.

Let us also add that 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 description composed in natural languages using parsing mechanisms enhanced with some approximation functions.

Bibliography

Barwise, K. J. & Perry, J. (1983) *Situations and Attitudes*. Cambridge: MIT Press.
Barwise, K. J. & Seligman J. (1997) "INFORMATION FLOW- the Logic of Distributed Systems", Cambridge University Press.
Benzécri, J.-P. (1984) *L'analyse des données*. Vol. 1: La Taxinomie ; Vol. 2: L'Analyse des Correspondances. Ed. Dunod, Paris, 4ème éd. (1ère édition en 1973).
BOJAR, B (1979) "Opis semantyczny czasowników ruchu oraz pojęć związanych z ruchem" (Description of Motion Verbs and of Motion-related Concepts), *Dissertationes Universitatis Varsoviensis Series*, Warszawa.
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.
Emelyanov, G. M. and Stepanova, N. A. (2005) "Semantic Relation Modeling using Formal Concept Analysis in Russian Lexical Databases", Proceeding, in *Automation, Control, and Information Technology* (489) ACIT — Software Engineering, 9-12.
Ganter, B. & Wille, R. (1999) *Formal Concept Analysis: Mathematical Foundations*, Berlin: Springer.
Gigerenzer, G. (1981) *Messung und Modellbildung in der Psychologie*, Basel: Birkhäuser.
Greenacre, M. (1983) *Theory and Applications of Correspondence Analysis*. London: Academic Press.

[18] Wlodarczyk A. & Wlodarczyk H. – 2003 and 2006.

Ikehara S. et al. (1999)「日本語語彙大系」CD-ROM 版，岩波書店 (The Japanese Lexicon), CD-ROM, Iwanami Pub. House, Tokyo
Jambu, M. (1978) *Classification automatique pour l'analyse des données*. Vol. 1: Méthodes et algorithmes ; vol. 2: Logiciels (avec M.-O. Lebeaux). Ed. Dunod, Paris.
McCarthy, J. & Hayes, P. J. (1969) "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in *Machine Intelligence 4*, ed Michie D. and Meltzer B., Edinburgh University Press (1969).
Montague, R. (1974) "English as a formal language", *Formal Philosophy*, Selected papers of Richard Montague, edited by Richmond Thomason, New Haven, Yale University Press, pages 188-221
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. (1987) "O Analizie pojec" (About the Analysis of Concepts), *Od kodu do kodu* (From Code to Code), A. Boguslawski & B. Bojar, 249-252
Pawlak, Z. (1992) *Rough Sets : Theoretical Aspects of Reasoning About Data* (Theory and Decision Library. Series D, System Theory, Knowledge Engineering, and Problem Solution), Kluwer Academic Pub; ISBN: 0792314727
Pawłowski, Tadeusz (1978) *Tworzenie pojęć i definiowanie w naukach humanistycznych* (Concept Formation and Defining in Human Sciences), PWN publishing house, Warsaw, *Begriffsbildung und Definition*, German translation by Georg Grzyb, Berlin: De Gryters, 1979, 166.
Pogonowski, J. (1993) *Linguistic Oppositions*, Wyd. Naukowe UAM, Seria Językozawstwo Nr 17 Poznań, (p. 136)
Priss, U. (1998) *Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases.* (PhD Thesis) Verlag Shaker, Aachen 1998.
Priss, U. (2000) "Lattice-based Information Retrieval*", Knowledge Organization*, Vol. 27, 3, 2000, p. 132-142.
Priss, U. & Old, L. J. (2004) "Modelling Lexical Databases with Formal Concept Analysis", *Journal of Universal Computer Science*, Vol 10, 8, 2004, p. 967-984.
Priss, U. & OLD, L. J. (2006) "An Application of Relation Algebra to Lexical Databases". *ICCS*, 388-400
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.
Saumjan, S. K. & Soboleva, P. A. (1973) "Formal Metalanguage and Formal Theory as Two Aspects of Generative Grammar" (COLING-73).
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.
Wlodarczyk, A. (1998) "The Proper Treatment of the Japanese "wa" and "ga" Particles, *Proceedings of the International Workshop on Human Interface Technology 1998* (IWHIT '98) — Aizu-Wakamatsu, Japon
Wlodarczyk, A. (2003) "Les Cadres des situations sémantiques". *Études Cognitives / Studia Kognitywne V*, Warszawa: SOW Publishing House, 35-51.
Wlodarczyk, A. (2005) "From Japanese to General Linguistics — starting with the 'wa' and 'ga' particles", *Paris Lectures on Japanese Linguistics,* Kurosio Shuppan, Tokyo
Wlodarczyk, A. (2007) "CASK — Computer-aided Acquisition of Semantic Knowledge Project", in *Japanese Linguistics*, vol 21, The National Institute for Japanese Language, Tokyo (in Japanese). English version: http://www.celta.paris-sorbonne.fr/anasem/papers/

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, H. (2009) "Lingwistyka na polonistyce krajowej i zagranicznej w dobie filozofii informatyczno-logicznej" (Linguistics in Polish Stuides home and abroad in the epoch of information science and logics), LingVaria, Rok IV (2009), nr 1 (7), Księgarnia Akademicka, Krakow, 65-79.
Wolniewicz, B. (1982) "A formal ontology of situations", in *Studia Logica* 41, 381-413.