

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

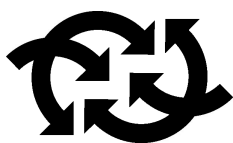
**Institute of Slavic Studies  
Polish Academy of Sciences**

# **Representing Semantics in Digital Lexicography**

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

## **Proceedings**

**Warsaw 2009**



**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



SLAWISTYCZNY  
OŚRODEK  
WYDAWNICZY

**Warsaw 2009**

Representing Semantics in Digital Lexicography.  
Warsaw, Institute of Slavic Studies, Polish Academy of Sciences, 2009.

The volume contains contributions presented at the Fourth Open Workshop “Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography”, held in Warsaw, Poland, on 29 June – 1 July, 2009. The workshop is organized by the international project GA 211938 MONDILEX Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources, Capacities — Research Infrastructures Design Studies for Research Infrastructures in all S&T Fields EU FP7 programme.

#### **Workshop Programme Committee**

**Violetta Koseska-Toszewa** (chairperson)

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland

**Ludmila Dimitrova** (co-chairperson)

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

**Peter Ďurčo**

St. St. Cyril and Methodius University, Trnava, Slovakia

**Tomaž Erjavec**

Jožef Stefan Institute, Ljubljana, Slovenia

**Radovan Garabík**

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

**Leonid Iomdin**

Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

**Maciej Piasecki**

Instytut of Informatics, Politechnika Wrocław University of Technology, Wrocław, Poland

**Adam Przepiórkowski**

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

**Volodymyr Shyrov**

Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, Kyiv, Ukraine

**Kiril Simov**

Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria

#### **Workshop Organising Committee**

**Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences**

Violetta Koseska-Toszewa & Roman Roszko

with Maksim Dushkin, Danuta Roszko, Joanna Satoła-Staškowiak

Editor of the volume and computer design: Sylwia Roszko & Roman Roszko [TEX]

Cover design: Radovan Garabík

© Editors, authors of the papers, Institute of Slavic Studies, Polish Academy of Sciences 2009

Slawistyczny Ośrodek Wydawniczy (SOW Publishing House)

Instytut Slawistyki PAN – Institute of Slavic Studies of PAS

[sow@ispan.waw.pl](mailto:sow@ispan.waw.pl)

<http://www.ispan.waw.pl/>

ISBN 978-83-89191-87-8

## Contents

Foreword .....	5
<b>Part 1. Common Theoretical Problems of Semantics</b>	
Semantic Interlanguage and Contrastive Studies..... <i>Violetta Koseska, Roman Roszko</i>	9
The Issue of Interlanguage in Contrastive Studies .....	18
<i>Małgorzata Korytkowska</i>	
Constructing Catalogue of Temporal Situations .....	24
<i>Violetta Koseska, Antoni Mazurkiewicz</i>	
Automated Extraction of Lexical Meanings from Corpus: A Case Study of Potentialities and Limitations .....	32
<i>Maciej Piasecki</i>	
Interactive Discovery of Ontological Knowledge for modelling Language Resources .....	44
<i>Andrzej Włodarczyk</i>	
Ontological Issues for Modelling Aspect and Modality Semantics .....	56
<i>Hélène Włodarczyk</i>	
<b>Part 2. Problems of Semantics and their Representaion in Slavic Digital Lexicography</b>	
Representing Semantics in the Digital Combinatorial Dictionaries of the ETAP-3 System: New Developments .....	69
<i>Leonid Iomdin</i>	
Bulgarian-Polish online Dictionary — Design and Development .....	76
<i>Ludmila Dimitrova, Violetta Koseska, Ralitsa Dutsova, Rumjana Panova</i>	
Theory of Lexicographic Systems. Part 2. ....	89
<i>Volodymyr Shyrovokov</i>	
Towards Semantic Concordances in Slovene .....	106
<i>Darja Fišer, Tomaž Erjavec</i>	
Syntactic-Semantic Treebank for Domain Ontology Creation .....	115
<i>Kiril Simov, Petya Osenova</i>	
Design of a Multilingual Terminology Database Prototype.....	123
<i>Mária Šimková, Radovan Garabík, Ludmila Dimitrova</i>	
Dictionary of Slovak Collocations .....	128
<i>Peter Ďurčo, Radovan Garabík, Daniela Majchráková, Matej Ďurčo</i>	
A Comparison of Two Morphosyntactic Tagsets of Polish.....	138
<i>Adam Przepiórkowski</i>	
Morphosyntactic Specifications for Polish and Lithuanian. [Description of Morphosyntactic Markers for Polish and Lithuanian Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)] .....	145
<i>Danuta Roszko, Roman Roszko</i>	
Description of Morphosyntactic Markers for Polish Verbs within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) .....	159
<i>Roman Roszko</i>	

### Part 3. Some Semantic Problems of Contrastive Studies of Slavic Languages and Related Topics

Lexical Functions in Bulgarian and Russian: a Sketch to Digital Comparative Lexicography <i>Svetlana Timoshenko, Olga Shemanaeva</i>	169
Translation of Polish Uninflected Present Participle in Bulgarian Literature — on the Basis of Pan Tadeusz by Adam Mickiewicz..... <i>Joanna Satola-Staskowiak</i>	180
Exponents of Adnumeral Approximation in Polish and Russian..... <i>Maksim Dushkin</i>	189
Definitions of Prepositions, Conjunctions and Particles in the Explanatory Dictionaries..... <i>Oleg Bugakov</i>	194
Quelle description pour les préverbes polonais..... <i>Ewa Gwiazdecka</i>	198
On the Lexicographic Representation of Relational Nouns..... <i>Petya Osenova</i>	205

### Part 4. Abstracts

The Lexicographic Description of Modals in Polish..... <i>Björn Hansen</i>	211
Situational and Information Structures of Discourse..... <i>Ewa Miczka</i>	213
Part of Speech Assignment as a Type of Semantic Information about a Word..... <i>Jadwiga Wajszczuk</i>	214
Facilitating Access to Digitalized Dictionaries..... <i>Janusz S. Biń</i>	215
The Confluence of the Dative and Middle Voice in Croatian And Polish..... <i>Mateusz-Milan Stanojević, Barbara Kryżan-Stanojević</i>	216
(Mini) Portraits of the Words <i>mistrz</i> and <i>uczeń</i> . Semantic Relations..... <i>Zofia Zaron, Katarzyna Drózdź-Łuszczuk</i>	217
Idiom Variability in Croatian: the Case of the CONTAINER schema..... <i>Jelena Parizoska</i>	218
RAMKI or How Verbs Were Framed..... <i>Magdalena Derwojedowa, Jadwiga Linde-Usiekniewicz, Magdalena Zawistawska</i>	219
Dictionary Sense Division and Relation to Frames..... <i>Jadwiga Linde-Usiekniewicz, Dorota Kopcińska</i>	220
Description of Verbs in Polish FrameNet Project Based on the Example of <i>İĆ</i> ('to go')..... <i>Witold Kieraś</i>	221
Some Questionable Issues of the FrameNet: the Case of the Death and Killing Frames..... <i>Magdalena Zawistawska</i>	222
<b>Authors</b> .....	223

## Foreword

This volume contains contributions presented at the MONDILEX project fourth open workshop “Representing Semantics in Digital Lexicography Innovative Solutions for Lexical Entry Content in Slavic Lexicography”, held in Warsaw in 29 June – 1 July, 2009. This workshop is organized in the framework the international project GA 211938 MONDILEX Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources, Capacities — Research Infrastructures (Design studies for research infrastructures in all S&T fields), a project developed under EC Seventh Framework Programme.

The workshop’s purpose was to study the representation of semantics, phraseology, etymology and related issues in bi- and multilingual digital dictionaries, especially for Slavic languages. The papers discuss current trends and achievements in the field of digital Slavic lexicography, especially representation of semantics and phraseology.

**Part I** of the volume, “Common Theoretical Problems of Semantics”, discusses some theoretical achievements in the field. The paper by V Koseska and R Roszko presents a semantic interlanguage for contrastive studies in the context of the MONDILEX project. The paper by M. Korytkowska discusses some interlanguage problems in contrastive studies. The paper by V Koseska and A Mazurkiewicz is dedicated to the constructing catalogue of temporal situations. M. Piasecki’s paper discusses a problem of the automated extraction of lexical meanings from a corpus, namely a case study of potentialities and limitations. The interactive discovery of ontological knowledge for modelling Language resources is considered in the paper by A. Włodarczyk. The paper by H. Włodarczyk presents some ontological issues for modelling aspect and modality semantics.

**Part II** of the volume is “Problems of Semantics and their Representaionin Slavic Digital Lexicography”. The paper by L. Iomdin discusses new developments of representing semantics in the digital combinatorial dictionaries of the ETAP-3 System. The paper by L. Dimitrova, V. Koseska, R. Dutsova and R. Panova describes the design and development of the first Bulgarian-Polish online dictionary. V. Shyrovkov’s paper deals with theoretic problems of lexicographic systems. The paper by D. Fišer and T. Erjavec presents a proposal for semantic concordances in Slovene. A syntactic-semantic treebank for domain ontology creation is described in the paper by K. Simov and P. Osenova. M. Šimková, R. Garabík and L. Dimitrova present the design of a multilingual terminology database prototype. The dictionary of Slovak collocations was described in the paper by P. Ďurčo, R. Garabík, D. Majchráková and M. Ďurčo. A. Przepiórkowski’s paper presents a comparison of two morphosyntactic tagsets of Polish. The paper by R. Roszko and D. Roszko describes morphosyntactic specifications for Polish and Lithuanian. The paper by R. Roszko analyses differences between the morphosyntactic descriptions for Polish verbs and MULTEXT-East morphosyntactic specifications from the point of view of a prospective unification.

**Part III** of the volume is dedicated to some semantic problems of contrastive studies of Slavic languages and related topics. The paper by S. Timoshenko and O. Shemanaeva discusses lexical functions in Bulgarian and Russian. J. Satoła-Staškowiak’s paper deals with the translation of Polish uninflected present participle in Bulgarian literature — on the basis of Adam Mickiewicz’s ‘Pan Tadeusz’. M. Dushkin’s paper presents exponents of adnumeral approximation in Polish and Russian. The paper by O. Bugakov analyses definitions of prepositions, conjunctions and particles in explanatory dictionaries. E. Gwiazdecka’s paper examines the semantic description of Polish verbal prefix. The lexicographic representation of relational nouns is discussed in the paper by P. Osenova.

**Part IV** of the volume is a collection of abstracts of short presentations at the Workshop.

The workshop in Warsaw was highly useful and efficient. The editors hope that the presented contributions will be of interest to both lexicographers and computer scientists.

V. Koseska, L. Dimitrova, R. Roszko



Part 1  
**Common Theoretical Problems of Semantics**





# Semantic Interlanguage and Contrastive Studies\*

Violetta Koseska-Toszewa, Roman Roszko

Institute of Slavic Studies, Polish Academy of Sciences, Poland  
amaz@inetia.pl ; roman.roszko@ispan.waw.pl

**Abstract.** The analysis of the language confrontation issues presented in the paper shows the imperfection of the results of research where two or more languages are compared based on a formal inventory, i.e. the so-called morpho-syntactical features and values. The use of interlanguage as a language of consistent and simple notions helps overcome the formal barrier, and ensures that the individual confronted languages are always referred to the same meaning plane, known traditionally as *tertium comparationis*. The results of research on natural languages obtained based on a confrontation with a semantic interlanguage are comparable and have an equal status.

## 1 Interlanguage

Interlanguage (*tertium comparationis*) is a language used for comparing two or more natural languages.

## 2 Contrastive linguistics

Contrastive (confrontative) linguistics is a field of synchronous linguistics with both theoretical and practical applications[14]. When contrastive studies deal with analysing differences and similarities for practical purposes (didactic or translation-related ones), we refer to them as a field of applied linguistics, connected first of all with teaching foreign languages. We can also single out the stream of research on the machine translation theory with a high degree of materialization.

**2.1.** In turn, we speak of theoretical contrastive studies in the case when they concern universal linguistic issues and use methods of language studies, aimed at isolating from languages the elements which are either common or different for them.

With respect to research methods used, as well as the use of synchronous approach, theoretical contrastive studies are close to typological studies, but differ from the latter in the aim of description. Typological studies lead to classification of languages, while contrastive studies — to systemic analysis of the compared languages. Moreover, typological classification of languages is based on revealing such differences between languages which can be used as a basis for exhaustive classification of many natural languages, while contrastive analysis is limited to just a few (most often, two) languages. Hence the sphere of contrastive studies is more modest than that of typological studies, see ([15], [14], [3, p. 366–456]).

**2.2.** Applied contrastive studies have used, and still use, the following basic notions, which are worth reminding here:

*Primary language* (or home language, native language) is the first language system which we master in childhood. Learning of any language later in life is based on mastering equivalents for the home language.

The *starting language*, according to A. Szulc [13], is the language being the point of reference in the practical process of teaching a foreign language. As a rule, it is the primary language, though exceptions from that rule may occur. This can be especially the case when learning the second foreign language, where the starting language may be the first foreign language (this is, for

---

\* The study and preparation of these results have received funding from the FP7 under grant agreement Mondilex.

example, the language situation of a Pole, who after mastering Serbian and Croatian is learning Albanian).

The *target language* is the opposite of the starting language. This term denotes the foreign language, either first or some in a row, being the target of teaching. The *object language* is the language which directly expresses the contents. Its opposite is the meta language, i.e. a language used for describing another language. In Szulc's opinion [13], one should point out the fact that in bilingual dictionaries the target language represents the object language, while the starting language (attention!) is the metalanguage for the former.

The *interlanguage* is not only related to theoretical contrastive studies. The term itself was coined by Selinker in 1969, in his talk at the 2<sup>nd</sup> International Congress of Applied Linguistics in Cambridge. During it, Selinker said that interlanguage is the "type of competences in the target language which is the product of the competences in the home language and the target language system" (See [12]). However, this definition fails to tell us what type of competences in the target language are referred to. We also have a problem of another nature, which we will discuss in more detail below.

As we can see, both the term "interlanguage" and the notion itself are relatively new. Together with the progressing development of the contrastive grammar theory, they may have been used not necessarily in line with Selinker's intention. In the hitherto developed contrastive descriptions, a selected language, usually a foreign language for the recipient, was compared to another language, usually the recipient's home language. With such an approach, the description consisted first of all in translating the surface constructions, characteristic for the foreign language and unknown to the recipient, and their comparison with the constructions of the recipient's home language. This type of studies focused on a very detailed (and providing a lot of valuable information, by the way) description of selected means for expressing given contents, whereby other means for expressing the same contents, often equally important for the characteristics of the whole language system, were totally disregarded.

However, in the case when the starting point for the contrastive description is the system of formal categories of the language, the researchers often reach a quite erroneous conclusion, e.g. the conclusion that Polish lacks the imperceptive modality or the definiteness/indefiniteness category viewed as a semantic one. As a further consequence, comparison of languages which are relatively remote typologically (which is the case with Bulgarian and Polish) in line with the traditions of language confrontation does not give and has never given any guarantee for revealing problems not noticed or described yet, and has never offered chances for viewing the issues discussed in single-language descriptions from another, new perspective.

**2.3.** Traditional contrastive studies were treated in a reserved way, because researchers had in mind solely studies comparing formal facts in one language with formal facts in another language. Typology had necessarily an advantage over this type of contrastive studies, for it had at its disposal a richer and more diverse language material. However, typological studies might be of greater importance for the natural language description theory only when observations of the compared languages will proceed in the direction from the meaning to the form, and will refer to an equal extent to each of the studied languages. This requirement is doubtlessly difficult to meet in the cases when a researcher has at his/her disposal material from, say, 40 languages, but knows only a few of them well, and interprets facts from others based on the subject literature only. Indeed, it is well-known that the language phenomena considered in it are not treated in a uniform way. Another problem is the phenomenon of terminological polisemy.

**2.4.** Doubtlessly, also in this case a good methodological solution would be an interlanguage used for objective and equal comparison of meanings and forms of the examined languages. However, development of such a language is an extremely difficult task, even if we are comparing two languages only. An equally difficult task is a description leading from analysing the content plane towards formal analysis of the considered languages, but such a description guarantees the maximum advantage for the recipient.

### 3 Interlanguage in the *Bulgarian-Polish Contrastive Grammar*

The *Bulgarian-Polish Contrastive Grammar* team undertook to execute these two methodological tasks, well aware of the difficulties involved in them (see [1]). The team rejected a description going from language A as the starting point to language B as the goal, and started working on development of an interlanguage. In order to separate descriptive descriptions of a single language from contrastive descriptions, it was necessary to clearly distinguish between the notion of a metalanguage describing a single language from that of an interlanguage, which constitutes a tool for comparing at least two language systems.

**3.1.** Thus the notion of a metalanguage differs from that of an interlanguage first of all in the fact that a metalanguage is used for describing one given language, while an interlanguage is a tool for comparing at least two language systems. In our approach, it is also a semantic language, which consists of semantic categories and notions necessary for their description.

**3.2.** It is worth noting that an interlanguage keeps developing and acquiring new notions as the research progresses. In our opinion, the most important requirement during its development is that the interlanguage be built based on theories which do not lead to a contradiction. For example, when building the basic semantic units used to describe the linguistic category of definiteness/indefiniteness in the interlanguage, we can use either the reference theory or the definite description theory. However, a simultaneous use of both the theories is not recommended, since it leads to internal inconsistencies in the concept system of the interlanguage. This can be seen in the works which do not separate the notions chosen here as an example, such as reference and definite description. Already from Volume 2 GKBP [5] we can see that a description choosing as a starting point Bulgarian formal language means is quite different from a description oriented at Polish formal language means. One the reasons for this is the more expanded morphological plane of the means expressing the notions of definiteness and indefiniteness in Bulgarian compared to Polish (see also [6]). This is, among others, why replacing the interlanguage by one of the contrasted languages together with its metalanguage would be a major methodological error — and this is how this issue is treated in most of the contrastive works we know.

**3.3.** The interlanguage for comparing Polish and Bulgarian within the semantic category of definiteness/indefiniteness is based on the assumption on the quantificational character of that category. Its basic notion of uniqueness (uniqueness of an element or uniqueness of a set) can be written down using the linguistic iota-operator, that of existentiality — using an existential quantificational expression, and of universality — using a universal quantificational expression (see [5], [4]).

**3.4.** The interlanguage needed for comparing Polish and Bulgarian within the semantic category of time and modality is based first of all on Petri net theory. For example, the notions of **state, event and process** are distinguished as units of the interlanguage in exactly the same way as they are defined in the net theory. Also the metalanguage for that interlanguage, i.e. the language expressing the above notions, see e.g. places, transitions and arrows in the net, is described in accordance with the definition of the interlanguage. The notions corresponding to modality types, such as conditionality, hypotheticality or imperceptiveness, are also distinguished, and interpreted in accordance with the net-based (granular/discrete/non-continuous) description of the semantic category of modality adopted in our network volume. For example, conditionality is captured in terms of branchings and forks in the net and the cause-effect law combining states and events; hypotheticality is connected with free choice nets; and imperceptiveness is directly connected to the global state.

### 4 State, event, process

The scope of the notions of state, event and process, known from the literature describing temporal, aspectual and modal phenomena in natural languages is not uniform. “Unfortunately,” wrote

Lyons in 1977, “there is no appropriate term which would encompass states on the one hand, and events, processes and actions on the other hand” [7, p. 101]. The above quotation illustrates the multifarious applications of the English terms state and event in linguistics (see also [8]).

**4.1.** Also in logic, from which linguistics has borrowed these notions, they were not distinguished very strictly before the development of Petri net theory in the 1960s [9]. As a result, the terms “state” and “event” were also used in different senses, and even interchangeably. For example, B. Russell was of the opinion that the world could consist of events, with each of them occupying a specific dimension of the timespace (...) and that one could make a bundle of events which could be considered as appearances of a “single thing”. He assumed that “each event occupies a determined and limited part of space and time, and that it occurs simultaneously with an infinite number of other events, which partially, but not wholly, occupy the same section of the timespace. A mathematician who wants to operate with points-moments can construct them with help of mathematical logic out of overlapping sets of events” [11, p. 11, 15, 116]. When introducing temporal notions to his theory of grammatical tenses, Reichenbach spoke of event, reference and speech points. In the sentence *Piotr szedł*. [Piotr was walking], an “event point” is the moment when Piotr was walking, and the “reference point” is the time between the event point and the speech point. For Reichenbach, the notion of “event spread” existed too. “English” — he wrote — “uses the active participle of the present tense, known as Present Participle, to mark that a given event covers a certain period of time”, see e.g. the sentence with the continuing event *I had been seeing John*. As we can see, Reichenbach does not use the notion of “state”, extending instead the meaning of the “event” notion. In the literature, the terms *state* and *event* were used interchangeably, without being separated like e.g. in the works of Petri [9].

## 5 Grammar dictionaries

In grammar dictionaries, terms of the type: *state*, *event*, *process* appeared in various meanings, depending on the individual theory. This cannot be allowed in the interlanguage describing contrasted languages — the interlanguage can only contain unambiguous terms and strictly defined notions. Indeed — this requirements follows from the purpose for which it is used. Thanks to the interlanguage, we can compare different languages in a reliable way, i.e. so that specific contents are represented by different formal means in the compared languages, and the languages are treated equally. See the semantic interlanguage in *Bulgarian-Polish Contrastive Grammar* [4].

## 6 Selected notions of the interlanguage with elements of its metalanguage

**Collective quantitative quantification / collective quantification** (concerns quantification<sub>2</sub>). This kind of quantification is related to multiple quantification, which reveals itself as detailing of the quantitative characteristics of multi-element sets. Collective quantification consists in assigning the property following from predication  $P$  to the whole given set  $((A)P(A))$ . Collective quantification does not imply distributive quantification.

**Current state.** A state including the speech state; besides the speech state, it includes also other states, coexistent with that state.

**Differentiation between states and events.** This is an essential feature of Petri nets. Each event ends or begins a state; two different states following one another must be divided by some event, which ends one of them and begins the next one. Similarly, between two events following each other there is always a state (which can be described e.g. as follows: <<the first event has already occurred, and the second one has not occurred yet>>).

**Definitive quantitative quantification / quantitative definiteness** (concerns quantification<sub>2</sub>). Precise determination of the number [or quantity] of objects, events and states. Singularity always satisfies the conditions of quantitative definiteness. In case of multiplicity, an indefiniteness characteristics is also possible.

**Distributive quantitative quantification / distributive quantification** (concerns quantification<sub>2</sub>). It is related to multiple quantification, which reveals itself as detailing of the quantitative characteristics of multi-element sets. Distributive quantification consists in assigning the property following from predication  $P$  to each element of the given set  $((\forall a \in A)P(a))$ . In other words, for each  $a$  belonging to the set  $A$ , it holds that  $a$  possesses the properties following from predication  $P$ .

**Eternal state.** This is a state that neither has nor ever will be broken by any event. However, such states are of no importance for dynamic aspects of the described situations.

**Event.** A point on the time axis being a border between states. An event cannot be extended in time — it does not last. An event ends the existence of some state and/or begins the existence of another one. For example, the four seasons of the year are states; the equinoxes and solstices are events; the spring equinox (event) separates winter (state) from spring (state). Each event begins or ends some state. Between two events following one another, there is always some state.

**Existentiality,** see scope-based existential quantification

**Future state.** A state being one of the possible consequences of the speech state.

**Global state.** Global state consists of the states of all objects in a given situation, in opposition to a local state which only involves one or a few objects in that situation. For example, in the situation: “door, windows” a local state is  $\langle\langle$ the door is closed $\rangle\rangle$ , and a global state will be  $\langle\langle$ the door is closed, the windows are open $\rangle\rangle$ . We can say that a global state is a special case of a local state since it includes, as mentioned above, all objects of the situation, while a local state includes one or some of them.

**Incomplete quantification** Insufficient formal differentiation of the quantification types (unique, existential, universal), resulting from ambiguity of quantification exponents. The kind of quantification is determined in strict cooperation with the context or situation (which includes also a minimum knowledge of the extra-language world common for the sender and the recipient).

**Indefinitive quantitative quantification / quantitative indefiniteness** (concerns quantification<sub>2</sub>). Approximate determination of the number or quantity of objects, events and states. Quantitative indefiniteness is never connected with singularity — it concerns multiplicity only.

**Local state.** A property of a certain selected object (or objects) of the described reality. Local states constitute interpretations of single network places: a marked place corresponds to occurrence of the property, and an unmarked place — to its non-occurrence. Events occur locally, i.e. they change local states. If we want to describe the real world in a natural language, we must refer in it to local states; certain modal forms of a natural language reflect effects of the locality of states. According to Petri net theory, with a given local state we can associate a set of global states — namely, all the states which are compliant with the local state in the given fragment of the universe.

In the semantic structure of an imperceptive sentence, a local state is connected with occurrence of an obligatory feature — participation of more than one observer in the net, which constitutes a representation of a primary information act (primary information situation) that is being related by the current sender.

**Multiplicity.** See multiple quantitative quantification.

**Numerical quantification,** see quantification<sub>2</sub>

**Past state.** A state whose consequences include the speech state.

**Precedence — succession relation.** Succession relation, see the arrow in networks. Precedence, succession and simultaneousness do not depend on the speech state only. This is because the states and events mentioned can refer to states and events which had occurred before the speech state, to those simultaneous with the speech state, and finally to states and events which occurred after the speech state, rather than directly to the state of speech.

**Present state.** See current state.

**Process.** A process is a configuration of states and events in the network representation. The basic argument of the proponents of the linear approach to describing time is their perceived need

for introducing a notion of a process in which a certain quantity (or quantities) change(s) in a continuous way. Let us consider, for example, the process described in words as “ $X$  is growing”. In the linear (continuous) representation, the process is represented by a section expressing the period of growing. In the network (discrete) representation, this notion is represented by the state of  $\langle\langle\text{growing}\rangle\rangle$ , either together with the events beginning that state or without them. The description of phenomena on a linear scale is a special example of the network description, i.e. the network description is a generalization of the well-known and commonly used method of describing temporal phenomena. All that can be expressed in a linear model can be expressed in a network model of the same complexity degree. However, the network model allows for a concise description of situations which in a linear description would require introducing a great number of variants.

**Quantification**, see quantification<sub>1</sub> and quantification<sub>2</sub>. Operation of binding variables with a quantifier (numerical, universal, existential one, or iota-operator).

**Quantification<sub>1</sub> / (=logical quantification) / scope-based quantification**, see quantifying. Quantification<sub>1</sub> is related to logical quantifying. The use of generally accepted definitions of logical quantifiers (the universal and existential ones) and of the iota-operator allows us to single out three basic notions whose meanings are determined by the language exponents of logical quantification [10, p. 211–255] and definite description [11, p. 253–293], see [5]. Quantification of natural language expressions can concern names (first order logic), but also predicates (second order logic). A quantifier transforms a logical predicate into a logical sentence — hence predication is not identified with quantification. Quantification is not a “syntactical operation” which transforms a sentential function into a sentence, but a mechanism which reveals a semantic relationship between the quantified object (a single one or a set) and the truth-based method of forming sentences. For example, the set ( $\iota X$ )  $\langle\langle\text{is a two-legged featherless being}\rangle\rangle$  is the only set satisfying the predicate  $P(X)$ , where  $P(X)$  means ‘ $X$  is the set of all humans’.

Each quantification used in a sentence decreases the number of (free) variables of the quantified predicate. It seems natural to classify a unique substitution of an object for a certain variable of a predicate as quantification, since such an operation also decreases the number of free variables. This is the way the iota-operator was treated in the works of Barwise and Cooper [2, p. 159–219], where the quantificative model concerned, though, the nominal phrase only, and in Volume 2 of the *Bulgarian-Polish Contrastive Grammar* [5], where the quantificative model concerned also the verbal phrase and the whole sentence. The notions of the iota-operator and the unique quantifier are synonymous there.

**Quantification<sub>2</sub> / quantitative quantification (=numerical quantification)**. Quantification<sub>2</sub> is not related to logical quantifying. This is an operation of binding individual variables occupying an argument position with a numerical quantifier. It assigns objects, events and states quantitative characteristics, which in case of discrete things is determined by counting them (e.g. *two books, he was late twice, he has read this book twice*), and in case of non-discrete things — by their measurement based on agreed units (e.g. *hectare of land, litre of milk, bottle of beer*). In opposition to scope-based quantification, it does not transform a logical predicate into a logical sentence, and does not bind variables. The value of quantitative quantification is read outside the context and the situation — but requires a minimum knowledge of the world.

**Quantifying (logical quantifying)**. A unary function transforming free variables into bound variables. This phenomenon concerns both the level of the verbum group (e.g. *to come — he came exactly then*) or the nomen group (*tree — exactly this tree*) and the level of both these groups together (*Jan/to read — (this) Jan is reading exactly now*). Quantifying transforms a sentential form (= logical predicate) into a sentential function (=logical sentence). The expression bound in the process of quantifying at the same time determines the scope of that quantification process (*man — this man | some man | some man over twenty | some people | every man | all people, etc.*). For more details on that subject, see quantification 1.

**Quantitative multiple quantification / multiple quantification / multiplicity** (concerns quantification<sub>2</sub>). This kind of quantification is opposite to quantitative quantification of singularity, and is a basic subcategory of the semantic category of quantity which consists in binding an

individual variable occupying an argument position with a numerical quantifier having a value different from one, exactly one, once, exactly once.

**Quantitative quantification**, see quantification<sub>2</sub>

**Reachable state**. This is a state which can be reached from the initial configuration.

**Scope-based existential quantification / existential quantification / existentiality** (concerns quantification<sub>1</sub>). A unary function transforming free variables into bound variables of the form  $(\exists x)P$ , which precedes predicate  $P$  in the semantically-logical structure of the sentence. Existentiality concerns objects, states, events and processes.

**Scope-based quantification**, see quantification<sub>1</sub>, quantifying. This kind of quantification is connected with quantifying and encompasses all quantification issues except numerical quantification.

**Scope-based quantification of time / quantification of states and events** (concerns quantification<sub>1</sub>). It is connected with quantifying taking place within the verbum group. Elements subject to quantifying include single states, single events and processes, which are defined as sequences of states and events. It can take unique, existential or universal values.

**Scope-based unique quantification / unique quantification / uniqueness** (concerns quantification<sub>1</sub>). A unary function transforming free variables into bound variables of the form (1)  $(\iota x)P(x)$  or (2)  $(\iota X)P(X)$ , which precedes predicate  $P$  in the semantically-logical structure of the sentence. Uniqueness concerns objects, events, states and processes.

**Scope-based universal quantification / universal quantification / universality** (concerns quantification<sub>1</sub>). A unary function transforming free variables into bound variables of the form  $(\forall x \in X)P$ , which precedes predicate  $P$  in the semantically-logical structure of the sentence. Universality concerns objects, states, events and processes

**Singularity**, see singular quantitative quantification

**Singular quantitative quantification / singular quantification / singularity** (concerns quantification<sub>2</sub>). This is a basic subcategory of the semantic category of quantity opposite to multiplicity, which consists in binding an individual variable occupying an argument position with a numerical quantifier having the value one, exactly one, once, exactly once.

**Speech state**. This state coincides with the information sender state. The speech state determines all states continuing in the present, and indirectly determines the states continuing in the past and the events occurring in the past, as well as the possibility of existence of states and events in the future.

**State**. This is a property of a certain object of reality. In the discrete approach to process description, a paradigm of a state is its duration. Each state lasts for a certain time. Two different states following one another are separated by some event, which begins the new state and ends the old one.

**Strength of a quantification meaning** (concerns quantification<sub>1</sub>). This is separation of meaning differences within the same quantification using the labels of strong and weak quantification meaning, singled out based on secondary semantic properties of expressions. The strength of a quantification meaning is determined by the position of the quantifier in the semantic structure of the sentence. If the quantifier has the broadest scope in the semantic structure of the sentence, i.e. if it covers with its scope all other quantifiers present in the semantic structure of the sentence, then we speak of a strong quantification meaning. If the quantifier's position is within the scope of other quantifiers contained in the semantic structure of the sentence, then such a quantification meaning is termed weak.

**Uniqueness**, see scope-based singular quantification

**Universality**, see scope-based universal quantification

## 7 Conclusions

Traditional dictionaries of grammatical notions find their reflection only in the language for which they have been developed. Hence we cannot say that the individual dictionaries of grammatical



notions for arbitrary two languages are comparable with each other. The more languages we compare, the greater disproportion we note between the grammatical notions contained in the dictionaries of those notions for the individual languages. Such divergences can be illustrated on the example of the morphological definiteness/indefiniteness category. So let us compare:

(a) In English, the morphological definiteness/indefiniteness category is based on the opposition between the use of the definite prepositional article *the* (used for both singular and plural nouns) to the indefinite article *a/an* (positional variants — used for singular nouns only).

(b) The same category in Bulgarian is based on the opposition between the use of the postpositional article *-sm/-a* to the so-called morphological zero.

(c) In turn, in Baltic languages, this category is formed by the opposition between qualitative adjectives and participles with complex (pronoun-based) flexion, and qualitative adjectives and participles with simple flexion. Its scope does not cover simple expressions founded on a noun alone.

The examples given above reveal substantial differences of not only formal but also meaning-related character (built based on a formal plan different for each language). Hence one can neither compare the formal exponents, nor — even less so — the meaning planes created based on non-uniform exponents with different usages in each of the presented languages.

The recently fashionable dictionaries of morphosyntactical features and values developed for multiple languages do not go beyond the formal plane either. Hence the individual categories, features and their values, though given the same name for several languages, need not describe the same language phenomenon. An example of this is the newest transformation of the meaning of the Croatian aorist form, whose use in the texts refers to the use of e.g. the Lithuanian perfectum form. Another example we can also give here is the problem of participles, e.g. in Polish and Lithuanian. The differences in the formal plane are substantial: 4 forms of Polish participles (with high restrictions on their creations) compared to 18 Lithuanians forms (created arbitrarily for each verb). What is more, Lithuanian formally differentiates the form of the same participle depending on the function it performs in the sentence, e.g. *dirbąs* – *dirbantis* (both participium praesentis activi, sg. masc.): *Jis dirbąs*. ‘He is allegedly working now’ and *dirbantis žmogus* ‘a working man’. No analogous formal operation is known in Polish. In the content plan, each use of a Polish participle can be rendered using a Lithuanian participle — but the reverse operation is not possible. The fact that there is no reflexivity in the use of Polish and Lithuanian participles discredits the use of traditional methods of so-called formal language confrontation.

A commonly known, though seemingly only too frequently unnoticed fact, is that languages differ from each other first of all in the formal plane — while the meaning plane is the universe which connects both genetically related and unrelated languages.

## Bibliography

- [1] Projekt (1984). Projekt gramatyki konfrontatywnej bułgarsko-polskiej i serbskochorwacko-polskiej. Wstęp. In *Studia polsko-południowosłowiańskie* (ed. Polański, K.), Wrocław.
- [2] Barwise, I., Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219.
- [3] Heinz, A. (1978). *Dzieje językoznawstwa w zarysie*. Warszawa.
- [4] Koseska-Toszeva, V., Korytkowska, M., Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Dialog, Warszawa.
- [5] Koseska, V., Gargov, G. (1990). *Семантичната категория определеност/неопределеност, Българско-полска съпоставителна граматика, т. 2*. София.
- [6] Koseska, V., Mazurkiewicz, A. (1988). Net representation of sentences in natural languages. In *Lecture Notes in Computer Science 340. Advances in Petri Nets 1988*, pages 249–266. Springer-Verlag.
- [7] Lyons, J. (1989 (1977)). *Semantyka 2*. Warszawa.
- [8] Miller, G., Johnson-Laird, Ph. (1976). *Language and Perception*. Cambridge – London – Melbourne.

- [9] Petri, C.A. (1962). Fundamentals of the theory of asynchronous information flow. In *Proc. of IFIP'62 Congress*, Amsterdam. North Holland Publ. Comp.
- [10] Rasiowa, H. (1975). *Wstęp do matematyki współczesnej*. Warszawa.
- [11] Russell, B. (1967). Denotowanie, deskrypcje. In *Logika i język*, pages 259–293, Warszawa.
- [12] Selinker, L. (1972). Interlanguage. *Iral*, 10(3):209–213.
- [13] Szulc, A. (1984). *Podręczny słownik językoznawstwa stosowanego*. Warszawa.
- [14] Weinsberg, A. (1983). *Językoznawstwo ogólne*. Warszawa.
- [15] Zabrocki, L. (1970). Uwagi o problemach spornych gramatyki kontrastywnej (konfrontatywnej). *Lingua Posnaniensis*, XIX:2–23.

# The Issue of Interlanguage in Contrastive Studies

Małgorzata Korytkowska

University of Łódź

**Abstract.** The approach presented in the paper motivates syntactical phenomena by semantic features of predicative units, as well as by the specific structure of the argument places opened by those units. The objects contained within those places are carriers of states in the sense of net theory, but describing that sphere solely by its relations to states and events does not exhaust the whole problem area. The paper presents an outline of an apparatus of analysis starting with the semantic plane, in which the elements of the interlanguage are classes of semantic predicates and the set of *predicate-argument positions* defined by positions at possibly simple predicates. The paper also shows problems connected with the condensation phenomena. As an effect of such phenomena, some elements of the semantic structure are only realized superficially, and functions of argument phrases might be sometimes neutralized.

1. It is well-known that defining the functions of sentential elements (and hence their semantic statuses) has an extremely long tradition in linguistic. By way of example, we can quote here terms like *agens*, *paciens*, *logical subject*, *subject*, etc. For a very long time, linguists defined this type of terms using definitions having an intuitive character and not formulated explicitly, which prevented verification. Hence e.g. *agens* was defined as the process initiator / process source / actor, etc.; *paciens* — as a goal/ object / substance at which the *agens*' action was targeted, etc. Though the functions of phrases were only defined in such a detailed way for the formal class of verbs allowing the so-called passive transformation, one can note that such an approach failed to take into consideration the multi-functionality of language forms, and hence a certain conventionality of the language with respect to the placement of phrases in the sentential structure. Thus e.g. in case of the Polish verb *wypraszać* [wheedle, plead]:

*wypraszać coś u kogoś* [wheedle sb. out of sth.] — e.g. *Piotr wyprosił u przyjaciół milion złotych.* — the phrase *milion złotych*—as an accusative one, has the *paciens*' function in this approach, while in case of *dopraszać, prosić* [plead,beg;ask] — the phrase does not have such function any longer, see

*dopraszać się czegoś u kogoś* [beg sb. for sth.] — e.g. *Piotr doprasza się miliona złotych u przyjaciół.*

*prosić kogoś o coś* [ask sb. for sth.] — e.g. *Piotr prosi przyjaciół o milion złotych.*

The reasons are formal here: a) the impossibility of carrying out a passive transformation in case of *dopraszać* (see the grammatically incorrect *\*Milion złotych jest dopraszany przez Piotra.*), or the choice of the phrase *przyjaciół* for the syntactic position of the direct object (whence that phrase rather than *milion złotych* has the *paciens*' function). In fact, such an approach often led to identifying the position of the syntactic subject (a phrase congruent with *verbum finitum*) with the *agens*' position, and the position of the direct object's phrase—with the *paciens*'<sup>1</sup> position.

A serious attempt to develop a theory which would enable defining the functions of phrases in semantic terms, and allow for equal treatment of all phrase positions in the sentence (including the subject position), was the theory developed by L. Tesnière [9], who recognized the *verbum* as the core whose features determine the number and the value of the individual actants. This breakthrough was of a revolutionary character; it was also the first attempt to create terms which were to enable comparison of different language systems. However, the definitions developed by Tesnière for the individual actants were still of non-explicit character, and the values assigned to the individual actants in the analysed examples clearly point out that also this time the position of the so-called 1<sup>st</sup> actant was identified with the subject position, of the 2<sup>nd</sup> actant — with the

<sup>1</sup> A broader review of theoretical positions on those issues can be found in [5].

direct object's position, of the 3<sup>rd</sup> actant — with the indirect object's position (as a rule, this is a prepositional phrase), etc. Thus e.g. in the sentence *Le livre me plaît.* the phrase *le livre* is recognized as the 1<sup>st</sup> actant, which contradicts the definition of the latter position, assigning the 1<sup>st</sup> actant active participation in a given process. Such a theoretical apparatus turns out to be ineffective as well in the analysis of sentences from different languages which correspond to each other. Hence e.g. for the sentences eng. *I miss you.* and French *Vous me manquez.*, the 1<sup>st</sup> actant's position is assigned to the phrase *I* in English and to the phrase *Vous* in French [9, p. 288], though any person who can understand those sentences refers the phrase *I* to the phrase *me*, and the theory should be able to interpret that fact.

It is difficult to outline her even most briefly the consecutive theoretical propositions referring to that problem area [8], [2], [1]. Doubtlessly, the next breakthrough attempt to refer to it was the theory of so-called semantic cases proposed by Ch. Fillmore [3]. However, also in that case the definitions were not based on the structure of the language (including the structure of the dictionary of a given language), but rather referred to the extra- language world and did not have the character of an explication. Thus e.g. *Dative* was defined as 'the case of an animate being encompassed by an action', and e.g. *Objective* — as the case of 'things that are encompassed by an action or state', etc. [3, p. 24]. The theory also failed to take into consideration the fact that the semantic structure of the verbum often results from more or less advanced condensation processes — and hence semantic analysis should reach deeper, giving the chance to take into consideration the fact that opening of certain argument positions might stem from reduction of certain fragments of the semantic structure.

2. Doubtlessly, it is difficult to interpret functions of the syntactical positions of phrases in the structure of the sentence. The said functions, analyzed based on the surface structure, are not semantically distinct, and the syntactical positions of phrases are multi-functional. The basis for the analysis seems to be the unquestionable and already widely popular thesis on the influence of semantic features of the predicate, i.e. the number and type argument places opened by the predicate, on the shape taken by the structure of the sentence. However, when taking that thesis into consideration, we cannot disregard the semantic structure of the verbum that realizes the predicate's position. Hence we can easily show the complexity of the semantic structure of e.g. the *causatives* class, and refer it to the surface realization. It turns out that a number of components of that structure, which is reflected by a paraphrase with analytic features, permanently fails to be realized in the surface structure, and that only some of its fragments can be selected and placed in the surface structure. Basically, a causative predicate opens two positions for propositional arguments: *p'* causes that *p''* happens, see e.g.: *To, że Kasia zrobiła awanturę(p')*, *spowodowało to, że powstało wielkie zamieszanie (p'')*.

The above implies that the argument positions relevant for the surface structure can be determined on the level of paraphrases with analytic features which contain lexical units functioning in the language and possibly simple semantically — implementations of semantically simple predicates. On that level, we can establish in an explicit way the definitions for the individual so-called predicate-argument positions<sup>2</sup>. These will be kinds of labels which are associated with a specific place at a certain type of possibly simple semantically predicate<sup>3</sup>. Placement of the individual types of predicate-argument positions in the surface structure of the sentence (i.e. features of the surface valence of the verbum) is, as already pointed out above, to a large extent a conventional (or: idiomatic) matter in a given language. Hence we can examine the tendency for placing certain types of arguments in certain syntactical places (e.g. in the subject), but it is easy to show that this is not the one and only position. Thus, for example, verba with a causative structure often admit to the subject position phrases located in that position in various ways, which is disclosed by paraphrasing, see e.g.: *Paweł wgniół maskę samochodu. — Paweł zrobił coś, co spowodowało, że maska samochodu jest wgnieciona. / Skata wgniotła maskę samochodu. — Stało się coś ze*

<sup>2</sup> A detailed description of such a model and of its application to contrastive analysis is given in [5].

<sup>3</sup> It should be pointed out that the terms used for specifying predicate-argument positions recall Fillmore's ones, but their definitions and usage are different.

*skatą* (see *Spadająca z góry skata*), *co spowodowało, że maska samochodu jest wgnieciona*. Hence it is also difficult to agree with the conception of anthropocentric features of the sentential structure, since in causative structures subject phrases often constitute a partial realization of the argument  $p'$ , and at nominalization of that argument, the realization of both the predicate and its arguments is possible, see e.g. *Stoczenie się skatły ze zbrocza wgniotło maskę samochodu*. Of course, a number of verbal units which admit causative interpretation represent such an advanced degree of semantic structure condensation that the full structure of  $p'$  can no longer be realized superficially, see e.g. *Piotr poinformował Annę o decyzji syna. — Piotr sprawił, że Anna wie o decyzji syna*.

**3.** The composition of the set of predicate-argument positions should, it seems, be suggested by the analysis of the degree of their relevance for the description of both the semantic plane and the formal plane. Hence in case of e.g. a description of two languages we should take into consideration the need to distinguish such categories for both the systems (at least in one of them, the given value should be characterized by a certain degree of grammarization)<sup>4</sup>. In so short a study, we can only give a general idea of the applied analysis apparatus, illustrating it on a few selected examples taken from Polish and Bulgarian.

**3.1.** The argument position Experiencer (Exp) is determined by the class of predicators (that is, verbs and predicative verbo-nominal units) that refer to processes taking places in the individual's mind or the emotional sphere of an individual, as well as sensations received by the senses. That position is determined by the position of  $x$  at such predicators as Pol. *x czuje / widzi / słyszy / wie (że  $\hat{S}'$ )*; Bulg. *x чувствува / вижда / чува / знае (че  $\hat{S}'$ )*. Hence those predicators represent second order predicates, opening one of the positions for a propositional argument. The Exp position can be fulfilled by entities characterized as [+Anim] (or, more narrowly, [+Hum]), and the predicator refers to states / events beyond the control of that individual, which is witnessed by the impossibility of providing a context for a causative sentence characterized voluntarily, or for an intentional sentence (occurrence of a voluntative context implies the possibility of controlling the process / state), see e.g. the grammatical incorrectness of: Pol. *\*Tęsknię, ponieważ tak chcę. — \*Tęsknię, abyś był zadowolony.* / Bulg. *\*Тъгувам, понеже искам така — \*Тъгувам за да си доволен*. The argument position defined in this way is assigned to predicators with which sentences allow for an explication of the above type, regardless of the location of the phrase. Hence, for example, that position will occur in the argument structure of the verb Pol. *podobać się* and Bulg. *харесвам*, though the location of Exp in the structure of sentences from both the languages may differ, and the issue is determined by a paraphrase — assignment of the position of  $x$  at a predicate referring to feelings to a certain phrase (i.e., the decision which of the phrases performs the function of  $x$  at the predicator *czuć*), see e.g.:

Pol. *Anna spodobała się Marysi.* — here: *Marysi* (Exp) as NP. in the dative.

Bulg. *Мария харесала Иванка.* — *Мария* (Exp) as the subject NP.

Such an analysis of the set of predicators which open a position for Exp on the semantic level allows for determining its distribution in each of the examined languages, the ways of its placement, and possible preferences or conditionings for its location, etc. We should distinguish here two basic sets, formed by non-causative and causative verbs. An extensive analysis of material from both languages allows us to determine its position at non-causative predicators as systemically labile. Most often, it is placed in the subject phrase or in a position beyond the subject (direct or indirect object)<sup>5</sup>. In case of causative predicators, a phrase with the value Exp is placed beyond the subject position, see e.g. the class of information verbs, which open a position for a propositional argument and for the Exp position, and fit within the basic paraphrase for that class, of the type  $y$  (Ag) *causes that  $x$  (Exp) knows that  $p'$* .

Pol. *Anna wyjaśniła mi (Exp), dlaczego nie przysłała.*

Bulg. *Ана ми (Exp) обясни, защо не е дошла.*

<sup>4</sup> Such a principle was adopted in the work on the *Confrontative Bulgarian-Polish Grammar* [GKBP].

<sup>5</sup> An extensive analysis of the material is given in [5].

**3.2.** The argument position Agentive (Ag) is determined by possibly simple semantically predicators referring to facts of actions, activities. In Polish this is the position of *x* in expressions of the type *x działa / robi*; Bulg. *x прави / действа*. This position contrasts with the Exp position discussed above — Agentive is an argument of predicators which refer to processes under the control of an individual, with respect to which the individual can take a volutative stand. Hence the context of sentences with predicators opening an Agentive position may contain causative clauses with volutative features or intentional sentences, see e.g.

Pol. *Anna wyjeżdża za granicę, ponieważ chce poprawić sobie humor (/ aby poprawić sobie humor).*

Bulg. *Ана заминава за чужбина, защото иска да си подобри настроението (/ за да си подобри настроението).*

Here we should note the possibility of including the position Ag within a propositional argument opened by the predicator of the main sentence. In such cases, the said predicate-argument position may be obligatory — the semantic features of the predicate may require the occurrence of predicators opening a position for Agentive within the propositional argument (and hence in a clause), see e.g.:

Pol. *Piotr zmusił Adama, aby (Adam — Ag) wyszedł z pokoju.*

Bulg. *Петър принуди Адам (Адам) — (Ag) да излезе от стаята.*

Paraphrases of this type of sentences indicate the inclusion of a *necessity* component within the propositional argument, see: *Piotr zrobił tak, że Adam musiał wyjść z pokoju*. The phrase *Adam / Адам* at the object position can be treated here as a specific lifting of the argument position *x* (Ag) *z p'* (realized by a clause). However, there are cases where the position Exp occurs in the semantic structure, and the predicator belongs to a class that opens a position for a propositional argument, which also contains obligatorily an Agentive position. These are information predicators which mark a propositional argument with regard to *necessity / advisability of p'*; paraphrases of those sentences fit within the type *y does so that x knows that he/she must/shuold p'*, see e.g.:

Pol. *Piotr każe Adamowi wyjść z pokoju. (= aby Adam wyszedł z pokoju), / Piotr prosi Adama, aby (Adam) wyszedł z pokoju.*

Bulg. *Петър заповядва на Адам (Адам) да излезе от стаята. / Петър моли Адам (Адам) да излезе от стаята.*

Most often, only one of the positions is realized in such sentences (they are denotatively identical), and one can say that in the surface structure we can observe a kind of flow-down of the Ag and Exp functions. However, the apparatus used here allows us interpret that fact.

**4.** Hence the application of the presented apparatus to analysis led from the semantic plane to the syntactic one allows for a consistent interpretation of phenomena within a single language, as well as within contrastive analysis. In such an analysis, isolated semantic units constitute the basic notions, which are units possibly simple semantically (taking into consideration the specifics of the studied lexical systems). They are later used for the analysis of cases more complicated semantically — that is, for complex, semantically expanded predicators. Of essential importance here is the assumption that the basis for interpretation of the set of sentences with a given predicator (i.e., a unit with one function = one meaning) is the characteristics in the form of opened positions assigned to that predicator. Hence e.g. the verb Pol. *zachwycać się* — *възхищавам се* has the structure  $P_{(x,p')}$ , and each sentential structure with that verb can be reduced to this semantic schema, see:

Pol. *Anna zachwyca się tym, co widzi przez okno. / Anna zachwyca się górami. / Anna zachwyca się widokiem gór. / Anna zachwyca się odwagą kolegi. / Anna zachwyca się kolegą.*

Bulg. *Ана се възхищава от това, което вижда през прозореца. / Ана се възхищава от планините. / Ана се възхищава от гледката на планината. / Ана се възхищава от храбростта на колегата си. / Ана се възхищава от колегата си., etc.*

Such an approach enables interpretation of the individual types of sentential structures in reference to the initial pattern, which amounts to a description of the different processes taking place in the basic structure of the sentences which realize that pattern. These are processes of multifar-

ious transformations of the basic structure, including ones which do not result in content losses, and ones which lead to omitting (for various purposes) parts of the content of that structure<sup>6</sup>.

4.1. Among processes which do not result in content losses we can rank the process of the so-called splitting of the propositional argument position (also with its nominalization taking place), see e.g.:

Pol. *To, że Adam zachował się nieodpowiednio, zdziwiło mnie.* / *Nieodpowiednie zachowanie się Adama zdziwiło mnie.* — *Adam zdziwił mnie tym, jak się zachował* (/ *swoim nieodpowiednim zachowaniem*).

Bulg. *Това, че Адам се държеше лошо, ме изненада.* / *Лошото държане на Адам ме изненада.* — *Адам ме изненада с това, че се държеше лошо* (/ *със своето лошо държане*).

4.2. However, there can also be transformation processes which affect only part of the set of sentences with a given predicator. This part of the set can be determined in various ways—for example, by a condition of undefinedness (existentiality or universality) imposed on the phrase (which needs not be the only limitation, by the way), as in case of the structures of some types of subject-free sentences:

Pol. *Ten dom zbudowano bardzo starannie.* (*Ktoś* / *Jacyś ludzie zbudowali ten dom bardzo starannie*).

Bulg. *В това легло е спано.* (*Някой* / *Нещо* (животно) *е спал(о) в това легло.*) / *Ядено е от паничката ми.* (*Някой* / *Нещо* *е яло от паничката ми.*)

In case of a propositional argument, a substantial content reduction might take place<sup>7</sup>, which can be related with its undefinedness, often deliberate incomplete statement (though the content may be also e.g. clear from the context), see:

Pol. *To, że Adam zachował się nieodpowiednio, zdziwiło mnie.* — *Adam zdziwił mnie.*

Bulg. *Това, че Адам се държеше лошо, ме изненада.* / *Лошото държане на Адам ме изненада.* — *Адам ме изненада.*

5. While in case of analysis progressing from the level of the initial structure to derivative structures it is relatively easy to interpret the ongoing condensation processes, the reverse direction of analysis — from the surface structure to the initial structure—may involve certain difficulties. As shown by the reasoning up to now, in the adopted model unique identification of the functions of the individual element in the sentence structure is only possible on that level. Hence the analysis starting from the surface structure should first of all take into consideration semantic features of the predicator, and reproduce the creation “history” of the surface structure. It must also distinguish between phrases which constitute realization of the semantic features of the predicator (=predicate) and the phrases which constitute the so-called added elements (which are fragments of other predications, like e.g. *Sąsiad produkuje zabawki w garażu.* = *Sąsiad produkuje zabawki i odbywa się to w garażu.*).

5.1. This process may lead to a certain neutralization of phrase functions, which can cause difficulties in assigning them an argument position in the initial structure. So, for example, the position of the phrase *Karol* is unclear in sentences of the type:

Pol. *Karol mnie zaskoczył.* — see: a) *To, co zrobił Karol, mnie zaskoczyło.*; b) *To, jaki jest Karol, mnie zaskoczyło.* Only a) admits interpretation of the phrase *Karol* as an Agentive argument position. This is connected with the fact that the verb does not impose the obligatory condition of opening that position on the set of predicators admitted to the subject sentence.

In case of the verb Pol. *szkodzić*<sup>8</sup>, which is a causative verb with the argument structure  $P(p, q)$ , content reduction processes can concern both arguments (the semantic structure contains a negative assessment of (the possibility of) the realization of  $q$ ), see:

<sup>6</sup> For analysis of nominalization process for sentences which realize a propositional argument, see [7]

<sup>7</sup> More broadly on processes which lead to content reduction see [6].

<sup>8</sup> Not in the sense ‘szkodzić zdrowiu’ [to be bad for health], like e.g. *Brak snu szkodzi.* / *To lekarstwo szkodzi.*

*To, że Adam wysyła skargi, szkodzi temu, jak Piotr awansuje (/ wygrywa konkurs / aby był zdrowy).*

= *To, że Adam postępuje w ten sposób sprawia / powoduje, że to, że Piotr awansuje (/ wygrywa konkurs / jest zdrowy) jest utrudnione / jest realizowane w niekorzystny sposób (etc.).*

This verb ten admits a far-reaching reduction in the elements of the basic sentential structure and realization of both *p* and *q* through their object arguments. See:

***Wysyłanie skarg przez Adama szkodzi awansowi / wygranej / zdrowiu Piotra.***

***Adam szkodzi awansowi / wygranej / zdrowiu Piotra wysyłaniem skarg.***

***Adam szkodzi awansowi / wygranej / zdrowiu Piotra.***

***Adam szkodzi Piotrowi.***

In case of that verbum, the *szkodzić* verb imposes the requirement for occurrence of the Agentive argument position on the right-hand side propositional argument (*p*) only, so the function of the phrase *Adam* in the structure ***Adam szkodzi Piotrowi.*** is clear.

6. Despite a number of detailed problems arising in that context, the essential thing in a contrastive study of the syntactical plane is the development of an interlanguage which is based on the semantic plane and clearly separated from the formal plane (also by the use of different terminology for both levels of the analysis). Its units should be defined explicitly and should ensure consistent interpretation of sentential structures, independently of their formal features. In the model outlined here, an important role belongs to analytic paraphrase, which constitutes a tool for analysis enabling identification of argument positions. Of course, in a contrastive study of that level it is important to point out not only differences, but also similarities between the studied languages. Such an approach facilitates a holistic view of the examined scope of problems, and helps pose basic questions regarding the form of the analysis apparatus. In case of studying realization of the semantic features of predicates, the similarities / differences stem, of course, on the one hand from the similar / different vocabulary structures, and hence from similar / different lexicalization processes (if they are to be treated as a process leading to emergence of vocabulary units), on the other hand — in admissibility / non-admissibility of analogous condensation processes reducing the basic (initial) structures.

## Bibliography

- [1] Daneš, F. (1968). Sémantická struktura větného wzorce. In *Otázky slovanské syntaxe II*, pages 45–49, Brno.
- [2] Daneš, F., Hlavsa, Z. (1973). Hierarhizace sémantické struktury věty. In *Československé přednášky pro VIII. Mezinárodní sjezd slavistů v Zahřebu*, pages 67–77, Praha. *Lingvistika*.
- [3] Fillmore, Ch. J. (1968). The case for case. In *Universals in Linguistic Theory*, Bach, E., Harms, R. T. (eds.), pages 1–88.
- [4] Korytkowska, M. (2004). Wokół problemów opisu kategorii kauzatywności i sposobów jej realizacji (na przykładzie języka bułgarskiego i polskiego). *Slavia Meridionalis*, 4.
- [5] Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska, t. 5, Typy pozycji predykatowo-argumentowych*. SOW, Warszawa.
- [6] Korytkowska, M., Kikiewicz, A. (to appear). O szczególnym typie zjawiska kompresji i o interpretacji struktur zdaniowych będących jej efektem (na przykładzie języka białoruskiego i języka polskiego). *Slavia Orientalis*.
- [7] Korytkowska, M., Maldziejewa, W. (2002). *Od zdania złożonego do zdania pojedynczego. Dopuszczalność nominalizacji argumentu propozycjonalnego w języku polskim i bułgarskim*. Toruń.
- [8] Pauliny, E. (1943). *Struktura slovenskeho slovesa*. Bratislava.
- [9] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris.



# Constructing Catalogue of Temporal Situations\*

Violetta Koseska<sup>1</sup>, Antoni Mazurkiewicz<sup>2</sup>

<sup>1</sup> Institute of Slavic Studies, Polish Academy of Sciences

<sup>2</sup> Institute of Computer Science, Polish Academy of Sciences

**Abstract.** The present paper is a continuation of the authors' papers submitted to the Moscow and the Bratislava meeting of MONDILEX project group. In this paper we formulate some rules of the catalogue construction as well as a number of temporal situations expressed in the proposed formalism. These situations are also explained in an informal way and some examples of corresponding phrases expressed in different languages are given.

## 1 Introduction

The aim of this paper is to continue work on creating a language independent list of basic temporal situations. This list is supposed to be a common framework for comparing linguistic forms used for describing the listed situations. As it has been already said [2] the comparison should be made on the basis of situations rather than grammatical (linguistic) forms.

Recall here the basic elements of the description formalism of temporal situations suited for the linguistic purposes. These elements, following [5] are: (1) states, representing physical or mental phenomena that are extended in time, (2) events, which represent some changes of states and taking no time, and (3) the flow (ordering) relation, binding states with events and indicating their mutual succession. It turns out that the three elements mentioned above are sufficient for expressing many of every-day situations in a language-independent way, as indicated in [2]. In fact, the language of states, events, and flow is a sort of an artificial language, serving as an intermediate (go-between) language, also known as "*tertium comparationis*". Call it the *Petri net language*, or the net language for short. The reader can consult some source texts on Petri Nets, eg. [7], or some earlier papers of the present authors, as e.g. [1], [4], to find the detailed description of the net formalism.

## 2 Formalism of nets

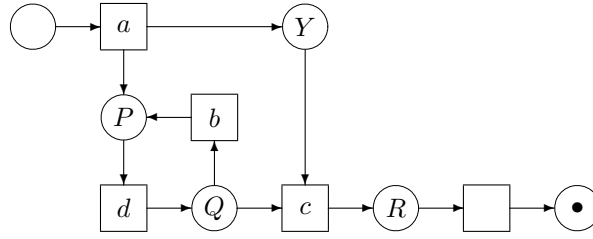
For the purpose of the present paper it suffices to recall the following facts.

Petri nets as used here are built of three basic elements: events (symbolized by boxes), states (symbolized by circles) and flow relation (symbolized by arrows). Any finite structure consisting of these elements, with some of them joint by flow relation in such a way that it connects a state with an event, or an event with a state (neither two states nor two events are directly connected by the flow). An alternating sequence of states and events connected directly by the flow relation is called a *path* through the net and indicates the sequence in which these elements appear in time. Any two elements of the same path are ordered in time: either one of them precedes the other, or the other way round. Nets with places from which there is only one arrow leaving it or pointing to it we call deterministic. Example of a deterministic net is given below:

---

\* Work supported by EU FP7 project GA211938 MONDILEX "Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources".

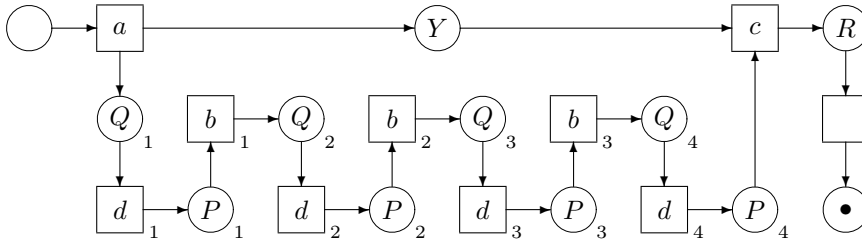




as a tool for representation of an arbitrary sequence of alternating states  $P, Q$  separated by events  $b, d$ , starting with state  $P$  and ending with state  $Q$ :

$$(a, P, d, Q, b, \dots, P, d, Q, b, P, d, Q, b, \dots, d, Q, c, R, \dots)$$

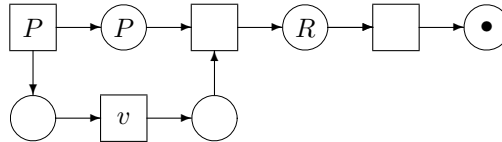
Events  $a$  and  $c$  serve for indicating the beginning and ending the repetition. In other words, the following net is an instance of the scheme given above, with indices 1,2,3, and 4 indicate successive instances of events and states being repeated:



**Past situations.** In this section we give a formal description of situations placed in the past with respect to the state of speech (the state of utterance). The number of such situations results from the number of mutual positions of (a) state (or point) of reference, see [6], (b) object state (or states), to which the utterance refers, (c) their ordering (or lack of ordering, i.e. the coexistence), (d) possible repetitions of object states and events. The list of situation schemes is certainly not exhaustive; but we hope that it offers a pattern for further extensions.

### Scheme 1

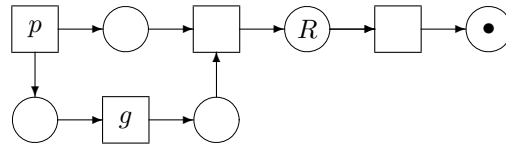
(Past perfective resultative)



- (eng) He had been drunk ( $P$ ) when visitors entered ( $g$ )  
 (bul) Kogato gostite vljazoha( $v$ ), toj beshe veche pijan ( $P$ )  
 (pol) Gdy goście weszli ( $v$ ), on już był pijany ( $P$ )  
 (rus) Kogda gosti voshli ( $v$ ), on uzhe byl p'jan ( $P$ )

In Bulgarian, event ( $g$ ) is expressed by aorist form of perfect verbs, while in Polish and Russian by praeteritum form of them. In all four languages state  $P$  is expressed by the past participle form.

**Scheme 2**  
(Past before past)



- (eng) He had got drunk ( $p$ ) before visitors entered ( $g$ )  
 (bul) Toj se be napil ( $p$ ), predi {da dojdāt, idvaneto na} gostite ( $g$ )  
 (pol) On się upił jeszcze ( $p$ ) przed przyjściem gości ( $g$ )  
 (rus) On napilsja ( $p$ ) do prihoda gostej ( $g$ )

In Polish and Russian this temporal situation is expressed similarly by praeterital form of perfective verbs. It can also be expressed by a noun with a preposition, in Polish *przed*, in Russian *do*. However, in Bulgarian the preceding event  $p$  is expressed by the pluperfect form of perfective verbs. Even  $q$  can be expressed in similar way in Polish and Russian by nouns with prepositions and also by the infinitive form of perfective verbs.

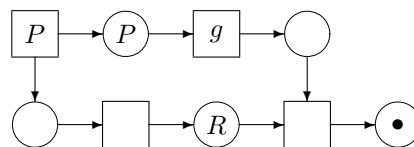
Other examples of expressing the above temporal situation are:

- (bul) Toj beshe dosh'al ( $p$ ), predi tja da dojde ( $q$ )  
 (eng) He had come here ( $p$ ) before she did ( $q$ )  
 (pol) On tu {był} przyszedł ( $p$ ) zanim przyszła ona ( $q$ )  
 (rus) On prishel ( $p$ ) pered tem, kak prishla ona ( $q$ )

This temporal situation is precisely expressed in English and Bulgarian by pluperfect form of perfective verbs. In Polish and Russian, additionally, it is supplemented by prepositions with praeterital form of perfective verbs, as *zanim*, *przed*, (Pol) *pered tem, kak* (Rus) or similar. In Polish pluperfect form is archaic.

**Scheme 3**  
(Past before past perfective)

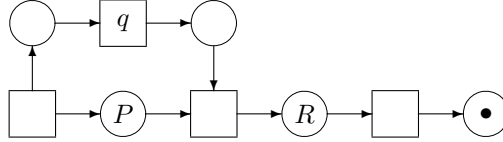
The difference between Scheme 2 and Scheme 3 consists in changing the position of event  $g$  and reference state  $R$ . Position of the reference state in Scheme 3 indicates its coexistence with state  $p$ ; it means that after  $g$  has happened, state  $P$  can still be holding, according to the story told by a speaker at place  $\bullet$  (the state of utterance).



- (eng) He had been drunk ( $P$ ) before visitors entered ( $g$ )  
 (bul) Toj stana ( $P$ ) pjan predi da dojdāt gostite ( $g$ )  
 (pol) On stał się ( $P$ ) pijany już przed przyjściem gości ( $g$ )  
 (rus) On zachmelel eshche ( $P$ ) do prihoda gostej ( $g$ )

**Scheme 4**  
(Past before past)

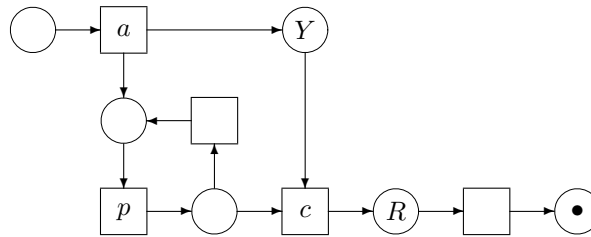
Scheme 4 is similar to Scheme 2, but with a state ( $P$ ) instead of event ( $p$ ) in its temporal structure:



- (bul) *Toj veche beshe dosh'al ( $P$ ) kogato tja dojde ( $q$ )*  
 (eng) *When she came ( $q$ ) he {was, has been} already here ( $P$ )*  
 (pol) *On już tu był ( $P$ ), gdy ona przyszła ( $q$ )*  
 (rus) *On uzhe byl ( $P$ ), kogda ona prishla ( $q$ )*

**Repetitive situations.** Schemes discussed in this section differ from the ones given above. Namely, Scheme 4 given below contains two states (1 and 2) that can be started or terminated with two different events each, henceforth (according to the net properties) excluding each other. State 1 can be initiated by event  $a$  beginning the cycle, or by  $b$  repeating the cycle. State 2 can be terminated with event  $c$ , continuing the cycle, or with  $b$ , closing the repetitions of the cycle. The number of repetitions is undefined, and is left to the "decision" made at state 2: to leave repetitions by event  $b$ , or to return to them by event  $c$ . Therefore, Scheme 4 describes a class of situations rather than a single one; which one of them is actually expressed is irrelevant from the speaker's point of view.

**Scheme 5**



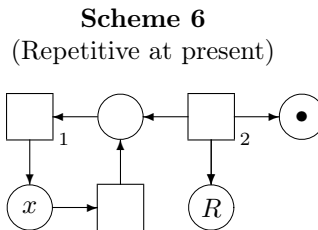
- (bul) *Minalata godina ( $Y$ ) {ponjakoga, често, rjadko} toj ja poseshtavashe ( $p$ )*  
 (eng) *Last year ( $Y$ ) he was visiting her ( $p$ ) {occasionally, frequently, sporadically}*  
 (pol) *W zeszłym roku ( $Y$ ) on {czasem, często, rzadko} ją odwiedzał ( $p$ )*  
 (rus) *W proshlom godu ( $Y$ ) on {inogda, chasto, redko} jeje poseshchal ( $p$ )*

In Bulgarian, the above temporal situation is expressed by the imperfect form of imperfective verbs. In Polish and Russian it is expressed by the praeterital form of imperfective verbs.

According to the general rules of net understanding, there is a number of histories consistent with such a scheme, in each history all states terminate or start with only one event. In this way the above scheme describes several possibilities of the history course; therefore, it describes a situation where state  $p$  is repeatedly renewed, starting and ceasing to exist in an alternating way. All these actions take place during state  $Y$  is holding; state  $Y$  and all instances of states and

events in the cycle are coexistent; in particular, state  $Y$  is coexistent with all occurrences of event  $p$ .

Next schemes show various versions of expressing repeating situation in dependence on their position with respect to the reference state (which is always situated in the past). Symbol  $x$  denotes the action (state) the situation is referring to. Scheme 7 describes repeating situation coexistent with the reference state.



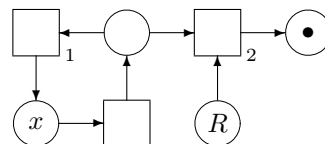
It is worth to notice that in the above diagram states  $\bullet$  (of utterance),  $R$  (of reference), and repetitive structure containing event  $x$  are independent, so to speak, coexistent. Therefore, all states and events of this repetitive structure are coexistent with both reference and utterance states. Examples of expressing this situation are as follows.

- (bul) *Tja sega ot vreme na vreme sjada pri prozoreca*  
 (eng) *Nowadays, she is sitting by the window from time to time*  
 (pol) *Ona teraz od czasu do czasu siaduje przy oknie*  
 (rus) *Sejchas {vremja ot vremeni, chasto} ona saditsja u okna*

In all languages mentioned above but Polish the repetitive character of the action in question is expressed by supplementing it with adverb *from time to time* (Eng), *ot vreme na vreme* (Bul), *vremja ot vremeni* (Rus). In Polish verb *siaduje* indicates explicitly the repeatability of the action, derived from imperfective verb *siadać*.

**Scheme 7**  
(Repetitive in the past)

Scheme 7 expresses the same situation, but shifted to the past with respect to the reference state. In this diagram state  $R$  of reference as well as the whole diagram of repeating states and events are in the past of the state of utterance  $\bullet$ . However, similarly to the scheme 6, state of reference  $R$  is coexistent with the whole repeating structure - between them there is no temporal dependency whatsoever.

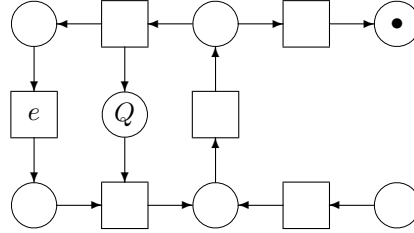


- (bul) *Tja ot vreme na vreme sjadashe pri prozoreca*  
 (eng) *She used to sit by the window from time to time*  
 (pol) *Od czasu do czasu ona siadywała przy oknie*  
 (rus) *Vremja ot vremeni ona sadilas' u okna*

In Polish and Russian the above temporal situation is expressed by praeterital form of imperfective verbs, while in Bulgarian it is expressed by the imperfect form of imperfective verbs.

### Scheme 8

The next scheme (Scheme 8) is quoted to indicate a possibility of use the reference as an event rather than as a state; that is, it is then a "point of reference" rather than "a reference state".



(bul) *Toj vinagi v tak'av edin moment (e) ja nenavizhdashe (Q)*  
 (eng) *He was hating her (Q) at any such a moment (e)*  
 (pol) *W kaźdej takiej chwili on j nienawidził (e)*  
 (rus) *On vsegda v takoj moment (e) jeje nenavidil (Q)*

In Polish and Russian this temporal situation is expressed praeterital form of imperfective verbs. In Bulgarian it is expressed by imperfect form of imperfective verbs.

### 3 Conclusions and further plans

In the present paper we dealt with net representation of temporal situations referring to the past. Some basic situations have been listed and explained using net formalism. With a single exception of repeating occurrences of situation elements there were no need to represent modalities such as different possibilities of the history courses, or uncertainty of some states Or events occurrences in the described situations. This issue is left for the forthcoming paper.

**Closing remarks.** we have presented a continuation of the work on creating a catalogue of temporal situations based on Petri nets theory. The list of situations presented in this paper is certainly not exhaustive; one can find a number of situations worth of listing and analyzing for the linguistic purposes. However, this list give a guidance for the next discussion on situation presentations and is open for a further augmenting and completion.

In the forthcoming paper we intend to create a similar list for situations related to the future and to various aspects of modality.

List of used situation functions.

Scheme	Situation	Temporal meaning
Scheme 1		Past perfective resultative, with state as the object
Scheme 2		Past before past, with event as object
Scheme 3		Past perfective, with object coexistent with reference
Scheme 4		Past before past, with state as object
Scheme 5		Past repetitive imperfective, with event as object
Scheme 6		Repetitive at present
Scheme 7		Repetitive in the past
Scheme 8		Repetitive situation in the past

Bibliography

- [1] Koseska-Toszeva, *Semantyczna kategoria czasu*, GKBP, SOW, Warszawa, 2007
- [2] Koseska V., Mazurkiewicz A.: *Net representation of sentences in natural languages*, Advances in Petri Nets, 1988, LNCS 340, Springer Verlag, pp 249-259
- [3] Koseska V., Mazurkiewicz A.: *Net Net Based Description of Modality in Natural Language (on the Example of Conditional Modality)*, Proc. of the MONDILEX Second Open Workshop, Kiev,(2008)
- [4] Mazurkiewicz, A.: *A Formal Description of Temporality (Petri Net approach)*, Lexicographic tools and techniques, Proc. of the MONDILEX First Open Workshop, Moscow, ISBN 978-5-990813 (2008) pp 98-108
- [5] Petri, C.A.: *Fundamentals of the Theory of Asynchronous Information Flow*, Proc. of IFIP'62 Congress, 1962, North Holland Publ. Comp., pp 386-390
- [6] Reichenbach, H.: *Elements of Symbolic Logic*, New York, McMillan Publ. (1944)
- [7] Reisig, W.: *Petri Nets — An Introduction*, New York 1985, Springer Verlag



# Automated Extraction of Lexical Meanings from Corpus: A Case Study of Potentialities and Limitations

Maciej Piasecki<sup>1</sup>

Institut of Informatics, Politechnika Wroclaw University of Technology  
maciej.piasecki@pwr.wroc.pl, www.plwordnet.pwr.wroc.pl

**Abstract.** Large corpora are often consulted by linguists as a knowledge source with respect to lexicon, morphology or syntax. However, there are also several methods of automated extraction of semantic properties of language units from corpora. In the paper we focus on emerging potentialities of these methods, as well as on their identified limitations. Evidence that can be collected from corpora is confronted with the existing models of formalised description of lexical meanings. Two basic paradigms of lexical semantics extraction are briefly described. Their properties are analysed on the basis of several experiments performed on Polish corpora. Several potential applications of the methods, including a system supporting expansion of a Polish wordnet, are discussed. Finally, perspectives on the potential further development are discussed.

## 1 Introduction

A technique or tool must have its purpose and justification for using it. Moreover, before we apply it, we should ask how it works and whether it does what was promised. We are going to approach these questions with respect to the automated extraction of formalised descriptions of lexical meanings from corpora.

Automatic methods are based on algorithms. An algorithm requires a formal model of the processed data: input and output. The crucial issue is a formal, or at least formalised, description of lexical meanings which will be discussed in Section 2. Having a formal model of lexical meanings defined, next we need to analyse what kind of evidence pertaining to the elements of the model can be automatically extracted from a corpus. A short review of the possible sources of evidence is presented in Section 3. Two basic paradigms of extraction of lexical meanings, namely a *pattern-based* and *distribution-based*<sup>1</sup> are introduced in Sections 4 and 5, respectively. Next, various applications of the introduced techniques and their hybrid combinations in the area of Computational Linguistics, but also Linguistics in general, are discussed in Section 6. In spite of significant improvements observed in the field during the last decade still in many aspects we should talk rather about potentialities instead of fully developed solutions. Perspectives on the possible development of the methods in the nearest future is briefly presented in Section 7 which concludes the paper.

## 2 Formalised Description of Lexical Meanings

From the historic perspectives, the oldest form of lexical meaning description is a lexical entry based on listing different *lexical units* (meanings) of a *lemma* in separate positions. Each lexical unit is given a short description in the natural language and often accompanied by a set of examples. This kind of lexical meaning definition is not very useful as an output for the extraction algorithms, since it is not formalised and a synthesis of a proper definition expressed in the natural language is a serious demand for Natural Language Processing. Proper lexicon definitions, e.g. in a sense postulated by [1], occur very rarely in a general corpus, unless it includes a dictionary. Automatic construction of precise definition would require detailed model of the structure and meaning of natural language definitions, i.e. utterances of special kind and purpose.

<sup>1</sup> The latter one is also called clustering-based [20].

On the opposite end of formalisation scale lays a technique based on the application of *meaning postulates* to the semantic description of logical symbols representing lexical units. Precise semantic representation of the natural language is based on a formal, logical language, in which lexical meanings are represented by logical predicates. As the predicate meaning is defined by its formal interpretation the only way to transfer meanings from natural language units to predicates is by constraining interpretation of the predicates, i.e. by shaping the formal structures of the interpretation model. A meaning postulate is an axiom constraining the possible interpretation of a given logical predicate used for representing meaning of a particular lexical unit, e.g. [8]

Defining a meaning postulate is usually a demanding task, which requires rich knowledge. The tasks complicates with the increasing size of the set of meaning postulates. One can hardly imagine the acquisition of meaning postulates from text corpora, as it would require extraction of detailed, formally expressed, knowledge about the world.

In Componential Semantics lexical meaning is defined as a set of features which distinguish it from other lexical meanings. The meaning can be also presented as a expression in a formalised language. The expression consists of basic meaning atoms linked by some operators, e.g. a short review of works in this area given by Dowty [7] but also works of Wierzbicka [34] or Pustejovsky [29] or Mel'čuk [18] can be mentioned here as examples of component-based analyses.

In traditional thesauruses, e.g. Roget's Thesaurus, lexical semantic relations like synonymy, hypernymy, meronymy etc. and grouping lexical units into semantic clusters are utilised in constructing large coverage descriptions of lexical meanings. Description by a network of relations defined on the set of lexical units is partial in comparison to the componential analysis, however appeared to be useful not only for human readers but also for Language Technology applications, e.g. hundreds of applications of WordNet – an electronic thesaurus of English [9, 26]. Lexical semantics relations occur also as elements of formalised lexical meaning descriptions, e.g. [17, 29].

In sum, there are two basic types of elements constituting formal descriptions: basic components and relation instances. However, the components must later be combined into complex expressions to form definitions. This complicates the process of automated extraction. That is why we will focus mainly on automatic extraction of lexical semantic relations.

### 3 Machine Tractable Evidence

Analysing a corpus we can take two possible perspectives. First we can concentrate on detailed analysis and drawing conclusions from particular occurrences of the given lemma in focus, trying to extract detailed information concerning its semantics from each context of occurrence. Secondly, we can take a global perspective and analyse all occurrence with lower accuracy each but taking into account statistical evidence. Both approaches are used and will be discussed in the following subsections.

#### 3.1 Embedded definitions

Since an utterance of a text performs often informative function, we can find language expressions in a corpus which describe meanings of particular lemmas in various ways, e.g. an example of quasi-definition found in corpus by Hearst [12].

The *bow lute*, such as the *Bambara ndang*, is plucked and has an individual curved neck for each string.

The above sentence relates *Bambara ndang* to the *bow lute* as its hyponym, but also characterises it by some details. In many cases, we can identify specific lexico-syntactic constructions that indicate a pair of lemma as an instance of a particular lexical semantic relation, e.g. hypernymy in the above example.

When we take into consideration more complex lexico-syntactic and possibly also semantic structures, we can also try to identify complex, descriptive definitions included in a text, e.g. the descriptive, ending part of the example above, or an example given in [10, pp. 3]:

A linguist is a scientist who investigates human language [...]

Such indicative language constructions are more likely to be found in texts of specific genres like encyclopaedias (almost every entry includes a helpful passage of text) or text books, but even experiments performed on a general corpus bring relatively good results, see Section 4.

In order to utilise the information expressed in this way, we need to apply some kind of structural analysis which is sensitive to the lexico-syntactic structures in focus. This issue will be discussed in details in Section 4.

### 3.2 Distributional Hypothesis

As the structure of language expressions is determined by the properties of constituents, including semantic properties, an analysis of a large number of language expression occurrences should allow us to identify some regularities of semantic nature. This general assumption has appeared in several linguistic theories. A well known version was proposed by Harris [11] in the form of Distributional Hypothesis.

Harris [11] in his *Distributional Hypothesis* expressed a strong belief that there is a direct relation between the observed use of language expressions and their meaning (cited after to [32]):

The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities.

The granulation of entities is not specified in the hypothesis. They can be any language items. Henceforth, we will concentrate our attention on one or multiple word *lemmas*, representing one or several *lexical units* each. A lemma is expressed in text by one of several possible *word forms*.

As in the whole corpus it is hard to identify strict restrictions on lemma combinations that are always preserved, it is more practical to perceive restrictions as preferences of different strength. It is used to describe a set of restriction imposed on a lemma in a kind of reversed way by identifying a set of particular *contexts* in which the given lemma can occur. However, we should describe each context by the strength of preference too.

A context is a textual situation involving a number of word form occurrences. A more detailed definition requires answering two questions:

1. what kind of features do we use to describe a context,
2. and how large is a context.

As for the first question, an occurrence context of the lemma  $x$  can be identify with:

- the association of  $x$  with a particular textual object, e.g. a particular document  $d$ ,
- the co-occurrence of  $x$  with a particular word form or lemma  $y$ ,
- the participation of  $x$  in an instance of a particular lexico-syntactic relation, e.g. an occurrence of  $x$  as an argument of a particular predicate together with some or all argument set, see detailed examples in Section 5.

The size of the context can be defined as the size of the whole textual object, a *text window* (i.e. a text snippet of certain number of words) or some syntactic construction like a sentence.

Identification of contexts in which a given lemma  $x$  can occur and the evaluation of the association strength result in a model of the distribution of  $x$ . According to the Distributional Hypothesis, comparison of the distributions of two lemmas allows us to draw conclusions concerning the similarity of their meanings. This claim is verified on the basis of empirical data in Section 5.

## 4 Pattern-based Relation Extraction

Traditional semantic lexicons consist of entries written in the natural language, but all definitions have similar structure and associate described lemmas with other lemmas occurring in entries in ways indicating certain lexical semantic relations, e.g. hypernymy (hypernyms are usually given at

the beginning of a lexicon entry description). Following this observation, a number of approaches to the extraction of lexical semantic relations from Machine Readable Dictionaries were proposed. A set of lexico-syntactic patterns was defined, where each pattern was a regular expression<sup>2</sup> defined over word forms, their morpho-syntactic properties and/or simple syntactic structures.

Text in a large corpus had seemed to be of much less predictable character, however, as the seminal work of Hearst [12] showed, similar patterns can be applied to the corpus text and produce valuable results. One of the productive Hearst's pattern is presented below:

$NP_0 \dots \text{such as } \{NP_1, NP_2 \dots (\text{and } | \text{ or } )\} NP_n$

The pattern or more precisely the language constructions identified by the pattern in corpus implies that each noun phrase  $NP_i$  is a hyponym of the noun phrase  $NP_0$ , i.e. the hypernymy relation holds between lemmas represented in the text by the given noun phrases. Hearst [12, 13] constructed manually only five patterns frequently matched in a corpus and appealingly accurate. The accuracy was measured as a number of lemma pairs linked by the hypernymy relation in WordNet [9] to all those extracted. For the pattern shown above, for example, 61 of 106 extracted lemma pairs from Grolier Encyclopedia were confirmed in WordNet [12].

The implicit assumption here is that one can construct patterns accurate enough to draw correct conclusions from single occurrences of lemma pairs. In general, however, it seems barely possible due, amongst others, to the presence of metaphor. Without deeper semantic and pragmatic analysis, instances of metaphor may be hard to distinguish from literal uses. Hearst extracted *aeroplane* as a hyponym of *target* and *Washington* as an instance of *nationalist*; such derived associations are clearly specific to particular documents from which they were extracted. Another problem is the scarcity of pattern instances in corpora; merely 46 instances were acquired from 20 million words of the New York Times corpus [12].

These patterns are expressed in a grammar of limited expressive power and work on the basis of an assumption of the fixed linear order of English sentence. For a highly inflected language, like Polish, a more sophisticated mechanism is required. Thus we applied a language of morpho-syntactic constraints called JOSKIPI utilised in the TaKIPI tagger of Polish [22]. JOSKIPI is equipped with operators for testing morpho-syntactic properties of particular words, their compatibility (including agreement), defining sequences and iterating tests over word groups of non-predefined size (e.g. until the beginning of a sentence or the fulfilment of a condition). Six productive patterns were defined, cf [26], and a scheme of one of them is presented below<sup>3</sup>:

$NP1 \text{ (Adj|Adv|Noun|,)*}$   
 $(\text{base} \in \{i, \text{oraz}\}(\text{and})) (\text{base} \in \{\text{inny, pozostały}\}(\text{other, remaining}), \text{nmb}=\text{pl})$   
 $(\text{Adj|Adv})* \text{ NP2}(\text{cas}=\text{cas}(\text{NP1}))$

The pattern identifies the lemma (potentially a multiword one) of NP1 as a hyponym of the lemma of NP2. Two other similar patterns were constructed on the basis of such lexical markers like: *taki jak* (*such as*) and *w tym* (*≈including*). An example of the construction covered by the pattern including *taki jak* is presented below:

Betondour doskonale nadaje się do wykończenia podłóg w pomieszczeniach takich jak garaże,  
*Betondour perfectly is suitable for dressing floors in rooms like garages,*  
 warsztaty samochodowe, magazyny, sklepy, pomieszczenia produkcyjne, piwnice czy wykonane  
*garages, stores, shops, manufacture rooms, cellars or made*  
 z betonu schody.  
*from concrete stairs.*

As all three types of language construction are used in a very similar role, these three patterns were merged together as three variants and tested jointly. We applied them to extract hypernymic pairs from three corpora:

- *IPI PAN Corpus* (including about 254 million tokens) [28],

<sup>2</sup> It has the expressive power of a regular grammar.

<sup>3</sup> In a simplified form, the original JOSKIPI expressions have been exchanged to labels describing words and word groups identified.

- a corpus of the electronic edition of a Polish newspaper *Rzeczpospolita* from January 1993 to March 2002 (about 113 million tokens) [31];
- and a corpus of large texts in Polish (about 214 million tokens) collected from the Internet; only documents containing a small percentage of erroneous word forms (tested manually) and not duplicated in the other two corpora were included in the collected corpus.

Henceforth, we will refer to all three corpora used together as the *joint corpus*. In order to increase precision, we limited the application of the patterns only to cases in which two nominal lemmas from the predefined list occurred in the same sentence. 13 285 nominal lemmas have been collected from: the core part of plWordNet [5, 6] – 5340, a small Polish-English dictionary [27], two-word lemmas from a general dictionary of Polish [30], and the IPI PAN Corpus [28] – only those that occur over 1000 times.

The results are presented in Table 1. The accuracy was manually measured on the basis of a representative sample of the produced pair list as a ratio of positively assessed pairs to all extracted. Each pair linked by a hypernymy relation – possibly not directly, with any number of intervening other lemmas – was counted as a positive case while others as negative ones.

IPI PAN Corpus		Corpus from the Web		<i>Rzeczypospolita</i> Corpus		the joint corpus	
No. of pairs	Accuracy	No. of pairs	Accuracy	No. of pairs	Accuracy	No. of pairs	Accuracy
14611	30.06%	5983	32.52%	6682	33.16%	24437	30.69%

**Table 1.** The results of hypernymy extraction by manually constructed lexico-morphosyntactic patterns.

The extracted list cannot be used directly as a list of hypernymic pairs – the accuracy is too low (however it matches the level achieved by other approaches). The accuracy should be increased above 50% to make the tool interesting for linguists. When this merged pattern is combined with two other ones, the accuracy can be increased up to 41.05% for the price of the reduced number of pairs to 8777 [26].

The pattern-based approaches most often target the hypernymy relation. However, manually constructed patterns were also applied to the extraction of meronymy, too, e.g. [2].

Manually constructed patterns achieve relatively high precision for the cost of the limited number of pairs extracted and some time spent on their manual tuning on the basis of corpus analysis. Due to their expressive power they are difficult to be extracted automatically. However, Pantel and Pennacchiotti [20] with their *Espresso* algorithm and recently Kurc and Piasecki with its modified version and adapted to Polish [15], called *Estratto*, showed that a set of simpler, general patterns can be effectively extracted and applied in a way resulting in accuracy even better on huge corpus than the manually build patterns. *Espresso/Estratto* can be applied to any lexical semantic relation that is manifested in corpus by some lexico-syntactic markers, e.g. Espresso was successfully used to extract hypernymy but also other types of relations like e.g. *part-of*, *reaction* (in chemical sense) or *production*. Both algorithms work according to the same schema:

1. A set of example instances of the relation in focus – pairs of lemma associated by the relation, is delivered to the algorithm.
2. All close co-occurrences of lemmas from the same pair (e.g. in the same sentence) are identified in the corpus and patterns are generated in a form of generalised descriptions of token sequences occurring in between the pairs of lemmas.
3. A measure of reliability is calculated for each pattern on the basis of instances covered by it and their reliability (the reliability of example instances is set to 1).
4. A subset of the highest ranked patterns is stored and next used to extract new instances.
5. Finally, the reliability of the extracted instances is calculated in similar way on the basis of the patterns matching the instances in corpus and their reliability; only the highest ranked instances are kept for the next iteration.

---

<i>Correct hypernymy instances</i>	
koncesja ( <i>concession</i> )	decyzja ( <i>decision</i> )
kapłan ( <i>priest</i> )	człowiek ( <i>human</i> )
maj ( <i>May</i> )	okres ( <i>period</i> )
kwestia ( <i>issue</i> )	problem ( <i>problem</i> )
sowa ( <i>owl</i> )	ptak ( <i>bird</i> )
klient ( <i>customer</i> )	osoba ( <i>person</i> )
pielęgniarka ( <i>nurse</i> )	osoba ( <i>person</i> )
profesor ( <i>profesor</i> )	człowiek ( <i>human</i> )
galeria ( <i>gallery</i> )	miejsce ( <i>place</i> )
matematyka ( <i>mathematics</i> )	przedmiot ( <i>subject</i> )
matka ( <i>mother</i> )	kobieta ( <i>woman</i> )
helikopter ( <i>helicopter</i> )	maszyna ( <i>machine</i> )
droga ( <i>way</i> )	szlak ( <i>track</i> )
zespół ( <i>team</i> )	grupa ( <i>group</i> )
mecz ( <i>game</i> )	spotkanie ( <i>meeting</i> )
restrukturyzacja ( <i>restructurisation</i> )	zmiana ( <i>change</i> )
konsument ( <i>consumer</i> )	osoba ( <i>person</i> )
tenis ( <i>tennis</i> )	sport ( <i>sport</i> )
festiwal ( <i>festival</i> )	impreza ( <i>event</i> )
dziennik ( <i>daily</i> )	dokument ( <i>document</i> )
medycyna ( <i>medicine</i> )	nauka ( <i>science</i> )
anioł ( <i>angel</i> )	istota ( <i>being</i> )
spółka ( <i>partnership</i> )	firma ( <i>firm</i> )
szczur ( <i>rat</i> )	szkodnik ( <i>pest</i> )
skorpion ( <i>scorpio</i> )	znak ( <i>sign</i> )
rak ( <i>cancer</i> )	choroba ( <i>illness</i> )
nagroda ( <i>prize</i> )	wyróżnienie ( <i>distinction</i> )
<i>Non-hypernymy associations</i>	
przepis ( <i>recipe</i> )	kwestia ( <i>issue</i> )
silnik ( <i>engine</i> )	jednostka ( <i>unit</i> )
człowiek ( <i>human</i> )	drzewo ( <i>tree</i> )
program ( <i>program</i> )	działanie ( <i>activity</i> )
muzyka ( <i>music</i> )	dźwięk ( <i>sound</i> )
istota ( <i>being</i> )	nic ( <i>nothing</i> )
wojsko ( <i>army</i> )	organizacja ( <i>organisation</i> )
stowarzyszenie ( <i>association</i> )	instytucja ( <i>institution</i> )
cień ( <i>shadow</i> )	wróg ( <i>enemy</i> )
książka ( <i>book</i> )	materiał ( <i>material</i> )
słońce ( <i>sun</i> )	czynnik ( <i>factor</i> )

---

**Fig. 1.** Examples of lemma pairs extracted from the joint corpus by the application of the merged group of patterns including the *i inny (and other/remaining)* pattern.

The application of *Estratto* initiated by a list of hypernymic pairs from plWordNet to the IPI PAN Corpus [15] produced 25 361 pairs with the accuracy of 41% (measured manually on a representative sample in the same way as for the manual patterns). The result obtained automatically is significantly better than the one produced by the manual patterns. Preliminary results of the application of *Estratto* to the extraction of meronymy and adjectival antonymy are promising, in spite of the significantly worse accuracy on the level around 30%.

A weak point of the pattern-based approaches is that each occurrence of a lemma pair matching the pattern results in extracting it. There are many accidental associations. However, this occurrence sensitivity can be also an advantage for a linguist, as we can easily trace back from an extracted pair to the place of its occurrence in corpus, e.g. one can list all pairs not supported by a thesaurus. In the case of pairs extracted by automatically created patterns, the situation is slightly different, as an extracted instance is mostly supported by more than one general pattern.

## 5 Distributional Semantics

According to the interpretation of the Distributional Hypothesis presented in Section 3.2 comparison of distribution models of particular lemmas can result in some assessment of ‘how close’ meanings of both lemmas are. It is important to emphasise that in the case of most distributional methods model of the distribution is built jointly for the whole lemma and the influence of its different lexical units is mingled in it.

The basic result of distributional methods is a *Measure of Semantic Relatedness* (MSR). MSR is a function which for a lemma pair returns a number expressing the strength the semantic relation between them, i.e.  $MSR : L \times L \rightarrow R$ , where  $L$  is a set of lemma and  $R$  is a set of real numbers.

Many methods have been proposed for MSR extraction, but they all contain four general steps, more or less clearly delineated.

1. *Corpus preprocessing* – typically up to the level of shallow syntactic analysis.
2. *Co-occurrence matrix construction* – in which rows correspond to lemmas being described and columns to contexts; each cell  $\mathbf{M}[x_i, c_j]$  stores the frequency of the occurrences of the lemma  $x_i$  in the context  $c_j$ .
3. *Matrix transformation* – a possible reduction of size and/or combination of feature *weighting* and *selection*.
4. *Semantic relatedness calculation* – lemma descriptions are compared by the application of an assumed measure of similarity between row vectors.

Depending on the type of the context used, [19] distinguishes between measures of *semantic relatedness* and *semantic similarity*. Semantic relatedness is obtained on the basis of contexts defined as co-occurrence with a particular lemma in one document or a text window, i.e. the types: 1 and 2 on the page 34. Semantic similarity is extracted on the basis of lexico-syntactic relations used as contexts – the type 3, e.g.

$x$  occurs as *subject\_of(a particular verb)* or as *modified\_by(a particular adjective)*. According to our experiments performed on the IPI PAN Corpus, cf [23], semantic relatedness encompasses broader semantic associations among lemmas, based on co-occurrences of both lemmas in the description of the same situation. According to [19] lemma pairs receiving high values of semantic similarity should represent lexical-semantics relation used in thesauruses, e.g., synonymy, hypernymy, meronymy, etc. However, intermediate methods are quite conceivable. For example, one can combine lexico-syntactic constraints with co-occurrences in the description of context. So, there is a continuum of methods with these two extremes. Semantic relatedness is a more general notion as among lemma pairs expressing high semantic relatedness one can also find lemma pairs expressing high semantic similarity.

An example of a list of 20 top semantically related lemmas to the given one is presented in Figure 2. The list was produced by a MSR extracted from the joint corpus and for 13 285 nominal, both discussed in Section 4. The MSR was based on the following types of lexico-syntactic contexts:

1. modification by *a specific adjective* or *a specific adjectival participle* (41 619 features),

2. co-ordination with a a specific noun (115 604),
3. modification by a specific noun in the genitive case (115 604),
4. occurrence of a specific verb for which a given noun lemma can be its subject (19 665),

There were 167 834 active features left after weighting and selecting.

In Figure 2 we can notice that the list includes hypernyms of *gaz ziemny* (*natural gas*), e.g. *gaz* (*gas*) and *kopalina* ( $\approx$ *mineral, resource, fossil*); co-hyponyms, e.g. *węgiel kamienny* (*coal (pit-coal)*) and *ropa* (*oil*); loosely related cousins from the broader part of the hypernymic structure, e.g. *azot* (*nitrogen*) and *cynk* (*zinc*) and also lemmas similar because of the similarity of the use of the respective substances e.g. *biokomponent* (*biocomponent*) (as an addition to car fuel), while *gaz ziemny* can be used as the car fuel by itself.

<b>gaz ziemny</b> ( <i>natural gas</i> )	
gaz ( <i>gas</i> )	0.258
węgiel kamienny ( <i>coal (pit-coal)</i> )	0.207
węgiel brunatny ( <i>brown coal</i> )	0.197
ropa ( <i>oil</i> )	0.193
olej opałowy ( <i>heating oil</i> )	0.164
paliwo ( <i>fuel</i> )	0.161
wodór ( <i>hydrogen</i> )	0.160
kopalina ( $\approx$ <i>mineral, resource, fossil</i> )	0.160
węgiel ( <i>coal</i> )	0.143
olej napędowy ( <i>diesel fuel</i> )	0.140
gaz płynny ( <i>liquid gas</i> )	0.140
koks ( <i>cox</i> )	0.127
ołów ( <i>lead</i> )	0.119
azot ( <i>nitrogen</i> )	0.119
tlen ( <i>oxygen</i> )	0.116
uran ( <i>uranium</i> )	0.116
biokomponent ( <i>biocomponent</i> )	0.115
cynk ( <i>zinc</i> )	0.114
łupek palny ( <i>slate (fuel)</i> )	0.113
benzyna ( <i>gasoline</i> )	0.110

**Table 2.** A list of the 20 lemmas most similar to the given one according to the MSR from [26].

There is a notorious problem with the evaluation of MSRs. Manual inspection is misleading – one can always find good and bad examples on the lists of the most related lemmas. Simple calculation of a kind of accuracy on the basis of lemma pairs occurring on the lists and representing particular lexical semantic relations tells only part of the truth. MSR generates a function assigning some strength of semantic relatedness to every lemma pair. The assignment of higher values to pairs representing some relations is very expected but it does not exhaust the potential of MSR. Manual assessment of the MSR values is not feasible and there are no similar manually created resources to compare with. Among several potential methods discussed e.g. in [25, 35], application of MSR as the only knowledge source in solving a synonymy test, which was originally conceived for humans, have been fruitfully applied in several experiments.

Besides MSR, another kind of lexical semantic knowledge extracted on the basis of statistical evidence are semantic selectional restrictions of predicates [16]. For each subcategorisation frame of a lemma and for each argument position of the frame semantic classes of language expressions occurring on the given position are identified. As extraction of the selectional restrictions is based on statistical evidence, the association of classes to frame argument positions is described by the strength of association. The main problems are: definition of semantic classes, recognition of the occurrences of frames and classes in text and, finally, identification of statistically significant associations: a position – a class. A set of classes is defined in relation to some semantic lexicon



(e.g. semantic codes assigned to lexical units) or a thesaurus. In the latter case, the hypernymy structure (of lexical units or semantic field) is used and a subset of the nodes of the structure is selected as a basis for the classes, i.e. a class corresponds to a node. Frame occurrences are identified in text by some means of parsing (mostly shallow parsing) and class occurrences by some form of Word Sense Disambiguation (as one lemma can represent several classes). Next, the frequencies of a particular frame argument position – a particular class co-occurrences are collected and a kind of transformation based on weighting and/or selection is applied in a way similar to the transformations described for MSR.

## 6 Applications

Accuracy of any single method of extraction is around 30% in comparison to manually constructed lexical semantic resources, e.g. to relations described in plWordNet – a Polish wordnet [26]. However, different methods extract lexical semantic relations of different types, e.g. a kind of semantic relatedness (MSRs) vs a particular relation extracted by the given lexico-syntactic pattern. The methods differ also in the character of the extracted data: continuous space of relatedness values vs discrete, binary information about lemma pairs. Finally, the methods differ also in types errors made. Thus, a combination of several methods, i.e. several sources of evidence should provide a broader perspective and more accurate description (for the latter it is important that the errors are not made in similar ways by the methods). The first example of successful combination is the system *Estratto* [15] discussed earlier. *Estratto* combines the application of patterns with their evaluation based on statistical evidence. As a result, the output is not only a list of lemma pairs but each pair is described by its reliability value based on statistical evaluation.

Another example can be the WordNet Weaver system supporting linguists in extending the nominal part of the plWordNet thesaurus [26]. For each new lemma (i.e. not described in plWordNet yet), WordNet Weaver generates a set of subsets of the hypernymy structure (i.e. subgraphs comprising several linked lexical units each) as attachment suggestions. A suggestion expresses a possible area in the hypernymy structure in which one of the lexical units of the given new lemma should be added as a hypernym, hyponymy or a synonym. Suggestions are based on the combined evidence coming from 4 types of knowledge sources. The sources were all extracted for 13 285 Polish nominal lemmas from the joint corpus discussed in Section 4 and are characterised below:

- a high accuracy MSR based on the Rank Weight Function, see [26],
- post-filtering lemma pairs produced by the selected MSR with a classifier presented in [24] – the percentage of lexico-semantic relation instances increases among the filtered pairs
- manually constructed lexico-morphosyntactic patterns described in [26], including the pattern presented in Section 4,
- and the results generated by the *Estratto* algorithm application [14, 15] discussed in Section 4.

All knowledge sources, except *Estratto*, were extracted from the joint corpus discussed earlier. *Estratto* was applied only to the IPI PAN Corpus and the corpus of *Rzeczpospolita*.

WordNet Weaver has been successfully applied in expanding plWordNet by 15 200 new lexical units (8 700 new synsets). The manual workload was 3.4 person-months. We observed significant improvement in the pace of work in comparison to a purely manual work. Detailed evaluation of WordNet Weaver on the basis of the analysis of the work of linguists, as well on the basis of automated tests can be found in [26]. Here, we would like to highlight only that in the case of 75.24% of new lemmas at least one generated suggestion was found to be helpful by the linguist.

WordNet Weaver is an example of the application of the extraction methods as a support for the construction of thesauruses. The automated methods can also serve as critiques of the existing thesauruses and lexicons facilitating discovery of lacking informations or identifying potential errors (cf [26] for the experience of this kind collected with WordNet Weaver).

A MSR can be used to generate clusters of lemmas associated semantically in the domain represented by the corpus. The automated methods can facilitate analysis and comparison of systems of lexical meanings characteristic for particular domains represented by domain corpora.

The automated methods supports naturally a data-driven approach in which data collected in a large corpus are the primary source. Thus, they deliver a perspective which is objective to the extent of the corpus dependence, but not influenced by the subjective interpretation of a researcher. In [21], we presented a comparison of the lemma associations obtained from surveys of human informants vs the associations extracted from the IPI PAN Corpus with the help of a MSR. The intersection of both lists is around 20% only, but the automatically extracted data presents also a valuable perspective on the understanding of Polish *collective symbols* and *flag words*.

## 7 Perspectives

In spite of more than 20 years of history of Distributional Semantics methods, there is still room for further development. First, still larger and larger corpora are available for many languages lacking such corpora earlier. Our experience with applying different types of contexts, e.g. [23, 25], shows that one can expect better accuracy of MSRs based on a deeper lexico-syntactic analysis. In the case of Polish, we have not had any robust parser at our disposal yet. Knowing the limitations of the lexico-morpho-syntactic constrains applied, we tried to increase their accuracy for the price of the unavoidable decrease in their recall. However, in this way only a portion of information included in text has been utilised. There are several potential factors influencing negatively the accuracy of MSR but probably the most important one is the fact that MSR is constructed for lemmas on the basis of evidence collected from lemma occurrences. As a result, in the majority of cases one sense, or most two senses (lexical units), dominate in the upper part of the list of the most semantically related lemmas generated for a given lemma. However, a potential improvement would require previous disambiguation of the corpus with respect to word senses (a task which is more difficult than the MSR construction) or a new method of MSR construction in which both processes, i.e. meaning extraction and word sense delimitation would be performed in parallel.

Combined methods seem to be in their initial stage of development, especially multi-criteria methods of automated resource construction, like WordNet Weaver [26] or [33]. We should see an intensive development of this subfield.

The extraction methods are gradually becoming more accessible to users not familiar with the technical details of the Language Technology, e.g. to linguists, as there are various attempts to make the language tools available and accessible to such users, e.g. the Clarin project<sup>4</sup>. Among the goals of Polish part of Clarin is to make the SuperMatrix system [3] available as a part of this research infrastructure. Among other options, SuperMatrix delivers tools for the pattern-based approach and Distributional Semantics. It was used for the generation of the majority of examples presented in this paper. Accessibility of the technology should boost the cooperation between linguists and researchers working in the area discussed here. Such a cooperation is needed.

Measure of Semantic Relatedness can be perceived as defining a specific perspective on lexical semantics.

An important future area of development for the extraction methods are multilingual applications i.e. extraction of lexical semantic relations and MSR from multilingual corpora in a way synchronised or mutually related. Potential results should be valuable for the construction of multilingual resources for lexical semantics and applications of the Language Technology, as well as possible linguistic comparative studies.

## Bibliography

- [1] Apresjan, J. D. (2000). *Semantyka leksykalna. Synonimiczne Środki języka [Lexical semantics]*. Warszawa. translated by Z. Kozłowska and A. Markowski.
- [2] Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

<sup>4</sup> [www.clarin.eu](http://www.clarin.eu)

- [3] Broda, B. and Piasecki, M. (2008). SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. In Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., and Trojanowski, K., editors, *Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, Advances in Soft Computing, pages 345–352, Warsaw. Academic Publishing House EXIT.
- [4] Calzolari, N., Cardie, C., and Isabelle, P., editors (2006). *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics.
- [5] Derwojedowa, M., Piasecki, M., Szpakowicz, S., and Zawisławska, M. (2009). plWordNet — The Polish Wordnet. Online access to the database of plWordNet: [www.plwordnet.pwr.wroc.pl](http://www.plwordnet.pwr.wroc.pl).
- [6] Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, Concepts and Relations in the Construction of Polish WordNet. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Global WordNet Conference, Seged, Hungary January 22–25 2008*, pages 162–177. University of Szeged.
- [7] Dowty, D. R. (1979). *Word Meaning and Montague Grammar*, volume 7 of *Synthese Language Library*. D. Reidel Publishing Company, Dordrecht:Holland/Boston:U.S.A./London:England.
- [8] Dowty, D. R., Wall, R. E., and Peters, S. (1981). *Introduction to Montague Semantics*, volume 11 of *Synthese Language Library*. D. Reidel Publishing Company, Dordrecht: Holland/Boston:U.S.A./London:England.
- [9] Fellbaum, C., editor (1998). *WordNet – An Electronic Lexical Database*. The MIT Press.
- [10] Fromkin, V., Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., Koopman, H., Keating, P. A., Munro, P., Hyams, N., and Steriade, D. (2000). *Linguistics: An Introduction to Linguistic Theory*. Blackwell Publishing.
- [11] Harris, Z. S. (1968). *Mathematical Structures of Language*. Interscience Publishers, New York.
- [12] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*, pages 539–545, Nantes, France. The Association for Computer Linguistics.
- [13] Hearst, M. A. (1998). *Automated Discovery of WordNet Relations*, chapter 5, pages 131–151. Volume 1 of [9].
- [14] Kurc, R. (2008). Automatyczne wydobywanie leksykalnych relacji semantycznych na podstawie prostych wzorców syntaktyczno-leksykalnych. Master's thesis, Faculty of Computer Science and Management, Wrocław University of Technology.
- [15] Kurc, R. and Piasecki, M. (2008). Automatic Acquisition of Wordnet Relations by the Morpho-Syntactic Patterns Extracted from the Corpora in Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, pages 181–188.
- [16] Manning, C. D. and Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- [17] Mel'čuk, I. (1988a). *Dependency Syntax: Theory and Practice*. State University of New York Press.
- [18] Mel'čuk, I. A. (1988b). *Dependency Syntax: Theory and Practice*. State University of New York Press.
- [19] Mohammad, S. and Hirst, G. (2005). Distributional measures as proxies for semantic relatedness.
- [20] Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In [4], pages 113–120.
- [21] Pawłowski, A., Piasecki, M., and Broda, B. (2008). Automatic extraction of word-profiles from text corpora. On the example of polish collective symbols. Prepared for the planned post-conference Proceedings of Trewir'07.
- [22] Piasecki, M. (2006). Handmade and Automatic Rules for Polish Tagger. LNAI. Springer.
- [23] Piasecki, M. and Broda, B. (2007). Semantic Similarity Measure of Polish Nouns Based on Linguistic Features. In Abramowicz, W., editor, *Business Information Systems 10th International Conference, BIS 2007, Poznan, Poland, April 25-27, 2007, Proceedings*, volume 4439 of *Lecture Notes in Computer Science*. Springer.

- [24] Piasecki, M., Marcińczuk, M., Szpakowicz, S., and Broda, B. (2008). Classification-based Filtering of Semantic Relatedness in Hypernymy Extraction. In *Proceedings of the GoTAL 2008 Conference*, LNAI. Springer.
- [25] Piasecki, M., Szpakowicz, S., and Broda, B. (2007). Extended Similarity Test for the Evaluation of Semantic Similarity Functions. In Vetulani, Z., editor, *Proceedings of the 3rd Language and Technology Conference, October 5–7, 2007, Poznań, Poland*, pages 104–108, Poznań. Wydawnictwo Poznańskie Sp. z o.o.
- [26] Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej. To appear.
- [27] Piotrowski, T. and Saloni, Z. (1999). *Kieszonkowy słownik angielsko-polski i polsko-angielski*. Wyd. Wilga, Warszawa.
- [28] Przepiórkowski, A. (2004). *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS.
- [29] Pustejovsky, J. (1991). Generative lexicon. *Computational Linguistics*, 17(4):409–441.
- [30] PWN (2007). Słownik języka polskiego. Published on the web page.
- [31] Rzeczpospolita (2008). Korpus Rzeczpospolitej. [on-line] [www.cs.put.poznan.pl/dweiss/rzeczpospolita](http://www.cs.put.poznan.pl/dweiss/rzeczpospolita). Corpus of text from the online edition of Rzeczpospolita.
- [32] Sahlgren, M. (2001). Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. In *Proceedings of the Semantic Knowledge Acquisition and Categorisation Workshop, ESSLLI 2001*, Helsinki, Finland.
- [33] Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In [4].
- [34] Wierzbicka, A. (2006). *Semantyka. Jednostki elementarne i uniwersalne*. UMCS.
- [35] Zesch, T. and Gurevych, I. (2006). Automatically Creating Datasets for Measures of Semantic Relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24, Sydney, Australia. Association for Computational Linguistics.

# Interactive Discovery of Ontological Knowledge for Modelling Language Resources

— prolegomena —

André Włodarczyk

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<sup>1</sup>. 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

---

<sup>1</sup> 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 1990<sup>th</sup> 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 optimisation 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<sup>2</sup> 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<sup>3</sup> 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

<sup>2</sup> See however the paragraph 4.1 below.

<sup>3</sup> 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 Włodarczyk 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 losing 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.

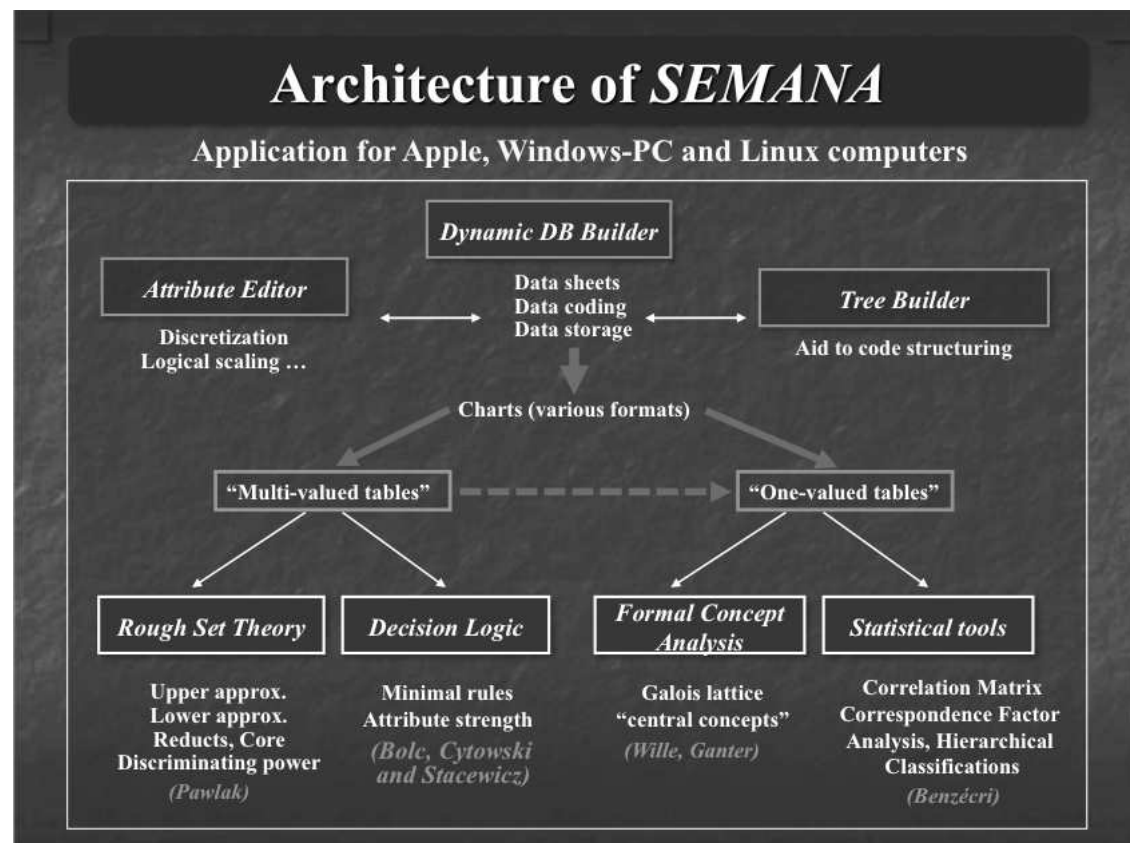
On the other hand, the difficulty of data input takes also its origin partly in the fact that linguistic expressions in context have also implicit meaning and entail as well presupposed as inferred knowledge. Namely, it is difficult to establish which part of the presupposed or inferred knowledge

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.

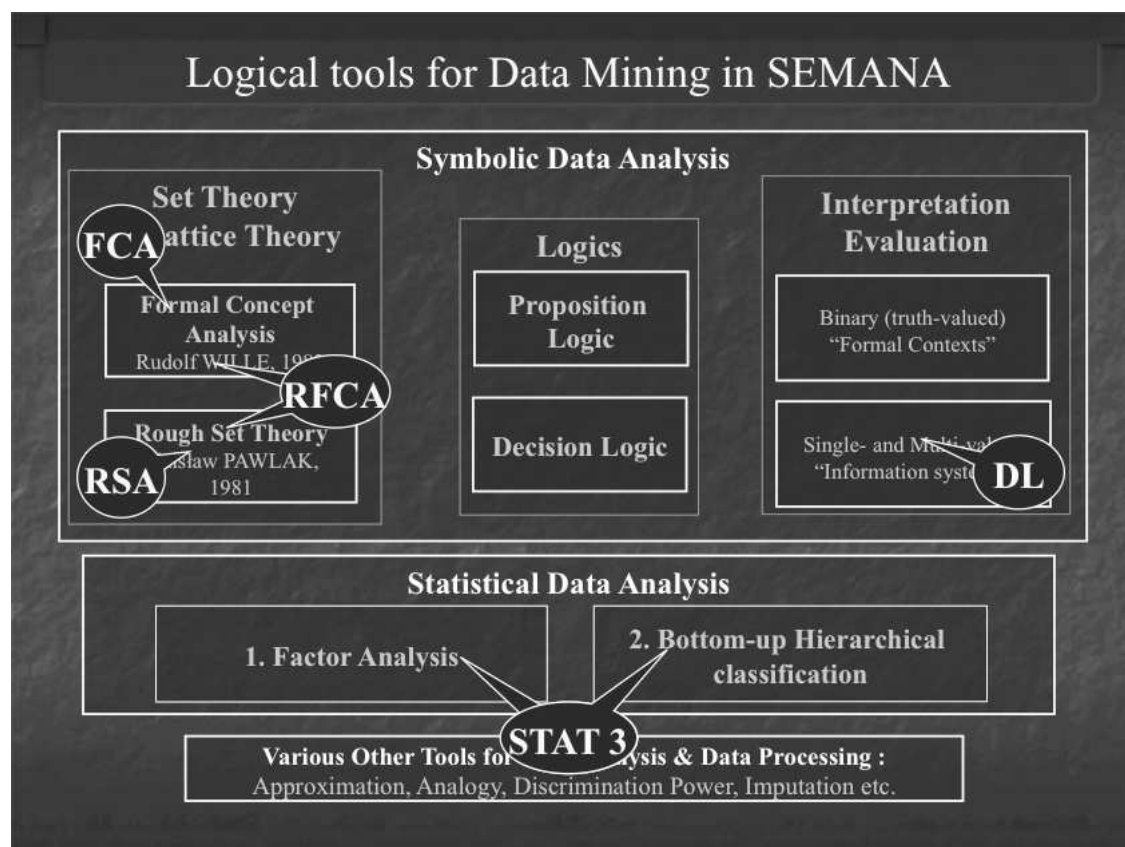
#### a) Symbolic Analysers

- 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 inconsistencies (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*<sup>4</sup> 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: *explanans* and *explanandum*. In this pair of notions, it is the definiens of the definition that corresponds to the explanandum of an explanation. Explanans guarantees mutual dependencies between conjunctions of partial definitions. However, we consider that, in order to make the explanans play this function, it is necessary that the concepts which represent the explanans 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)

<sup>4</sup> 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 authors. 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*<sup>5</sup>, *rough set information system analysers*<sup>6</sup>, *ascending hierarchical classifiers and correspondence factor analysers*<sup>7</sup>, 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<sup>8</sup>. 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*<sup>9</sup>  $S$  as a set of formulae or attributes or definitions and *Assignment*  $A$  as an assignment function from uses to semes ( $A: U \rightarrow S$ ).

$$\mathbf{Sign} = \langle U, S, A \rangle$$

<sup>5</sup> Cf. Wille R. (1982, 2001), Ganter B. & Wille R. (1999).

<sup>6</sup> Cf. Pawlak Z. (1981), Orłowska E. & Pawlak Z. (1984).

<sup>7</sup> Cf. Benzécri, J.-P. (1984), Jambu, M. (1978) and Greenacre, M. (1983).

<sup>8</sup> 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**<sup>10</sup> 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*<sup>11</sup> function from uses to semes and vice versa (Wille, R. – 1982).

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

Let  $S$  be a set of semes<sup>12</sup>  $S = \{\alpha, \beta, \gamma \dots\}$  and let  $\Sigma$  be a subset of semes in  $S$  ( $\Sigma \subseteq S$ ). The **Usage** of a concept is defined as its extension.

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

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

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

**Informal definition 3.** We will call *semion*<sup>13</sup> 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	<b>Usage</b>	<b>Sense</b>
Semion	<b>Use</b>	<b>Sememe</b>

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*<sup>14</sup> 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

<sup>10</sup> In Formal Concept Analysis, the term used is Formal Concept (Wille R. — 1982, 2001).

<sup>11</sup> Our presentation of this problem is very succinct. The dual character of formal concepts lies at the basis of the algebraic structure of lattice representation (cf. the literature which is now very rich on FCA — Formal Concept Analysis).

<sup>12</sup> Note also that the original Wille's terminology significantly differs from ours because we limited our theory only to the semiotic objects. What we call semes, Wille calls attributes.

<sup>13</sup> 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).

<sup>14</sup> 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<sup>15</sup> the individualized ontologies<sup>16</sup>. 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** ( $\pm A \rightleftharpoons \mp B$ ) 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):

$\mathbf{wa}^+$  **Topic** /  $\mathbf{ga}^-$  **Subject<sub>old</sub>**  $\rightleftharpoons$   $\mathbf{ga}^+$  **Focus** /  $\mathbf{ga}^-$  **Subject<sub>new</sub>**

(b) the Polish verb past morphemes *-li* and *-ły* have two<sup>17</sup> senses each (Włodarczyk H. — 2009):

<sup>15</sup> As a matter of fact, this function from goes across the internal semantic representations.

<sup>16</sup> From our perspective, the semantic interpretation function of linguistic expressions should be characterised by both refinement and blending.

<sup>17</sup> If we consider that the neutre gender’s meaning of nouns which refer to animate beings is *–Adult*, the number of binary oppositions, in this case, amounts to three (Włodarczyk, H. – 2009).

- (1)  $-li^+ + \text{Human} / -ty^- + \text{Human}$   
 (2)  $-ty^+ + \text{Feminin} / -li^- - \text{Feminin}$

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<sup>18</sup> 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, 4<sup>ème</sup> éd. (1<sup>ère</sup> 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.

<sup>18</sup> Włodarczyk A. & Włodarczyk 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.
- Włodarczyk, 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
- Włodarczyk, A. (2003)** "Les Cadres des situations sémantiques". *Études Cognitives / Studia Kognitywne V*, Warszawa: SOW Publishing House, 35-51.
- Włodarczyk, A. (2005)** "From Japanese to General Linguistics — starting with the 'wa' and 'ga' particles", *Paris Lectures on Japanese Linguistics*, Kurocio Shuppan, Tokyo
- Włodarczyk, 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/>

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

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

**Włodarczyk, H. (2009)** “Lingwistyka na polonistyce krajowej i zagranicznej w dobie filozofii informatyczno-logicznej” (Linguistics in Polish Studies 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.



# Ontological Issues for Modelling Aspect Semantics\*

Hélène Włodarczyk<sup>1</sup>

CELTA - Centre de Linguistique Théorique et Appliquée  
helene.wlodarczyk@paris-sorbonne.fr

**Abstract.** The CASK method (Computer-aided Acquisition of Semantic Knowledge) is being used at CELTA with the SEMANA (SEMantic ANalyser) software, which was especially designed for this purpose. We use the notation of semantic feature structures as a meta-language for describing the category of aspect in Polish regardless of its pertaining to various levels of expression (morphological, syntactical or lexical). To model the category of aspect we treat types of situations as conditions for two relevant aspect parameters we call analysis of the situation (or internal aspect) and control of the situation (or external aspect). In order to cope both with the lexical diversity of aspectual morphemes (prefixes) and the grammatical (dichotomous and obligatory) character of aspect in Slavic languages, we have proposed to define the perfective aspect as a hypercategory. Our database of Polish aspect was analysed using KDD statistical tools (G. Sauvet 2008, <http://celta.paris-sorbonne.fr/anasem/papers/>). We obtained the first preliminary experimentally fixed semantic definitions of the perfective and imperfective values of the category of aspect in a Slavic language.

## 1 Ontology and Semantics

Multi-lingual contrastive studies need to refer to ontologies as *tertium comparationis*, including those which were especially designed in order to account for linguistic objects (expressions). We claim that this task cannot but be interactive (computer-aided), if we want to avoid the lack of precision and the variability of semantic parameters in traditional linguistic research, especially hazardous in the semantic domain.

We define the semantic content of a linguistic expression as a function mapping this expression on ontological concepts. Therefore, to describe the semantic content of aspectual expressions used in utterances we need to specify the ontology to which they refer. A formal cognitive description aims at giving an ontological account of a semantic category by treating its definition in different languages as a finite set of strictly defined feature structures. We aim at giving a two-fold account of semantic categories by:

- building a general set of ontological abstract structures necessary to interpret aspect in different languages
- choosing a specific set of semantic feature structures for the described category in each human language, in our case for Polish.

To describe aspectual values of verbs in context, we had first to define relevant aspectual semantic attributes and their values (AV). At this stage, linguists must use both general knowledge of the category they are studying and specific knowledge of the language they are describing.

## 2 Computer-aided Acquisition of Semantic Knowledge

Our experience consists in using a software for the acquisition of semantic knowledge (SEMANA designed at CELTA) after years of traditional linguistic research on the topics of aspect. We have

---

\* Acknowledgment: the research on aspect with the SEMANA software has been conducted at CELTA (Paris-Sorbonne University) since 2005 by Hélène and André Włodarczyk (ontology and semantics of verbal aspect), Georges Sauvet and André Włodarczyk (conception of the SEMANA software), Georges Sauvet (analysis of the aspect database with SEMANA KDD tools). Doctoral and master students also take part in this research.

been conducting this research in the field of verbal aspect semantics in Slavic languages in contrast with Indo-European non-Slavic languages, mostly French and English. We turn to KDD methods in order to enhance linguistic research in two ways. 1) using a software as Semana can help grasping a large set of semantic features and make clear the relations between them, i.e. the structure or system they belong to ; 2) it provides the linguist with a universal (feature) language whose rules are logico-mathematical and with tools worked out by computer scientists to handle this language and perform calculation on it. And last but not least, a software offers the possibility to store large language data, to access them easily and share them with other researchers.

## 2.1 Towards Experimental Semantics

The method of Computer Aided Acquisition of Semantic Knowledge (CASK) is based on the idea that the meaning of linguistic units can be described only in context. In a given context (inside a text or discourse and in a particular speech situation) an expression is not ambiguous (it can be described by one feature structure). What is called ambiguity or polysemy is the possibility for an expression to be used with different senses in different contexts. For this reason, it is important to collect numerous samples of usages (build semantic databases), so that linguistic theories consist in describing and analysing these data with symbolic and statistical methods.

This approach differs radically from hypothetico-deductive linguistic theories (which use merely a few examples as illustrations) but it is confronted with the serious problem of meta-data input. In fact, we consider that bare facts do not exist as such, any description of linguistic “facts” relies on a chosen theoretical background; for this reason, we call the linguistic data we collect in our database “meta-data”.

At the stage of collecting and marking semantic data, the intuition of the linguist is irreplaceable although fallible. Interaction between man and machine (consisting of handling a list of fixed well defined monosemous features that demand conscious intervention to be modified or enlarged) can prevent from the subjectivity and variability of human appreciation of the meaning of expressions.

The problem of data input comes partly from that linguistic expressions exhibit in context not only explicit meaning, but also entail as well presupposed as inferred knowledge. It is sometimes difficult to establish which part of the presupposed or inferred knowledge is pertinent in a given context and should be taken into account in the description.

## 2.2 SEMANA: a Software designed for semantic research

The problem of communication between linguists and computer scientists comes from the times when the latter used to analyze the formers’ needs in their specific field at a given stage of their research in order to offer them adapted tools. Most often such tools reflected the image (knowledge) of the field at the time of the programmer’s analysis but were not flexible enough to be easily modified when the specialist’s view of the field changes. The SEMANA software<sup>1</sup>, especially designed at CELTA for the CASK project (Computer-aided Acquisition of Semantic Knowledge) offers two sorts of tools: (1) tools for interactive intelligent and dynamic database designing; (2) tools for automatic KDD research.

SEMANA’s Dynamic Database Builder (db builder) is highly interactive, relying on the linguist’s expertise in a given domain; it makes it possible to change features and their values (and the structure they belong to) as soon as progress in research proves it necessary. Each card in the db builder contains a field for the specimen described (an utterance chosen from a corpus) and a field with a list of attributes and values from which the linguist chooses the appropriate values for the sample he is describing. The characteristic morpheme of the analysed expression is used as index. The db builder is completed by a Tree Builder Assistant which allows the linguist to organise in a tree structure the attributes and values chosen for the description of a semantic field. Any change in the feature tree of the Tree Assistant is transferred to each record of the database after the linguist is asked whether he accepts changes in the tree builder to be echoed in the database. Since

<sup>1</sup> <http://www.celta.paris-sorbonne.fr/anasep/papers/>

the beginning of the research with Semana we could change several times the tree of attributes used to describe aspect in Polish. All specimens are automatically collected in a contingency table. The synthetic table has the form of a chart of attributes and values for each sample described in the database. In the synthetic table, linguists can observe which samples present the same value of the same attribute. For linguistic description it is an important assistance : it makes it possible to verify whether the same attribute and value were chosen rightly in different contexts and at different times of data description by the linguist. It may seem trivial but, in semantic annotation, the choice of attributes and values is very sensitive to narrow and broad context. This table is completed by tools which provide statistical information about the use of attributes and their values and suggest interactive restructuring of the attributes and their values: it checks objects with the same AV (duplicates), proposes to merge 2 or more attributes, shows types of objects by attributes or by values, checks the AV field (appearing in each card) and the feature tree (in the Tree Builder Assistant). Many automated checking procedures are available and help the linguist to check the consistency of his reasoning.

The second part of SEMANA consists in tools for KDD research integrating the following methods: Rough Set Theory (RST, Z. Pawlak), Formal Concept Analysis (FCA by R. Wille and B. Ganter), Statistical Data Analyses (by J.-P. Benzécri). These tools are described in André Włodarczyk's article in the same volume: "Interactive Discovery of Ontological Knowledge for Modelling Language Resources".

### 3 The Categorial and Contextual Meaning of Aspect

As regards aspect, it is often a very delicate task to distinguish the explicit categorial aspect meaning of an expression used in an utterance from its inferences and presuppositions<sup>2</sup>. The latter are referred to globally as the "pragmatics" of aspect or its conversational implicature. In my view, what is called the pragmatics of aspect refers to two different kinds of Aspectual uses in context.

In the first place, properly aspecto-temporal senses of verbs in context must be taken into account. Such senses do not depend only on the categorial meaning of the aspect category in the language system but also on the utterance in which the verb is used. For instance, in some languages, uses of verb forms expressing that a process has reached its finish point in the past (before the speech time point) often allow in context to infer that the situation is in its *after* stage (in a new state resulting from the process expressed by the verb) thus producing what is called a *resultative* meaning. e.g. In Polish *Zmarzłem* (*I have got frozen*) means *I am now cold* (at the speech time) or *was cold* (at the time period serving as reference point). Such kind of inferred meaning properly relies on ontological knowledge and reveals important in translation. Since every utterance content is partial as regards the situation it refers to, the explicit/implicit part of an utterance content is not always the same in two different languages. When translating, it is sometimes necessary to replace the implicit inferred meaning of the original expression by an explicit expression in the target language. As a matter of fact, the aim of the translation is to produce an expression which refers to some ontological knowledge that is similar to that of the original expression.

Secondly, we claim that the properly *pragmatic* meaning of aspect is related to the meta-informative (sometimes called cognitive) *old* or *new* status of the utterance. We have devoted and continue to devote much attention to this part of the aspectual problem, which reveals very important when contrasting languages (cf. Włodarczyk H. 1997, Włodarczyk A. & H. 2008). In this paper we do not develop this problem because, for the time being, in our database, we limit the description of aspect uses to the first sort of uses, i.e. properly aspecto-temporal uses.

<sup>2</sup> Much has been written about this problem by Jakobson 1932, 1936 (Gesamtbedeutung) , Bondarko 1971a, b, and others Primary and Secondary meanings of aspectual forms (Kuryłowicz 1977)

## 4 Aspect as a Hypercategory

Aspect in Slavic languages is based on the opposition of two values only (perfective and imperfective) but the aspectual opposition is expressed not only in aspectual pairs but also in what we call aspectual families (or derivational nests). In Slavic languages, all prefixed verbs derived from a simple verb become perfective<sup>3</sup>. Among them, two classes can be distinguished although the border between these two classes is not sharp: verbs with a new lexical meaning (lexical derivatives) and verbs, which keep the same lexical meaning but change aspect (aspectual derivatives). Among the aspectual derivatives, slavists distinguish traditionally (since the beginning of the 20th century, cf. Agrell 1908) between a series of derived verbs called *Aktionsart* (or *lexical aspect*) and *one* derived perfective verb considered as *the grammatical* perfective partner of the simple verb. Recently, this distinction has been reappraised by many researchers (including, Sémon 1986, Paduceva 1996, Karolak 1997, Xrakovskij 1997). Moreover, much work has been done to prove that not all so-called “aspectual pairs” are semantically similar, i.e. the opposition of perfective and imperfective may be based on different sets of semantic features depending on the lexical meaning of verbs (more precisely, depending on the ontological type of situation as defined below). The first pioneer work in that direction was that of Cezar Piernikarski (1969) who showed that there exist several different types of “aspectual oppositions”.

Our ontological approach makes it possible to bring closer the two sub-categories of aspect and *Aktionsart* and to replace the notion of “aspectual pair” by that of aspectual nest when defining the perfective aspect of Slavic languages (Włodarczyk A. & H. 2001). Most of the meanings traditionally assigned to the category of *Aktionsart* are part of what we call **control** parameters (see below). In prefixed verbs, these meanings combine with those considered as strictly grammatical perfective meanings and are characteristic of the aspectual nest that can often be derived from a simple imperfective verb. In former theories of aspectual pairs, only prefixed perfective verbs with a *resultative* meaning were considered *pure grammatical* perfective partners of a simple imperfective verb. In our view, we consider as aspectual pairs only those pairs consisting of a perfective verb and the suffixed imperfective verb that is derived from it, e.g. *przepisać* (perf.) / *przepisywać* (imp.), *to copy*.

Perfective as a hypercategory (a two-level category due to the derivational origin of aspectual morphemes in Slavic languages) subsumes all meanings of so called *pure* aspectual partners and *Aktionsart*. The concept of hypercategory allows us to describe each derived perfective verb as inheriting several aspectual features. Following this hypothesis, no verbal prefix can be viewed as semantically void because different configurations of features are always at hand. As a matter of fact, none of the perfective partners of a simple imperfective verb can be considered as entirely “synonymous” to the root verb.

Using the database of the Polish Frequency Dictionary (SFPW) we studied the relative frequency of simple imperfective verbs and the different perfective verbs derived from them (Włodarczyk A. & H. 2001). It appears that the frequency of the so-called “pure perfective partner” (very often the one with *resultative* meaning) is much higher than the frequency of verbs considered as *Aktionsart* and this is probably the reason why this perfective partner was generally considered as *the only pure grammatical perfective partner* of the simplex imperfective verb. Our treatment of aspectual prefixed verbs can be regarded as one more contribution to the long-lasting discussion about “aspectual pairs” in Slavic languages. We fully agree with the opinion of slavists who consider (although mostly on the ground of other arguments) that pairs are constituted only by an imperfective suffixed verb that is derived from a perfective verb, e.g. *przepisywać* imp. from *przepisać* perf. (both can be translated as “to copy”), *zamawiać* imp. from *zamówić* perf. (both can be translated as “to order”), etc.

The notion of hypercategory is derived from the idea that categories can be classified. As shown above, the aspectual meaning of a verb used in a given context can be described as a bundle of attributes. By using inheritance, we do not have to build disjoint classes of aspectual meanings (as it was the case with classes of *Aktionsart* verbs). In fact, aspectual meanings are irregularly

<sup>3</sup> Exceptions are extremely rare and due to diachronical reasons.

linked to various superior nodes. Moreover, one and the same verbal lexeme may have different links depending on the context in which it is used. This is often the case of *po-* prefixed verbs that can have several meanings (cf. Włodarczyk A. & H. 2001, II).

This approach allows us to take into account the lexical diversity in Slavic aspectual semantics and sheds light on the controversies about the grammatical or lexical nature of aspectual features. We conclude that in the aspectual derivation in Slavic languages there is no clear border between the lexical and the grammatical. Moreover, there is no need neither to consider this situation as “incomplete grammaticalisation”. On the contrary, this endows Slavic languages with a very systematic way of expressing a broad range of aspectual nuances

## 5 Three Kinds of Aspectual Parameters and Formal Definition of Aspect

The aspectual attributes and values we use in the database are the result of previous research on aspect (Włodarczyk H. 1997) and are defined in the framework of the theory of aspect we outlined for interactive semantic research (Włodarczyk A. & H. 2003, 2006). We adopted the notation of *semantic feature structures* as a meta-language for describing aspectual meanings in various languages regardless of linguistic levels (morphological, syntactical, lexical etc). We propose to describe the meaning of the aspect category as a pair of feature bundles : *Analysis* and *Control*. The “analysis of a situation” (viewed as a whole or as one of its moments or stages) is considered as its *endocentric aspect*, concerning the internal development of a situation as time passes by. We call the “control of a situation” a set of operations such as iteration, modifications of flow or intensity, composition of two or more situations into one. This control parameters are imposed on a situation from outside of its internal development in time and therefore we consider them as *exocentric aspect*. Moreover these internal and external aspectual features occur and combine diversely depending on the semantic type of the situation to which a verb is related in a given context. Each aspect use can therefore be described by a semantic feature bundle consisting of two parts: situation analysis and situation control. The situation type is considered as a condition for the use of aspect. This allows to propose a formal definition of aspect as follows.

Aspect = { Sit. Analysis , Sit. Control } condition : Sit. Type

### 5.1 Types of Semantic Situations

The classification we use (Włodarczyk A. 2003) differs from other classifications (Vendler Z. 1967, Laskowski R. 1998, among others) in that it does not incorporate features concerning situation participants and roles but only situation frames; it is based on four relevant features: space (three dimensions), time, progression and granularity.

SEMANTIC TYPES OF SITUATIONS				
Characteristic properties (dimensions)	Static Situations	Dynamic Situations (ACTIONS)		
	STATE	EVENT	Ordinary PROCESS	Refined PROCESS
Space (3D)	+	+	+	+
Time	–	+	+	+
Progression	–	–	+	+
Granularity	–	–	–	+

Table 1: Hierarchy of semantic situations (Włodarczyk A. 2003)

States differ from dynamic situations, in that they are not modified by time passing. Events have no progression: their beginning moment (“start”) coincides with their ending moment (“finish”) and it is impossible to observe an intermediary stage (“run”) between them. In other words, events are dynamic situations without progression or development. Progression characterises processes, which develop in time from a *begin* stage to an *end* stage through an intermediary stage we call *run*. A granular process is made up of a repetition of many identical grains. Situations are hierarchically ordered: each type of situation inherits properties of the preceding type.

## 5.2 Internal Aspect: Situation Analysis

Any situation is preceded by a preceding situation (we call the “before” stage) and a following situation (we call the “after” stage). But only processes, i.e. dynamic situations with progression, may be further analysed into inner moments and stages (Fig. 1). Static situations may have only a start and a finish moment without any further analysis between them. Events are dynamic situations conceived as instantaneous (without progression) and therefore they are described as only one moment into which the start moment and the finish moment are merged.

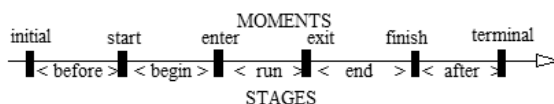


Fig. 1: Analysis of simple dynamic situations in moments and stages (constitutive parts)

The distinction that we make between moments and stages (represented on the figure as points and segments) may remind of the geometric metaphor of a “point” as opposed to a “line” often used in aspectology but, in our approach, we do not identify totally perfectivity with the point-view and imperfectivity with the segment-view. As a matter of fact, we consider the selection of a moment or a stage as only *one* of the semantic attributes of the aspect category. This attribute is combined in different configurations with other parameters in order to give account of different usages of the perfective and imperfective verbs.

A situation characterised as a process may be roughly analysed in three inner stages: *begin*, *run* and *end*. Moments serve as boundaries for stages, and we called them (arbitrarily) *initial*, *start*, *enter*, *exit*, *finish* and *terminal*. What is relevant in our theory is not the intuitive meaning of these words in English but the place they mark on the line representing the progression of a process in time

## 5.3 External Aspect: Control

Exocentric aspect (“control”) consists in a set of parameters that are combined with the endocentric ones. The control may concern the repetition of the situation, the modification of the flow or intensity (*interrupt*, *resume*, *keep*, *off-and-on*<sup>4</sup>, *trans*<sup>5</sup>), the composition of two (or more) sequential or parallel situations into one complex situation. In Polish, the composition of situations is expressed in the case of verbs with prefixes indicating that the situation is composed of two or several situations. As an example we may quote the so called *distributive Aktionsart*: situations performed simultaneously or successively by different subjects or on different objects are composed into one complex situation, e.g.:

*Pootwieriałem wszystkie okna.* (Lit. *I opened all the windows one after the other.*)

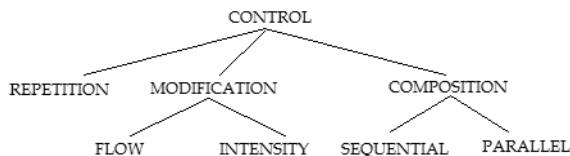


Fig. 2 Aspectual Parameters of Situation Control

Many of what we call control parameters were previously described as *limitative*, *intensive*, *iterative*, *distributive*, *completive* etc. “*Aktionsarten*”. However, each type of aspectual meaning that was previously called an *Aktionsart* was generally described by only one single (often dichotomous) label; in our approach, such meaning pertains to more than only one parameter because we define

<sup>4</sup> The “off-and-on” flow modification concerns the unfolding of a situation intermittently.

<sup>5</sup> We call “trans” the unfolding of the whole situation from start to finish (regardless of its stages)

aspect at least by the pair of two sorts of parameters: *analysis* of the situation and its *control*. Generally, in one verb usage, at least one control parameter (repeated or not repeated) or more (modification, composition) may be at hand, thus every aspect usage is defined by a structure of several semantic features.

## 6 Interactive Semantic Research on Aspect with the SEMANA Software

Hereafter we sketch out the interactive research on aspect both from the point of view of the linguist (aspect database building by Włodarczyk H.) and the computer-scientist (KDD analysis of the aspect database by Sauvet G).

### 6.1 From the work on the ontological tree of aspectual features

In the aspect database we use a tree of the attributes and values (AV) described above to annotate each aspectual specimen chosen in a corpus. This tree was first designed in the Tree Assistant as an ontology of aspect and modified several times as we collected more and more samples.

Let us quote just an example of the possibility of modifying the tree of ontological features during the collecting of data. In the first version of the aspect feature tree, the attribute ASPECT ANALYSIS was divided into *inner* and *outer* moments and stages:

```
ASPECT-*_ANALYSIS-*_MOMENT-----*_MINN-----*_{AMI}=[enter|exit|finish|start]
*           *           *_MOUT-----*_{AMO}=[initial|terminal]
*           *_STAGE-----*_SINN-----*_{ASI}=[begin|end|run]
*           *           *_SOUT-----*_{ASO}=[before|after]
```

After we collected data, both *outer moments* (*initial*, *terminal*) that were never used in the db were deleted. This led us to simplify the attributes moments and stages as follows

```
ASP-*_ANLS-*_MOM--*_MOMI-*_{AMI}=[ent|exi|fin|str]
*           *_STG---*_SI-----*_{ASI}=[beg|end|run]
*           *           *_SO-----*_{ASO}=[bef|aft]
```

On the contrary both outer stages were frequently used in the description : we can understand that linguistic aspectual expressions of situations take into account the situation itself and its immediate bordering situations (before and after) but does not point at any dividing moment between the preceding situation (what we call the before stage) and another even more anterior situation because this would lead to an infinite regression (and the same for the moment we called terminal as end of the after stage). Thus, only the moments (schematized by points on a line) indicating the border between the outer stages and the situation itself can be expressed in verbal expressions of aspect : the moment we call *start* constitutes both the end of the *before* stage and the beginning of the *begin* stage of the situation, and the moment we call *finish* is both the last moment of the situation (of its *end* stage) and the first moment of the *after* stage.

### 6.2 Consistency checks

The field *index* of the db builder contains the aspectual morpheme of the described expression: a prefix or a suffix, or a periphrastic aspectual expression (e.g., an aspectual verb as *zaczynać*, “to begin”, *kończyć* “to stop” or *nie przestawać* “do not stop”, an adverb as *wciąż*, “continuously” etc).

Each specimen is characterised by a set of AV and by its morpheme (used as index). It may be written as a rule: **if {given set of AV} then index**. This allows index consistency to be detected. As a matter of fact, in our database, the test of consistency detected several different prefixes that were described by the same set of AV. However the polysemous character of verbal prefixes in Slavic languages and the two-step categorial structure of aspect (the hypercategory) requires that the linguist check the detected inconsistencies and accept them under the following conditions: (1) in the case when different prefixes share a common bundle of semantic features

(with the necessity for the linguist to add relevant distinctive features to give account of fine-grained semantic differences) or (2) when the same prefix is characterised by two or more different feature trees (polysemous prefixes).

### 6.3 Successive versions of the Aspect database

As we progressed by trials and errors in different versions of the aspect db we obtained gradually improved statistical reports. In Table X, column 2 displays the result of the automatic deletion of duplicates in the db. As we collect samples from text corpora it is obvious that this random access to data leads to the input of different samples having the same aspectual feature structure. In columns 3, 4 and 5, we can see that, as our work proceeded, we limited the number of attributes so that the number of theoretical combinations decreased and the possibility of merging attributes was reduced to zero.

DB version	Distinct objects	Number of attributes	Number of theor. combin.	Number of “merging attributes”
HW-Aspect-V1	61	12	2,064,384	9
HW-Aspect-V2	60	11	1,032,192	9
HW-Aspect-V3	77	11	829,000	6
HW-Aspect-V4	79	9	408,240	1
HW-Aspect-V5	79	8	136,080	1
HW-Aspect-V6	69	8	45,360	1
HW-Aspect-V7	74	8	61,440	0
HW-Aspect-V8	78	7	58,320	0

Table 2. Improvements reflected by statistical reports

### 6.4 Analysis of the first database of Aspect in Polish using SEMANA KDD tools

The feature tree used in the aspectual database (version 8) analysed by G. Sauvet was the following:

```
ASPECT-*--ANALYSIS-*--{ANA}=[enter|exit|finish|start|initial|terminal|begin|end|run|before|after|nonanalyzed]
*-CONTROL-*--FLOWMODIF-*--FLOW-----*--{MOD}=[interrupt|keep|resume|stop|trans|OffandOn|parallel|sequential]
*--REPETITION-*--{CRE}=[defnb|indnb]
*-INTENSITY-*--{ITS}=[increase|decrease|strong|weak]
ASPVALUE-*--ASPVAL---*--{VAL}=[imperfective|perfective]
*-MRPHCOMP-*--{MCP}=[impPrf|prfImp|prfImpPrf]
SITTYPE-*--{TYP}=[event|ordProcess|refProcess|state]
```

The multi-valued synthetic table corresponding to version 8 (fig. 3) was exported to the STA3 device of SEMANA and then automatically transformed into a one-valued table.

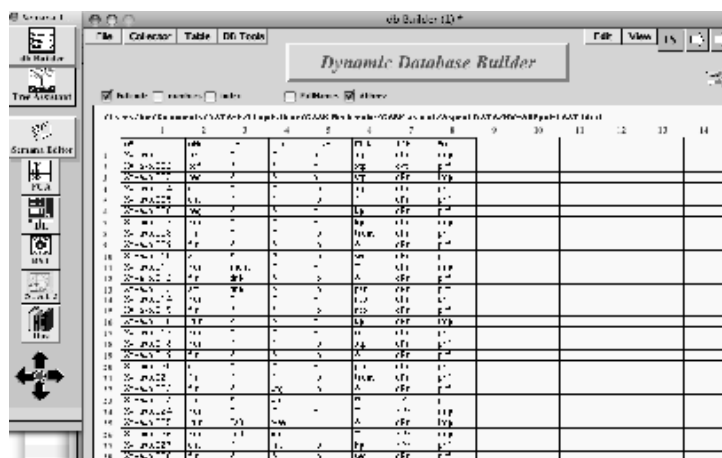


Fig. 3 Synthetic table of Aspect db Version 8



The Correspondence Factor Analysis (fig. 4) shows a clear partition of relevant features into two classes according to the attribute [VAL] = {perfective | imperfective}.

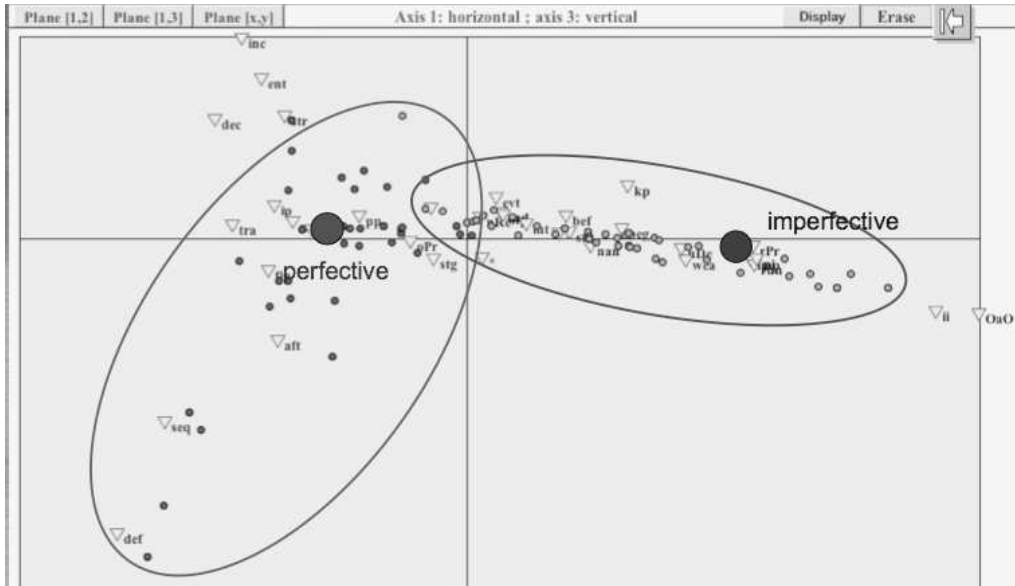


Fig. 4 Two classes of values of attributes around the perfective and imperfective

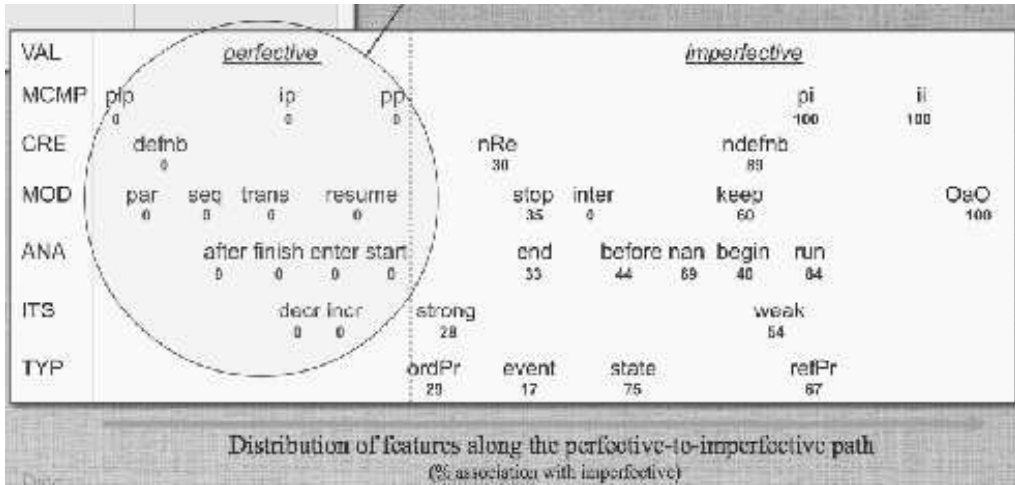


Fig. 5 Polish Aspect: Correspondence factor Analysis

The distribution of features along the perfective to imperfective path in Correspondence Factor Analysis (Fig.5) shows that a group of features imperatively requires the perfective value of aspect. Among them we find the *definite number of repeated situations* opposed to the *non definite number of repetitions* clearly situated in the imperfective zone. Three moment values of the attribute *situation analysis* are clearly associated mostly with perfective verbs: *start*, *enter* and *finish*; this captures the traditional view on perfective aspect as denoting either the end or the beginning of a process. On the other hand the values *nonanalysed* (nan) and the value of stage *run* of the same attribute are characteristic mostly of imperfective verbs which are known as able to feature situations as non analysed wholes or in progress in their run stage without taking into account any border moment, neither at the beginning nor at the end. As concerns situation types: ordinary processes are situated between the two zones (such type of situation may be expressed as well by an imperfective as a perfective verb) whereas refined processes appear clearly in the imperfective

zone. Among the different values of the *flow modification* attribute the *stop* and *interrupt* values are closer to the perfective whereas the *keep* and *off-and-on* values are closer to the imperfective.

These first promising results will have to be improved by collecting a larger amount of samples and defining some extra features in order to capture the nuances introduced into the perfective hypercategory by different prefixes. This task is carried on in the years 2008-2010 by a group of master students working on Polish prefixes at CELTA of Paris-Sorbonne.

## 7 Experimental Semantics and Language Contrasting by Alignment of Ontological Structures

With the possibility to use the techniques of knowledge discovery in databases (KDD), provided that the latter contain meta-linguistic information, our theory of the category of aspect in Polish can be seen as the first attempt of applying computational approximation-based methods in order to determine the relevance and relative importance of the semantic parameters used to model aspect. Only such detailed work with databases may be supposed to offer formal, experimentally tested and comparable cross-language definitions of semantic categories.

Our ontology-based semantic approach is appropriate for contrastive studies: the complete tree of ontological features used in different languages (language specific and universal features) can serve as intermediary comparison language (*tertium comparationis*). The interactive work with SEMANA consists in describing each language independently of the other(s) and exploring original text corpuses (not translations). For instance, we propose to describe aspect uses in a language L1 (collecting a database 1), obtain types (usages) defined by their feature structure (partial tree), compare these types with those obtained independently in language L2 (in database 2): language specific semantic feature structures are partial trees of the general complete ontological tree of aspect features. Computer-aided translation methods would thus consist in bringing together expressions from L1 and L2 with identical or similar feature structures. Approximation methods (RST and FCA) make it possible to compare not only totally *identical* but also *similar* feature structures.

## Bibliography

- Karolak St. (1997)** “Arguments contre la distinction: aspect / modalité d’action (Aktionsart)” in *Etudes cognitives / Studia kognitywne*, Vol.2, SOW, Warszawa, 175-192.
- Laskowski R. (1998)** “Uwagi o znaczeniu czasowników” in *Gramatyka współczesnego języka polskiego, Morfologia*, wyd. 2 zmienione, T. 1, P.W.N., Warszawa. p. 152-171.
- Paduceva E. V. (1996)** *Semanticeskije issledowanija*, Moskva.
- Piernikarski C. (1969)** *Typy opozycji aspektowych czasownika polskiego na tle słowiańskim*, Ossolineum, Wrocław.
- Sauvet G. (2008)** “Symbolic and statistical Analyses of meta-data using the “Semana” platform — a bundle of tools for the KDD research”, CASK SORBONNE 2008 (Language Data Mining) International conference, June, 13th-14th, 2008, Université Paris-Sorbonne – Paris 4 <http://celta.paris-sorbonne.fr/anasem/papers/>
- Sémon Jean-Paul (1986)** «Postojat’ ou la perfectivité de congruence, définition et valeurs textuelles» *Revue des Etudes Slaves*, T. 58/4, Institut d’Etudes Slaves, Paris.
- SFPW: Słownik frekwencyjny polszczyzny współczesnej** (1990), red. Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J., PAN, Instytut Języka Polskiego, Kraków.
- Vendler Z. (1967)** « Verbs and Times », in Z. Vendler, *Linguistics and Philosophy*, Ithaca, New York: Cornell University Press., pp. 97-121 (revised version of Vendler Z. « Verbs and Times », *The Philosophical Review*, 66 (1957), 143-160)
- Włodarczyk, A. (2003)** « Les cadres des situations sémantiques », *Etudes cognitives / Studia kognitywne* V, SOW, Polish Academy of Sciences, Warszawa.

**Włodarczyk, A. & Włodarczyk, H. (2001)** « La Préfixation verbale en polonais I. Le statut grammatical des préfixes, II. L'Aspect perfectif comme hyper-catégorie », *Etudes cognitives / Studia kognitywne* IV, SOW, Warszawa p. 93-120.

— **(2003)** « Les paramètres aspectuels des situations sémantiques », *Etudes cognitives / Studia kognitywne* V, SOW, Warszawa, p. 11-34.

— **(2006)** « Semantic Structures of Aspect (A Cognitive Approach) », *Od fonemu do tekstu, prace dedykowane Profesorowi Romanowi Laskowskiemu*, Instytut Języka Polskiego Polskiej Akademii Nauk, Lexis, Kraków, p. 389-408.

— **(2008)** « The Pragmatic validation of Utterances », in *Etudes cognitives / Studia kognitywne* VIII, SOW, Warszawa, 117-128.

**Włodarczyk, H. (1997)** *L'Aspect verbal dans le contexte en polonais et en russe*, Institut d'Etudes Slaves, Paris, 240 p.

— **(1998)** « Wykładniki wartości informacyjnej wypowiedzenia w j. polskim i francuskim (aspekt, okreslonosc, modalnosc) », Congrès des Slavistes Cracovie 1998, *Revue des Études Slaves* T. LXX/1, p. 53-66, Paris.

— **(2003)** «L'Aspect perfectif comme hypercatégorie (approche cognitive) », communication au XIIIe congrès des slavistes à Ljubljana en août 2003, *Revue des Études Slaves* LXXIV/2-3, p.327-338 Paris.

Xrakovskij V. S. (1997) “Mul'tiplikativy i semel'faktivy (problema vidovoj pary)”, *Semantika i struktura slavjanskogo vida*, red. S. Karolak, Wyd. Naukowe, Kraków.

Part 2  
**Problems of Semantics and their Representaion  
in Slavic Digital Lexicography**



# Representing Semantics in the Digital Combinatorial Dictionaries of the ETAP-3 System: New Developments

Leonid L. Iomdin

A.A.Kharkevich Institute for Information Transmission Problems,  
Russian Academy of Sciences, Moscow, Russia  
iomdin@iitp.ru

## 1 Introduction

The ETAP-3 multipurpose linguistic processor under development in the Laboratory of Computational Linguistics of the Kharkevich Institute in Moscow (see e.g. [1]) makes use of a number of large digital dictionaries that operate in all options of the processor, including machine translation, the system of synonymous and quasi-synonymous paraphrasing, the tagging tool for the creation of the syntactically annotated text corpus, the UNL enconverting and deconverting system, and a few others.

The four largest dictionaries of ETAP-3 are 1) the 130,000-strong morphological dictionary of Russian, 2) the morphological dictionary of English, counting 100,000 entries; 3) the 100,000-strong combinatorial dictionary of Russian and 4) the combinatorial dictionary of English, which has approximately the same number of entries.

We will consider the issues of semantic representation in the two latter dictionaries – the combinatorial dictionaries of Russian and English. Other ETAP-3 dictionaries, used for prototype MT systems working with several different languages (French, Spanish, German, Korean, and Arabic), are much smaller. There is also, within the ETAP-3 environment, a rather large UNL dictionary embracing almost 80,000 entries. Even though this dictionary, being part of the UNL project, is specifically designed to represent the semantics of the natural language ([5], [4]), its bulk is created semi-automatically from external resources (primarily the WordNet, see [3]), and another substantial part inherited many of the features of the English combinatorial dictionary, including those related to semantics. It is therefore secondary in nature, so we will disregard it in the discussion that follows. I

## 2 Combinatorial Dictionaries: General Layout

The ETAP-3 combinatorial dictionaries (CD) of Russian and English inherit their name and structure from the explanatory combinatorial dictionary (ECD) of Meaning  $\Leftrightarrow$ Text linguistic theory (see e.g. [8], [7]); the main difference being that CD dictionaries, unlike ECD, do not contain lexicographic definitions of words (which is why the attribute “explanatory” is omitted). Otherwise, the structures of CDs and ECDs are similar, with the natural exception that ECD is oriented at human readers whilst CD dictionaries are designed for computer applications.

An important characteristic feature of the ETAP-3 processor is high reusability of its linguistic resources. The dictionaries are no exception; in particular, the Russian CD is used by the Russian parser as the source dictionary in the Russian-to-English translation and by the Russian generator as the target dictionary in the opposite direction of translation, while for the English CD the reverse is true.

The general layout of a CD entry is as follows. Every entry is divided into several zones. The first, universal, zone contains linguistic data relating to the source language; while all the remaining zones present information referring to specific options: machine translation into a specific target language, paraphrasing etc.

The universal zone is further subdivided into several fields: 1) entry name field; 2) part of speech<sup>1</sup>; 3) syntactic features; 4) semantic features, or descriptors; 5) government pattern, or subcategorization frame, and 6) lexical functions. Additionally, an entry may have one or more operational fields where specific rules referring to particular stages of text analysis, or references to such rules, are given. Optional fields of discriminative comments can be used to distinguish between word senses of polysemous lexemes, or homonyms.

Fig. 1 and 2 below show the universal zones for the English CD entry *accusation* and the Russian CD entry *обвинение 1* ‘accusation’, respectively.

```

1 ACCUSATION
2 POR:S
3 SYNT:VOC,COUNT
4 DES:‘ФАКТ’,‘ДЕЙСТВИЕ’,‘АБСТРАКТ’
5 D1.1:BY1
6 D1.2:OF,‘ЛИЦО’
7 D2.1:AGAINST
8 D3.1:OF,‘ФАКТ’
9 _VO:ACCUSE
10 _SYN1:CHARGE3
11 _S1:ACCUSER
12 _ANTI:JUSTIFICATION
13 _MAGN:GRAVE3
14 _VER:JUST2/WELL-BASED
15 _ANTIVER:FALSE/GROUNDLESS/UNFOUNDED/BASELESS/UNJUST
16 _OPER1:MAKE1/BRING
17 _FINOPER1:DROP2
18 _REAL1-M:PROVE/SUBSTANTIATE
19 _OPER2:BE<UNDER1>
20 _REAL2-M:DENY/REFUTE/REPUDIATE
21 _ANTIREAL2-M:ADMIT
22 _CAUSFUNC1:LAY1/LEVEL2

```

Fig. 1. Universal zone of an English CD entry

In Fig.1, line 1 is the entry name; line 2 shows the part of speech of the word (S corresponds to the noun); line 3 offers the list of syntactic features (COUNT denotes a countable noun and VOC means that if the word is preceded by the indefinite article its form should be *an* rather than *a*); line 4 lists the semantic features ‘FACT’, ‘ACTION’ and “ABSTRACT THING”, lines 5 to 8 present the government pattern, which consists of three valency slots (AGENT, as in *accusation by the court* or *accusation of the court*, PATIENT, as in *accusation against the alleged robber*, and THEME, as in *accusation of embezzlement*).

Lines 9 to 22 list the values of lexical functions (LF) for which the headword *accusation* is the argument; e.g. the LF Magn conveys the meaning of high degree of an accusation (*grave accusation*), the LF Oper<sub>1</sub> represents the verb denoting what the AGENT (accuser) does with the accusation: he *makes* or *brings an accusation*, the LF Oper<sub>2</sub> represents the verb denoting what happens to the PATIENT (defendant) in the situation of an accusation: he *is under* it; the LF Real<sub>1</sub> denotes what the AGENT does in order to succeed in his accusation: he *proves* or *substantiates* it; the LF Real<sub>2</sub> denotes what the PATIENT (defendant) does to succeed when an accusation is brought against him: he *denies*, *refutes*, or *repudiates* it, and the LF AntiReal<sub>2</sub> denotes what the PATIENT does if he decides he will not strive to succeed when an accusation is brought against him: he *admits* it.

<sup>1</sup> Other morphological information on the word is contained in the morphological dictionary, whose entries are directly linked to CD entries.

```

1  ОБВИНЕНИЕ1
2  КОММЕНТ:“ВЫСКАЗЫВАНИЕ МНЕНИЯ О ЧЬЕЙ-ЛИБО ВИНЕ”
3  ЕХАМПЛЕ:“ОБВИНЕНИЕ В ХАЛАТНОСТИ”
4  POR:S
5  SYNT:СРЕДН,ИСЧИСЛ
6  DES: ‘ДЕЙСТВИЕ’, ‘ФАКТ’, ‘АБСТРАКТ’
7  D1.1:ТВОР, ‘ЛИЦО’
8  D1.2:РОД, ‘ЛИЦО’
9  D2.1:РОД
10 D2.2:ПРОТИВ1
11 D3.1:В2
12 _SYN1:ОБЛИЧЕНИЕ
13 _ANTI:ОПРАВДАНИЕ2
14 _VO:ОБВИНЯТЬ
15 _S1:ОБВИНИТЕЛЬ
16 _S2:ОБВИНЯЕМЫЙ
17 _MAGN:СУРОВЫЙ/ТЯЖКИЙ
18 _VER:ОБОСНОВАННЫЙ/ПРАВИЛЬНЫЙ/СПРАВЕДЛИВЫЙ
19 _ANTIVER:ЛОЖНЫЙ/НЕОБОСНОВАННЫЙ/НАПРАСНЫЙ/ПУСТОЙ
20 _OPER1:ВЫДВИГАТЬ/ПРЕДЪЯВЛЯТЬ/БРОСАТЬ2
21 _SO_INСЕРOPER1:ВЫДВИЖЕНИЕ
22 _INСЕРOPER1:ВЫДВИГАТЬ
23 _FINOPER1:ОТКАЗЫВАТЬСЯ<ОТ>/СНИМАТЬ1
24 _SO_FINOPER1:ОТКАЗ1<ОТ>/СНЯТИЕ
25 _SO_OPER1:ПРЕДЪЯВЛЕНИЕ
26 _OPER2:ПОДВЕРГАТЬСЯ
27 _REAL1-М:ДОКАЗЫВАТЬ
28 _REAL2-М:ОТВЕРГАТЬ/ОТМЕТАТЬ/ОТКЛОНЯТЬ
29 _ANTIREAL2-М:СОГЛАШАТЬСЯ2<СЭ>/ПРИЗНАВАТЬ
30 TRAF:АГЕНТ.10
31 TRAF:1-КОМПЛ.20

```

Fig. 2. Universal zone of a Russian CD entry

In Fig. 2, line 1 is the entry name; lines 2 and 3 are discriminative comments and examples (‘offering an opinion stating someone’s guilt’, as in *accusation of neglect of duty*) that help to distinguish between the word sense *обвинение 1* ‘accusation’ and other senses, such as *обвинение 2* ‘prosecution’; line 4 shows the part of speech; line 5 offers the list of syntactic features, line 6 lists the semantic features (they are the same as in the word’s English equivalent entry); lines 7 to 11 present the government pattern, which consists of the same valency slots as its English counterpart even though they are instantiated differently; lines 12 to 29 list the LF values; and lines 30 and 31 are references to syntactic rules that describe syntactic structures in which the word *обвинение 1* may appear.

### 3 Semantic Issues

It is easily seen that almost all fields of the dictionary are related to semantics, to a greater or lesser degree. The fields of descriptors and lexical functions are intrinsically semantic. However, many syntactic features have a semantic origin, too: in the sample entries in Fig. 1 and 2, the simple **count** and **исчисл** features, stating the countability of the English and Russian nouns, are semantically motivated even though they are primarily used by the syntax, ensuring e.g. the correct choice of the grammatical number of the noun, the article, or determiners (*an accusation* vs. *information*, *few accusations* vs. *little information*, etc).



Other examples of semantically motivated syntactic features are Russian syntactic features **мес** and **прини**.

The first feature, **мес**, is ascribed to the words denoting months of the year but exclusively used in syntactic rules as the word having this feature constitute the construction known as the genitive of the date: in *Он приехал первого февраля* ‘He came on the first of February’ the numeral adjective *первого* must be in the genitive case. Other types of modifiers of date never appear in the genitive case: *\*приехал вечера* lit. ‘came of the evening’, *\*приехал понедельника* lit. ‘came of Monday’ are all wrong.

The second feature, **прини**, is ascribed to many nouns that denote units of measurements and objects whose small number testifies to exceptional situations: *копейка* ‘kopeck’, *капля* ‘drop’, *крошка* ‘crumb’; *миллиметр* ‘millimeter’, *облачко* ‘cloud’ etc. This feature manifests itself in constructions with the negative particle *ни*  $\approx$  ‘not one, not a single’, as in *не дам ни копейки* ‘I won’t give a kopeck’, *нет ни капли воды* ‘there is not a single drop of water’, *на небе не было ни облачка* ‘there was not one cloud in the sky’. Other words with close but not identical semantics are not allowed in such constructions: *\*Я не купил ни стола* ‘I didn’t buy one table’, *?На небе не было ни звезды* ‘there was not one star in the sky’: to make these constructions valid, one needs to explicitly add the word *один* ‘one’ or *единый* ‘single’: *Я не купил ни одного стола*, *На небе не было ни единой звезды*.

Government patterns also resort to a substantial share of semantic data; in the sample entries above, selectional restrictions on the instantiation of valency slots are imposed by specifying the semantic features of words which may fill these slots. For example, the agent valency of the word *accusation* may be filled by a word that has the semantic feature ‘PERSON’, and its theme valency by a word with the semantic feature ‘FACT’.

A recent major innovation in the combinatorial dictionary of ETAP-3 permits the developers to use the operator of negation when imposing restrictions on valency instantiations: we can now easily state e.g. that a valency slot may NOT be filled by a word that has certain features. This is especially important when lists of semantic features ascribed to individual words are rather loose. For instance, the requirements imposed on the set of semantic features ascribed to words in ETAP-3 include the provision that any human is a physical object, which means that whenever a word is supplied with the feature ‘HUMAN’, the feature ‘PHYSICAL OBJECT’ is automatically added. This may lead to errors in valency slot instantiations: say, the entry for the Russian verb *стирать* ‘wash (clothes etc.)’ is supplied with a restriction on its instrumental valency: it is required that this slot may only be filled by a word in the instrumental case having the features ‘PHYSICAL OBJECT’ or ‘SUBSTANCE’: *стирать хозяйственным мылом <порошком>* ‘to wash with laundry soap <washing powder>’. Formally, a human is a physical object; so, to avoid instrumental interpretations of expressions like *стирается прачкой* (‘is washed by the laundry woman’ but not ‘is washed with a laundry woman’ we may now use the formula like)

D3.1: ТВОР, ‘PHYSICAL OBJECT’, NOT ‘HUMAN’.

## 4 Case Study

It follows from the above that semantic data are crucial for the creation, update and proper use of the combinatorial dictionary, and they have to be tackled with utmost attention and maximum precision. This is of special importance when the developer makes decisions on the number and choice of lexical entries that correspond to word senses of a particular polysemic word. I would like to illustrate this type of activity with the following example.

Russian explanatory dictionaries of conventional type (see e.g. [10] or [6]) represent the noun *история* as having at least seven word senses, further divided into subsenses, which can be briefly summarized as follows: 1) reality in the course of development, as in *законы истории* ‘laws of history’; 2a) consequent course of development or change, or course of events, as in *история нашего города* ‘history of our city’ or *история болезни* ‘medical history’; 2b) course of events, as in *история моей жизни* ‘the history of my life’; 3) the old times, the past, as in *история и современность* ‘history and contemporaneity’; 4a) the science that studies the past of human

society, as in *история средних веков* ‘the history of the Middle Ages’; 4b) the subject studied at school, as in *учитель истории* ‘a teacher of history’; 5) the science that studies the development of nature, culture, or knowledge, as in *история живописи* ‘the history of painting’; 6) a narration, as in *смешная история* ‘a funny story’; 7) an event or an accident, as in *со мной случилась странная история* ‘a strange story happened to me’.

It goes without saying that no automatic parser or other NLP system can afford to have a dictionary in which a common-type word has so many senses, for the obvious reason that such a system will never be able to distinguish between these senses and thus resolve the lexical ambiguity. Actually, more often than not, many of the senses defined in a way similar to that illustrated above are hardly distinguishable even by humans: if one hears, for instance, a sentence like

(1) *Школьники изучают историю российского государства* ‘Schoolchildren study the history of the Russian State’,

will he be able to confidently choose the adequate interpretation of the word *историю* from among the set of senses 2a), 2b), 4a), 4b) and 5) above? More generally: if not, which I think is the right answer, will it ever occur to a human that sentence (1) may be ambiguous due to the lexical ambiguity of the noun *история*? Again, I believe that the answer is not.

If anyone hears another sentence like

(2) *Друг рассказал мне грустную историю о том, как он не сдал ерекзамен* ‘My friend told me the sad story about how he failed in the exam’,

will he be able to see the lexically ambiguity of *историю* and choose unequivocally between the senses 6) and 7)? Hardly so.

A natural conclusion suggests itself that the list of senses offered for the word *история* by the conventional dictionary is too vague and very redundant, and it would be wrong to adopt this list in an application-oriented computer dictionary. Our lexicographic experience shows that such a situation is extremely typical: in lots of cases, the distribution of word senses in regular dictionaries is inappropriate for digital ones. This means that the authors of digital dictionaries need to develop adequate approaches to the selection of senses to be represented. We believe that a satisfactory trade-off approach to this issue can be developed from the principle of reasonable sufficiency: postulate as few word senses, or lexemes, as can be reliably distinguished in texts. Of course, this tradeoff is not always easy to achieve but in most cases it proves to be adequate enough.

Let us look at the Russian noun *история* from this point of view. When we first included this word into the CD, we decided that one sense would be enough: for the practical purposes of Russian-to-English machine translation, one single English equivalent *history* seemed to be adequate in most cases, even though it was clear that it should be replaced by another equivalent, *story*, in certain situations (for which a couple of *ad hoc* rules were written). It became clear, however, that this solution is not the best one. Differences in the behaviour of this word in certain texts became too obvious to be ignored. To name but a few, 1) certain contexts easily allowed the use of plural, as in *Он рассказал нам две страшные истории* ‘he told us two horrible stories’ while in other contexts the plural number was virtually impossible (*\*истории древнерусского зодчества* ‘histories of ancient Russian architecture’); 2) the government patterns proved to be different *история его семьи* ‘the history of his family’ but *история о его семье* ‘a story about his family’; 3) the lists of lexical functions turned out to be different, too: *заниматься историей Египта* ‘to engage in the history of Egypt’ vs. *рассказать историю про Египет* ‘to tell a story about Egypt’.

Accordingly, the decision was taken to have two separate entries in the Russian CD for the word *история*, whose universal zones (with minor abridgements) are presented in Fig. 3 and 4 below:

```
ИСТОРИЯ1
КОММЕНТ:“НАУКА; ПОСЛЕДОВАТЕЛЬНОСТЬ СОБЫТИЙ (HISTORY)”
ПРИМЕР:“ ИСТОРИЯ ДРЕВНЕГО МИРА, ИСТОРИЯ БОРЬБЫ КЛАССОВ”
РОД: S
СИНТ: ЖЕНСК, ИСЧИСЛ
```

```

DES: 'ДЕЯТЕЛЬНОСТЬ', 'НАУКА', 'ФАКТ', 'АБСТРАКТ', 'ПРЕДМЕТ'
D1.1: РОД, 'ЛИЦО'
D2.1: РОД
_GENER: НАУКА
_S1: ИСТОРИК
_AO: ИСТОРИЧЕСКИЙ
_OPER1: ЗАНИМАТЬСЯ1
_SO_OPER1: ЗАНЯТИЕ1

```

Fig. 3. Universal zone of the Russian CD entry for *история 1* 'history'

```

ИСТОРИЯ2
КОММЕНТ: "ПОВЕСТВОВАНИЕ; СЛУЧАЙ"
EXAMPLE: "СО МНОЙ ПРИКЛЮЧИЛАСЬ НЕПРИЯТНАЯ ИСТОРИЯ"
FOR: S
SYNT: ЖЕНСК, ИСЧИСЛ, НУМЕР
DES: 'ИНФОРМАЦИЯ', 'ФАКТ', 'ПРЕДМЕТ', 'АБСТРАКТ'
D1.1: РОД, 'ЛИЦО'
D2.1: 02
D2.2: ПРО1
D2.3: С3
_FUNCO: СЛУЧАТЬСЯ1/ПРОИСХОДИТЬ1
_OPER1: РАССКАЗЫВАТЬ

```

Fig. 4. Universal zone of the Russian CD entry for *история 2* 'story'

If we now compare the resulting CD entries with the list of senses from the conventional dictionary, we will easily see that *история 1* corresponds to the first five groups of senses of this dictionary, while *история 2* took in the remaining two senses.

A series of experiments with the deeply annotated text corpus of Russian, SynTagRus (see e.g. [2], [9]) showed that the parser was able to correctly choose between the two senses of *история* in the overwhelming majority of cases.

We firmly believe that the principle of reasonable sufficiency should be used in most lexicographic tasks involving multilanguage digital dictionaries.

## Bibliography

- [1] Apresian, J., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., Tsinman, L. (2003). Etap-3 linguistic processor: a full-fledged nlp implementation of the mtt. In *MTT 2003, First International Conference on Meaning - Text Theory. Paris, École Normale Supérieure, Paris, June 16-18 2003*, pages 279–288, Paris.
- [2] Apresjan, J., Boguslavsky, I., Iomdin, L., Iomdin, B., Sannikov, A., Sizov, V. (2006). A syntactically and semantically tagged corpus of russian: State of the art and prospects. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 1378–1381, Genoa.
- [3] Bekios, J., Boguslavsky, I., Cardeñosa, J., Gallardo, C. (2007). Using wordnet for building an interlingual dictionary. In *Proceedings of the Fifth International Conference "Information Research and Applications" i.TECH 2007. ISSN: 1313-1109 Varna, Bulgaria. V.1.*, pages 39–45, Sofia. Ithea.
- [4] Богуславский, И. М., Диконов, В. Г. (2009). Универсальный словарь компонентов. In *Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» Диалог'2009 (Бекасово, 27–31 мая 2009 г.) Вып. 8(15)*. ISBN 978-5-7281-1102-3, Москва. РГГУ.

- [5] Boguslavsky, I., Dikonov, V. (2008). Universal dictionary of concepts. In *MONDILEX First Open Workshop «Lexicographic Tools and Techniques»*, pages 31–41, Moscow.
- [6] Кузнецов, А. С. (2008). *Современный толковый словарь русского языка*. Москва.
- [7] Мельчук, И. А., Жолковский, А. К. (1984). *Толково-комбинаторный словарь современного русского языка*. Вена.
- [8] Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Lessard, A. (1984). Dictionnaire explicatif et combinatoire du français contemporain. *Recherches lexico-sémantiques*, I:183–193.
- [9] Nivre, J., Boguslavsky, I., Iomdin, L. (2008). Parsing the syntagrus treebank of russian. In *Coling 2008. 22nd International Conference on Computational Linguistics. Proceedings of the Conference. ISBN: 978-1-905593-47-7. Vol. 2.*, pages 641–648.
- [10] Ожегов, С. И. (1991). *Словарь русского языка. 23-е изд., испр.* Москва.

# Bulgarian-Polish Online Dictionary

## — Design and Development\*

Ludmila Dimitrova<sup>1</sup>, Violetta Koseska<sup>2</sup>, Ralitsa Dutsova<sup>3</sup>, Rumiana Panova<sup>3</sup>

<sup>1</sup> Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

<sup>2</sup> Institute of Slavic Studies, Polish Academy of Sciences, Poland

<sup>3</sup> Veliko Tărnovo University, Bulgaria

**Abstract.** In the present paper we describe problems of design and development of an experimental Bulgarian–Polish online dictionary. We focus our attention on the ongoing version of the web-based application representing the dictionary. The basis for the first Bulgarian-Polish experimental online dictionary is the ongoing version of the Bulgarian-Polish electronic dictionary, currently developed in MS WORD-format in the framework of the joint research project “Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary” between IMI-BAS and ISS-PAS. The current version consists of approximately twenty thousand dictionary entries. Some examples of dictionary entries are presented.

**Key words:** bilingual digital dictionaries, entry classifiers, lexical database, web-based application, online dictionary

## 1 Introduction

Computer lexicography encompasses computer methods and resources for the automation of lexicographic activity. Such activities include: setting up of basic principles for development, creation and maintenance of dictionaries, recording of linguistic information in databases, creation of electronic indices, etc.

A dictionary (commonly) is a list of words and their meanings arranged in alphabetical order. In it, information is given about the pronunciation, grammar, derivative words, history or etymology of the main word, as well as recommendations for its usage, examples, phraseological expressions, illustrations. Dictionaries are most commonly available as books, but lately ones in electronic form are also gaining recognition.

Dictionary classification is based on multiple criteria. Various classifications exist in different lexicographic and lexicological works.

In the scope of this paper, two classifications are of particular interest:

*According to the type of carrier:*

- traditional dictionaries — these are developed with a human/computer, but in their final form, they reach the user in paper form;
- electronic dictionaries — these exist in digital format and can be divided into two categories: online (web-based) and local (desktop) dictionaries.

*According to the number of languages:*

- monolingual dictionaries — dictionaries in which words are defined in the same language;
- bilingual dictionaries — dictionaries which contain a translation of the words in exactly two languages;
- multilingual dictionaries — dictionaries which contain a translation of the words in more than two languages.

---

\* The study and preparation of these results have been supported by the EC’s Seventh Framework Programme [FP7/2007–2013] under the grant agreement 211938 MONDILEX.

### 1.1 Why choose a Bulgarian-Polish Dictionary?

Our reasons are as follows:

- In the past 20–25 years neither Bulgarian-Polish nor Polish-Bulgarian dictionaries have been published.  
Existing dictionaries from Bulgarian to Polish or vice versa are a collector’s rarity and are outdated. The first Polish-Bulgarian dictionary, prepared by prof. I. Lekov had been exhausted long before the second one appeared in 1961 — by authors I. Lekov and F. Sławski. The next Polish-Bulgarian dictionary was only published in 1988; its author is S. Radewa. The first Bulgarian-Polish dictionary (PODREČZNY SŁOWNIK BULGARSKO-POLSKI), prepared by F. Sławski, was published in 1963 in Warsaw. The second (and last) edition of this dictionary dates from 1987.
- These dictionaries contain about 50–60 thousand words, but they have a significant number of disadvantages (from a contemporary point of view). They contain words that are no longer in use — mostly dialect words and words of Turkish language origin. Moreover, these dictionaries contain a lot of words that were created by the authors themselves (in Bulgarian, it is possible to create new words through the addition of suffixes to certain words, mostly verbs), which are practically unusable.
- There is no existing Bulgarian-Polish online dictionary. So far the communication between languages of the same language group (Slavic) is channelled via other languages such as English or German.

### 1.2 Why an online dictionary?

Advantages of online dictionaries over the local (desktop) electronic ones:

- Wide accessibility
- A possibility for a continuous update and editing
- Opportunities for search and hypertext cross-references
- Real-time use by a large number of users
- Real-time editing and update of the dictionary
- No restrictions on volume
- Economizing computer resources — local installation of the dictionary is not required.

## 2 Problems of the computer realization

The basis for the first Bulgarian-Polish experimental online dictionary is the ongoing version of the Bulgarian-Polish local electronic dictionary [1], [3]. The Bulgarian-Polish electronic dictionary is currently developed in MS WORD-format in the framework of the joint research project “Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary” between IMI-BAS and ISS-PAS under the supervision of L. Dimitrova and V. Koseska. The current version consists of approximately twenty thousand dictionary entries.

The main problems related to the computer realization of dictionaries arise from the fact that they are **simultaneously treated as text and databases**. They obviously look like text and have common points with other types of text. However, users do not normally read dictionaries, from A to Z, as they do with the majority of texts, but rather use them to obtain specific information through a given key (in this case a headword). The information associated with this key can include: pronunciation, grammar information, definitions, etymology, etc. Electronic dictionaries are capable of fulfilling users’ requests multiple times faster than paper dictionaries, as well as providing the possibility to return all entries whose title words contain the user-defined criteria. Despite the fact that dictionary entries resemble a text on the screen, the internal representation of electronic dictionaries is actually done as a database.

Dictionaries are among the **most complex text types** because of the high level of structuring and information content. Every dictionary entry is a structured object which uses different abbreviations and structural units in order to present the whole information succinctly. Furthermore, the structure of dictionary entries varies greatly within the dictionary itself as well as between different dictionaries. In spite of these variations some strict and constant structural rules exist so that the dictionaries can be understood by their readers.

**Another kind of problems is related to the part of speech (POS) classifications** of the headwords in bilingual digital dictionaries. As we already pointed out, [3], [2], the choice of *POS classifiers* of the headwords in the dictionary entries is very important. The development of a system of multilingual dictionaries on the basis of bilingual ones requires at first a unification of the classifiers in the dictionary entries. The problem turns to the harmonisation of the classifiers for various languages, and its solution has to present a unified selection of classifiers and a standard form of their presentation. The comparison of the Bulgarian and Polish material requires an explanation, which is important for the part of speech classifiers in the dictionary entries of the cited bilingual electronic dictionary. In the current paper we will analyze in short the specifications of verbal forms in both languages.

The POS classification varies across different languages. Often there is more than one possible POS classification for a given language. Next we will briefly review the POS classification of the *participle* (one of the important verbal forms) in the two languages, in comparison to another POS, the *adjective*.

#### **Functions of the participle**

The classification of a participle, not only as a verb form, is an important problem: the role of the participle varies significantly across languages, because its properties and functions are different. In the Slavic languages the forms of the participles are inflected, in contrast to English, for instance, where the participle are invariant. In the Slavic languages the forms of the participles contain information about the aspect and tense of the verbal form. The information about the aspect of the verb is important for the Slavic languages, but does not exist, for instance, in English. Bulgarian and Polish distinguish between the following functions of the *participle* form: predicative function, attributive function and adverbial (or semipredicative) function.

#### **Participles and verbs**

It is important to emphasize that participles preserve some properties of the main form of the verb, such as voice, tense and aspect. In Bulgarian and in Polish there are active and passive participles.

The properties presented above serve as a proof that participles deserve a separate treatment, different from that of adjectives.

#### **Features of the adjective**

Adjectives in Polish can be declined for gender, number and case, while in Bulgarian only for gender and number. The adjective does not express a temporal or aspect relation on its own, unlike the participle.

The main grammatical meaning of an adjective is its attributive meaning. Unlike a participle, which is closely related to the verb tense (with states or events in the past, present and future), an adjective describes properties or qualities of an object, like:

малко дете // małe dziecko // *a little child* //

Adjectives can have not only an attributive but also a predicative function (only as adjective clause). Adjective clauses perform the same function in sentences that adjectives do: they modify nouns. In this function, however, adjective clauses are just a nominal part of the subject, i.e. they do not express independently neither a temporal nor an aspect relation:

Mały dom // Малка къща // *A small house* //

Dom jest mały. // Къщата е малка. // *The house is small.* //

These arguments show that participles must be classified as a separate POS, and not be classified as adjectives. The unification of adjectives and participles is probably possible in languages that do not have verbal forms, or in which the system for description of the aspect and tense of the verbal form is simpler than that of the Slavic languages.

### 3 Design of Structure and Content of the Entry

**Model for dictionary encoding:** we choose the dictionary encoding model CONCEDE, developed in the framework of the EC project CONCEDE (Consortium for Central European Dictionary Encoding)<sup>4</sup>. CONCEDE is a model developed in accordance with the TEI standards (Text Encoding Initiative: [7]) and offers a standardized, understandable and intuitive structure and semantics of a dictionary article. The CONCEDE model was used to develop lexical databases in six European languages — Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovenian. The first lexical database in Bulgarian was created in the framework of this project and contains more than 2700 lexical units from source “Bulgarian Explanatory Dictionary”, for more details see [5].

The build-up of electronic dictionaries is a complex and strenuous process, associated with overcoming various difficulties: (1) Lack of a sufficient number of formal models that allow words to be divided into formal language classes and a given word to be automatically included in one or another class. The creation of electronic dictionaries can be done through manual input of the dictionary articles — a process through which paper dictionaries are input into a computer (also possible with a scanner) or new dictionaries are prepared for print (they get printed after being typed up). These dictionaries, known as “machine-readable dictionaries” are different from their paper counterparts mostly in that they exist on magnetic carriers as files and can be processed as files. They follow a certain order and the articles have a concrete structure. As they are meant to be used by a human, their disadvantage from a computer point-of-view is that they are not sufficiently formalized (formal structures are missing from their descriptions) and the extraction of knowledge from them requires the development of special computer modules. (2) A great variety of structures and content, which presupposes a conflict between universality and detail. The conflict between universality and detail is particularly strong for dictionaries due to the large diversity in structures and content, which turns the creation of a standard for dictionary encoding into a major challenge, [8]. With the scope of overcoming this conflict the TEI workgroup created a universal standard for coding different types of dictionaries which encompasses fundamental principles of high degree of structure and diversity of dictionary entries [7].

A dictionary entry — in terms of structure and content — is a complex unit. The structure of dictionary entries varies a lot within the dictionary as well as between separate dictionaries. The external structure (presentation of text) does not completely determine the internal structure (information content in the database).

There is a great diversity of hierarchical structures: in some dictionary entries the hierarchy organization of their structure may be deeply embedded (i.e. it allows many levels), whereas in other cases some structural elements from this hierarchy may be missing.

This makes the database supporting the dictionary logically complex and difficult to create.

### 4 Lexical Database of the Bulgarian-Polish Dictionary

The structure and content tags of the designed structural unit should fully meet international standards so that the LDB [4] and the electronic dictionaries be compatible with language resources created in other projects and for other languages.

#### Structure of a dictionary entry:

- Headword
- Formal Features — phonetics, grammar, morphology, syntax, etymology, style
- Semantic information
- Quotations
- Additional information:
  1. Derivatives

<sup>4</sup> <http://www.itri.brighton.ac.uk/projects/concede/>



2. Phrases
3. Examples — phrasal and sentence usages, illustrations

The CONCEDE model [6] is used in the development of LDB for 6 European languages — Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The first LDB for Bulgarian was created in the above framework and contains more than 2700 lexical units from “Български тълковен речник” [5].

The *structural tags* are **alt**, **entry**, **struc**, and the *content tags* are **case**, **def**, **domain**, **eg**, **etym**, **gen**, **geo**, **gram**, **hw**, **itype**, **lang**, **m**, **mood**, **number**, **orth**, **person**, **pos**, **q**, **register**, **source**, **subc**, **time**, **tns**, **trans**, **usg**, **xr**.

*New content tags* for **Bulgarian verbs** were added: the **<conjugation>** tag (to represent the conjugation of verbs) and the **<type>** tag (for the type of conjugation).

*New information* for the **aspect** of verbs in the tag **<gram>** (for perfect aspect and progressive aspect) and for the **transitivity/intransitivity** of verbs in the tag **<subc>** was also added.

#### *Realization of homonyms:*

The meanings of homonyms are entered in the dictionary as different database records. On the word-entry page, there is a field where the user must specify a homonym index — a number which shows the order of the meanings. For the representation of the homonym it is necessary to fill in the value of the attribute **n(homonym index)** in the tag **<entry>**:

**<entry n="1">**

**<entry n="2">**

#### **Representation of an entry in the LDB**

We choose the following entries from the Bulgarian-Polish electronic dictionary:

**завъ'ршва|м, -ш** vi. kończyć, zakańczać; kończyć się

**завъ'рш|а, -иш** вр. v. **завъ'ршвам**

**завъ'ршен** part. adi. skończoney, zakończoney, wykończoney

#### **Representation of the dictionary entries in the LDB:**

**<entry>**

```

<hw>завъ'ршва|м</hw>
<pos>v</pos>
<gram>i</gram>
<subc>transitive</subc>
  <conjugation>
    <orth>-ш</orth>
    <type>III</type>
  </conjugation>
<struc type="Sense" n="1">
  <trans>kończyć</trans>
  <alt>
    <trans>zakańczać</trans>
  </alt>
</struc>

```

**</entry>**

**<entry>**

```

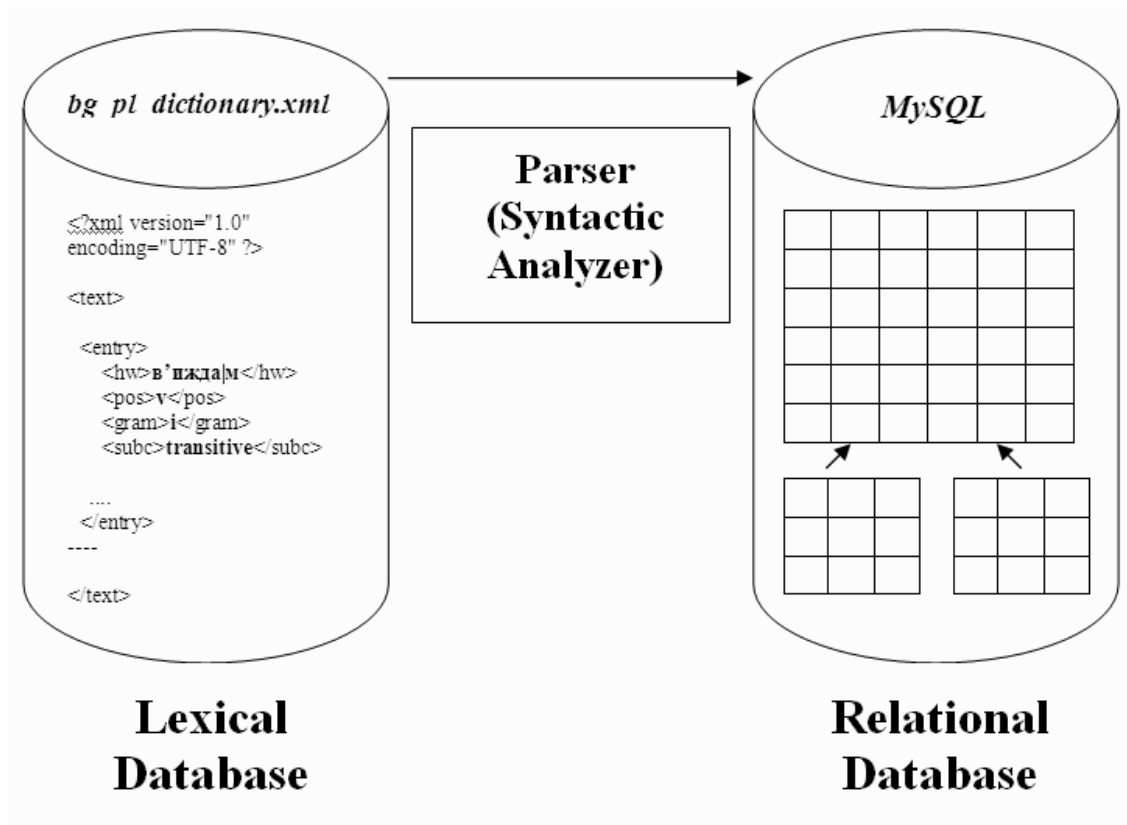
<hw>завъ'рш|а</hw>
<pos>v</pos>
<gram>p</gram>
  <subc>transitive</subc>
  <conjugation>
    <orth>-иш</orth>
    <type>II</type>
  </conjugation>

```

```

</entry>
  <xr>завъ'ршвам</xr>
</entry>
<entry>
  <hw>завъ'ршен</hw>
  <pos>part</pos>
  <alt>
    <pos>adi</pos>
  </alt>
  <struc type="Sense" n="1">
    <trans>skończony</trans>
  <alt>
    <trans>zakończony</trans>
  </alt>
  <alt>
    <trans>wykończony</trans>
  </alt>
</struc>
</entry>

```



Transformation of the Lexical Database to the Relational Database

Column / Word	завъ'рш а	завъ'ршва м	завъ'ршен
id	662	663	664
homonym_index			
bg_word	завъ#рш	завъ#ршва	завъ#ршен
suffix	а	м	
bg_word_search	завърша	завършвам	завършен
plural			

is_plural_rare			
conjugation	иш	ш	
conjugation_type	2	3	
has_gender			
gender_feminine			
gender_neuter			
id_explanation			
id_bg_word	582		
referent_bg_word	завъ#ршвам		

Table *bg\_word*

id	id_bg_word	pl_word	sense_index	alternative_sense_index	latin_translation	id_explanation
1110	663	kończyć	1	1		
1109	663	zakończyć	1	2		
1113	664	skończony	1	1		
1112	664	wykończony	1	2		
1111	664	zakończony	1	3		

Table *pl\_word*

id_bg_word	id_characteristic
662	18
662	57
663	17
663	57
664	5
664	44

Table *mm\_bg\_word\_characteristic*

id	abbreviation_bg	abbreviation_pl	description_bg	description_pl	description_lat	id_characteristic_type
5	прил	adi	прилагателно име			6
17	мин. нсв.	vi	глагол от несвършен вид			5
18	мин. св.	vp	глагол от свършен вид			5
44	прич	part	причастие			6
57	прех	transitive	преходен глагол			7

Table *characteristic*

id	name_bg	name_pl
5	Граматически категории за глаголи	

6	Граматически категории (части на речта)	
7	Граматически категории за преходност на глаголите	

*Table characteristic\_type*

## 5 Web-based Application for the representation of the Bulgarian-Polish online dictionary

The technologies used for the implementation of the web-based application are Apache, MySQL, PHP and JavaScript. We use free technologies originally designed for developing dynamic web pages with a lot of functionalities. With the help of HTML and CSS we created the designs of both administrative and end user modules. The following web based application is experimental, and the structure of the text fields is not permanently determined.

The current version of the Bulgaria-Polish online dictionary works optimally with Internet Explorer 6.0+ (Windows), and with Firefox 2.0.1+ (Windows, Linux). The website resolution is 1024/768 pixels.

Future expansion of the Bulgaria-Polish dictionary is a precondition for any changes in the database and web application.

### Main Modules:

The web-based application consists of two modules: an **administrator module** and an **end-user** module.

**The administrator module** — is intended for the person updating the dictionary, access only for authorized users. The administrator module is used to fill in the database and to offer user-friendly interface to the user who will be responsible for the word management.

**The end-user module** is bilingual, the user can choose the input language (Bulgarian or Polish) and according to their choice, a virtual Bulgarian or Polish keyboard is displayed.

### 5.1 Goals and tasks of the end-user module are:

- To present correct and up-to-date information to the user
- To be convenient and easy for searching and finding the meanings of words
- To allow an opportunity for translation from Polish to Bulgarian
- To allow the end-user to report missing words
- To create a user interface in both languages — Bulgarian and Polish (our idea is that both the end-user and administrative parts of the web-based application be bilingual).

The **end-user module** includes three sections.

Section “**Dictionary**”: this section is the home page — the first page of the web-based application. It contains a text field where the users fill in the word whose translation they are searching for. There are a Bulgarian and a Polish virtual keyboards built into this section.

“About the project” section: this section represents information concerning the project — authors, ideas, etc. Only the administrator could manage the text in this section. “Support” section that visualizes a form containing some text fields. Users can report missing words or inaccuracies in translation.

### 5.2 Goals and tasks of the Administrator Module

Goals and tasks:

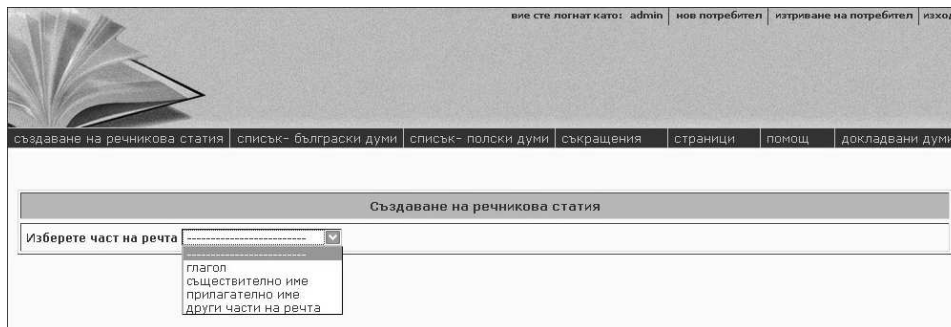
- To allow for enlargement of the number of words in the database (volume of the dictionary)
- The dictionary must be easy to use and must not require a programmer’s background from the administrator.

- The administrator must be able to receive, save and store missing words reported by the end-users, to insert new words in the database, to update all pages of the web-based application for the end-user.

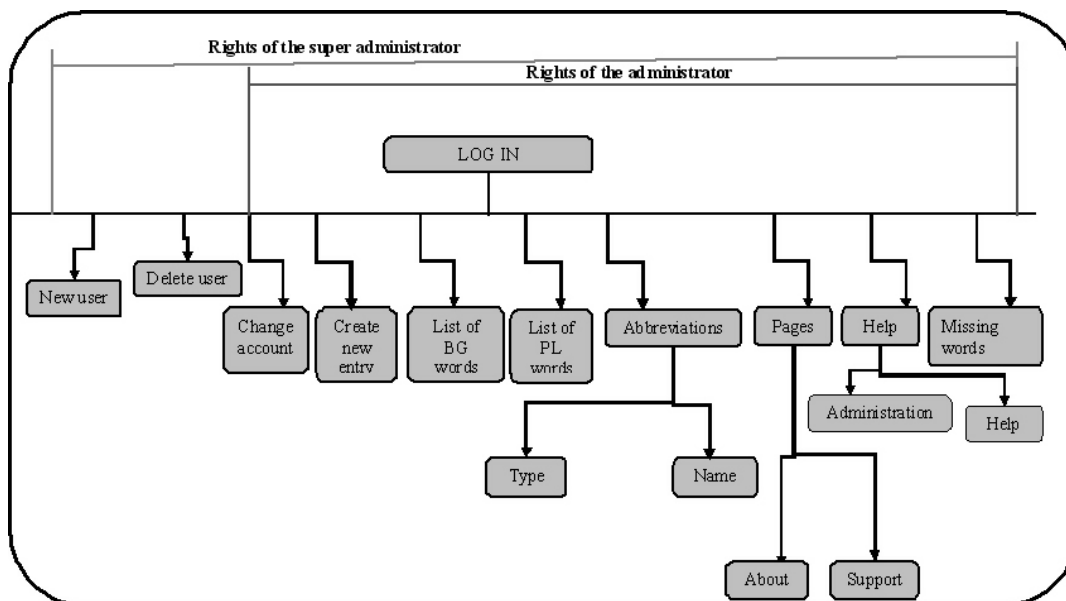
The **administrator module** is intended for the person updating the dictionary. We recognize two types of users here: (1) “**super administrator**”- who has all rights of adding, editing, deleting and searching for words; adding, editing and deleting users and (2) “**administrator**”- who has all rights except creating a new user and deleting an existing one.

Access to the administrative module is permitted only to authorized users. After a user’s password and username have been verified, the user is redirected to the administrative module where there are **several sections**: a **section** for entering a new word, **sections** for searching for Bulgarian or Polish words, a **section** where the user can enter new abbreviations, a **section where** end-users report the missing words. The **Help section** serves both the administrators and the end users.

At the beginning the user must choose from a combo box what he/she wants to enter, (what part of speech) and only the corresponding text fields are loaded.



Administrative panel — choosing the type of the word which will be added



The structure of the **Administrator module**

## 6 How to fill in the LDB?

We present here examples of entering verbal forms: two transitive verbs and a participle. In the Bulgarian-Polish electronic dictionary the corresponding entries are as follows:

завъ'ршва|м, -ш vi. kończyć, zakańczać

завъ'рш|а, -иш vp. v. завъ'ршвам

завъ'ршен part. adi. skończony, zakończony, wykończony

(1) Adding the verb *завършвам* / *end, finish off, complete, terminate* /

The entry in MS WORD format:

завъ'ршва|м, -ш vi. kończyć, zakańczać

In the **first step** the administrator or authorised user must fill in the text fields for headword, conjugation of the verb in 2<sup>nd</sup> person, singular and the conjugation type I, II, III from a drop-down list, to point if the verb is transitive /intransitive, and to choose again from drop-down list the perfect aspect (*vp*) or imperfect aspect (*vi*) of the verb (see the picture below).

There are explanations for the determination of the conjugation type and the definition of transitivity or intransitivity of verbs in the *Help-section* of the administrative module.

Въвеждане на глагол	
Индекс за омоним	<input type="text"/>
Заглавна дума *	завършвам <input type="button" value="търси в списък с думи"/>
2 л. ед.ч. сег. време *	аш <input type="text"/> Спряжение на глагола <input type="text" value="III"/>
Св. / несв. вид на глагола*	<input type="text" value="vi"/>
Преходен / непреходен глагол	<input type="text" value="transitive"/>
добавяне на обяснение към думата *	
<input type="button" value="&gt;&gt;"/>	

Administrative panel — 1<sup>st</sup> step of adding the verb *завършвам*

It is not necessary to perform the **second step** for this kind of verb, so we move to step 3. In the **third step**, we must fill in the text fields the Polish meanings of the headword. With a button “add” we can enter as many as meanings as necessary. There are also added drop-down fields and extra text fields which can be used to give detailed information for the Polish verbs usage.

Значение на полски						
№ група на точни значения*	Значение на полски*	(напр./пр.)	Сфера на употреба	Стилистично значение	Латинско значение	
1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="добави"/>
1	kończyć					<input type="button" value="изтрий"/>
1	zakańczać					<input type="button" value="изтрий"/>
<input type="button" value="&gt;&gt;"/>						

Administrative panel — 3<sup>rd</sup> step of adding the verb *завършвам*

In the **forth step** we must add examples, derivations and phrases for the current verb. As can be seen from the example above, such information is not given for the verb **завършвам**. That is way one can press the button” finish” and the new word is added to the database.

**(2) Adding the verb *завършиа* / end, finish off, complete, terminate /**

The entry in MS WORD format:

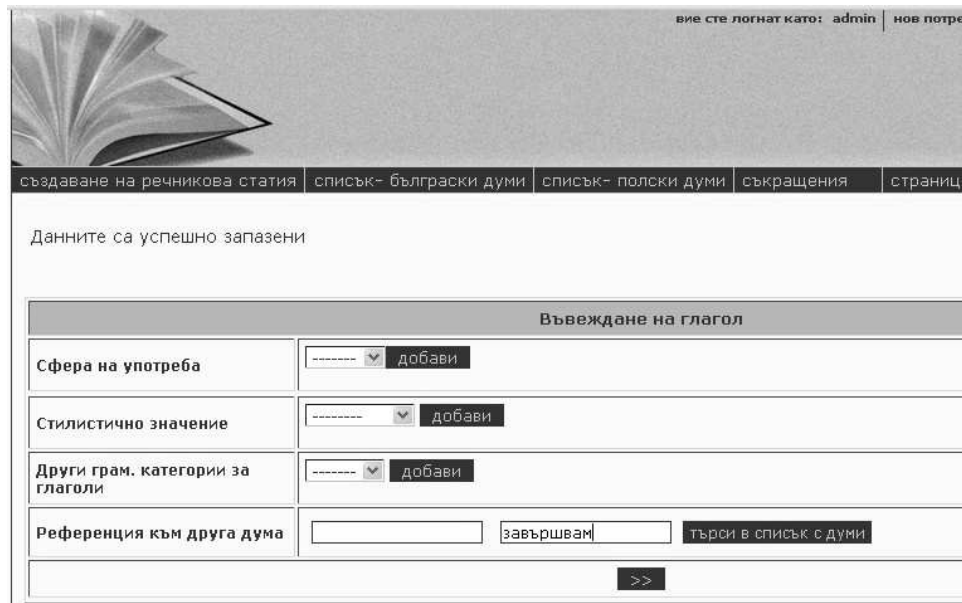
**завърш|а, -иш** вр. v. **завършвам**

In the second entry of our example — headword “*завършиа*” - there is a reference of type “*see to another verb*”. Due to this fact one must enter at first the verb from the reference — **завършвам**. To add this verb to the database of the web-application, one must perform several steps, some of which are not compulsory while others are obligatory. From the drop-down list the administrator (authorised user) must choose again the option to add a verb. The **first step** is the same as it was described for the previous word.

*Administrative panel –1<sup>st</sup> step of adding the verb **завършиа***

In the **second step** one must show that there is reference to another verb. After the administrator clicks on the button “*search from a word list*” a pop up window appears where the administrator (authorised user) can search for a word to create the reference. There is a text field and a “*search*” button. The user can search for a certain criteria. On the image below is the result of a search for all words, which begin with the prefix “*за*”. One can create a reference of type “*see only to words*” that are already stored in the database.

*Administrative panel –2<sup>nd</sup> step of adding the verb- pop up window for adding a reference to the verb **завършиа***



вие сте логнат като: admin | нов потре

сздаване на речникова статия | списък- български думи | списък- полски думи | съкращения | страници

Данните са успешно запазени

**Въвеждане на глагол**

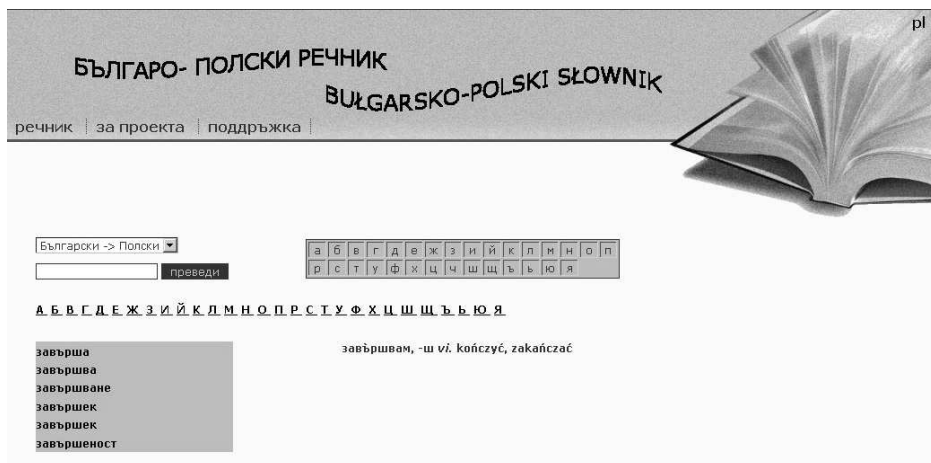
Сфера на употреба	<input type="text" value="-----"/> <input type="button" value="добави"/>
Стилистично значение	<input type="text" value="-----"/> <input type="button" value="добави"/>
Други грам. категории за глаголи	<input type="text" value="-----"/> <input type="button" value="добави"/>
Референция към друга дума	<input type="text" value="завършвам"/> <input type="button" value="търси в списък с думи"/>

Administrative panel –2<sup>nd</sup> step of adding the verb *завършвам*

## 7 How the online dictionary works?

If an end-user starts to search for a word, on the left side of the screen a list of words, starting with the given entry, is displayed. When clicking on any of these words in the list the translation is visualized in the right frame.

If one translates from Bulgarian to Polish, the whole information saved in the RDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized.



**БЪЛГАРО- ПОЛСКИ РЕЧНИК**  
**BULGARSKO-POLSKI SŁOWNIK**

речник | за проекта | поддръжка |

pl

Български -> Полски

а б в г д е ж з и й к л м н о п  
р с т у ф х ц ч ш щ ъ њ я

**А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ш Щ Ъ Ы Ю Я**

завърша  
завършва  
завършване  
завършек  
завършек  
завършеност

завършвам, -ш vi. kończyć, zakończyć

Web page for end user — translation of a Bulgarian word *завършвам*

## 8 Possibilities for the Future Improvements

- Improvement and extension of the databases (lexical and relational).
- Increase in speed of the database search.
- User feedback via a web-based application, which would allow a continuous update and extension of the dictionary resources.



## Bibliography

- [1] Dimitrova, L., Koseska-Toszewa, V. (2007). Digital dictionaries — problems and features. In *Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics, 6 July 2007, Sofia, Bulgaria. ISBN 978-954-8986-28-1*, pages 25–34, Sofia.
- [2] Dimitrova, L., Koseska-Toszewa, V. (2008a). The significance of entry classifiers in digital dictionaries. In *Lexicographic Tools and Techniques. Proceedings of the MONDILEX Open Workshop “Lexicographic Tools and Techniques”, Moscow, 3–4 October 2008. ISBN 978-5-9900813-6-9*, pages 89–97, Moscow.
- [3] Dimitrova, L., Koseska-Toszewa, V. (2008b). Some problems in multilingual digital dictionaries. *Études Cognitives, ISSN 1641-9758*, 8:237–254.
- [4] Dimitrova, L., Panova, R., Dutsova, R. (2009). Database of the experimental bulgarian-polish online dictionary. In *Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open International Workshop, Bratislava, Slovak Republic, 15–16 April 2009. ISBN 978-5-9900813-6-9*, pages 36–47, Bratislava.
- [5] Dimitrova, L., Pavlov, R., Simov, K. (2002). The bulgarian dictionary in multilingual data bases. *Cybernetics and Information Technologies*, 2(2):33–42.
- [6] Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2000). The concede model for lexical databases. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC’2000*, pages 355–362, Paris. ELRA.
- [7] Ide, N. M., Sperberg-McQueen, C. M. (1995). The tei: History, goals, and feature. *Computers and the Humanities*, 29:5–15.
- [8] Ide, N., Véronis, J. (1995). Encoding dictionaries. In *The Text Encoding Initiative: Background and Context (Eds. Ide, N., Veronis, J.)*, pages 167–179, Dordrecht. Kluwer Academic Publishers.
- [9] CONCEDE: <http://www.itri.brighton.ac.uk/projects/concede/>
- [10] TEI: <http://www.tei-c.org/index.xml>

# Theory of Lexicographic Systems. Part 2.

Volodymyr Shyrov

Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine

**Abstract.** The continuation of the theory of lexicographic systems is expounded. The connection between the general lexicographic structures and the linguistic dictionary structures is considered. The concept of lexicographic environment, which is a basis of the theory of integrated lexicographic systems, is introduced. The connection between the lexicographic effect in the information systems and the Löwenheim-Skolem theorem is established. The concepts of generalized linguistic variable, linguistic system and lexicographic calculus are introduced and considered.

**Key words:** lexicographic system, lexicographic environment, generalized linguistic variable, linguistic system, lexicographic calculus.

## 3 Lexicographic structures and dictionaries

Lexicographic systems acquire simple and pellucid interpretation when applying to the traditional dictionaries and dictionary complexes. In fact, the L-system alphabet is identified with a dictionary sign system (including wildcard characters), EIU  $I^Q(D)$  class – with a set of register units (objects of lexicographing –  $x$ ), a set of descriptions  $V(I^Q(D)) = \{V(x)\}$  – with a set of dictionary entries where register units run the  $x$  set, while  $A(x)$  and  $P(x)$  – with left and right parts of the relevant dictionary entries, etc.

The examples of designing the lexicographic systems for the concrete lexicographic works will be given in the next sections. We will receive evidence that any traditional dictionary can be presented as an L-system. At the same time, the use of L-systems constructives gives the means for the generalizations of the traditional problems of lexicography. And many of the unsettled problems are solved in the L-systems paradigm.

For example, let us consider the problem of ordering the dictionary entries by various criteria that cannot be solved in the traditional dictionary. Such a problem in the L-system is reduced to the setting the system of classifications on the  $I^Q(D)$  set, that is a basis for creating the relevant search instrument and is realized by the relevant means of the L-system internal and external models. The simplest classification – alphabetical – generates the so called lexicographic ordering the  $I^Q(D)$  set and the whole dictionary. The morphemic classification that lies in the selection of the word classes with the same stems, leads to the run-on layout typical to the dictionaries of the word-formative type. Such classification is also used in the bilingual dictionaries. The grammatical classification may generate a set of the dictionaries. Naturally, the traditional hard-copy dictionary is ordered only by one of the classifying principle (yet there are attempts to combine various classifications in one dictionary). Quite other possibilities are opened up for the general lexicographic systems and their realizations – computer dictionaries, the search instrument of which may include all the mentioned above and a chain of other classifications at the same time.

Let's define an  $\mathbf{A}$  subset of the special type in the  $\sigma[\beta]$  structure of the  $V(I^Q(D))$  L-system, the  $A$  ( $A \in \mathbf{A}$ ) elements of which are named the *automorphisms* of the  $V(I^Q(D))$  L-system. The sense of these elements is that they provide internal mappings of  $V(I^Q(D))$ , that is mappings:

$$(2.1) \quad A: V(I^Q(D)) \rightarrow V(I^Q(D))$$

of the special type – the mappings between certain dictionary entries  $A: V(x) \rightarrow V(y)$  for various  $x$  and  $y$ . In the concrete lexicographic works the  $A$  automorphism can state the availability of the reference dictionary entries like  $x \text{ } \partial u \text{ } \delta. \text{ } y(x \text{ see } y)$ . The pointed automorphism defines the following mapping of the dictionary entries:  $V(x) \rightarrow V(y)$ . Its identifier, as a rule, is a certain reference pseudoword (in the specified example – " $\partial u \delta. \mathbf{y}$ "), which confronts the  $V(x)$  dictionary

entry with its  $V(y)$  correspondence. But the  $A$  automorphism structure may be more complex than in this example.

First, the length of the references chain may be more than one, i.e. it should have chain or recursive unwinding type:

$$V(x) \rightarrow \{V(x')\} \rightarrow \dots \rightarrow \{V(x'')\} \rightarrow \dots$$

Besides, the mapping  $V(x) \rightarrow V(y)$  may represent a set of references. It is realized, for example, when the  $V(x)$  dictionary entry has the following structure:  $\mathbf{x}, x', x'', \dots$  див.  $y, y', y'' \dots$

In this case a set of mappings is defined in one  $V(x)$  dictionary entry:

$$V(x) * V(y); V(x') \rightarrow V(y'); V(x'') \rightarrow V(y'') \dots$$

For example, the dictionary entry in the explanatory Ukrainian Language Dictionary (ULD) (СЛОВНИК, 1970–1980):

**УГВИНТИТИ, УГВИНТИТИСЯ, УГВИНЧЕНИЙ, УГВИНЧУВАТИСЯ**

див. **ВГВИНТИТИ, ВГВИНТИТИСЯ** И Т.Д.

represents a set of representations:

$V(\mathbf{УГВИНТИТИ}) \rightarrow V(\mathbf{ВГВИНТИТИ}),$

$V(\mathbf{УГВИНТИТИСЯ}) \rightarrow V(\mathbf{ВГВИНТИТИСЯ})$

etc.

The elements of the  $A$  automorphisms set should not be necessary set in the explicit way. Moreover, dictionary automorphisms establishment, as a rule, is not formalized, as this is rather complex task, connected with uncovering the internal regularities and hidden structure (symmetry) of the L-system.

The  $\mathbf{H}$  and  $\mathbf{A}$  mappings sets generate the L-system *macrostructure*.

The  $\mathbf{F}\sigma[\beta] = \lambda[\beta]$  and  $\mathbf{C}\sigma[\beta] = \varrho[\beta]$  elements defined with the formula (1.25), refer to the L-system macrostructure elements. The local mappings (they are constructed as restrictions of the relevant macrostructures on  $V(x)$ ):  $H|_{V(x)}$ , and  $\lambda[\beta]|_{V(x)} \equiv \lambda(x)$ ;  $\varrho[\beta]|_{V(x)} \equiv \varrho(x)$  – the two latter are defined with the formula (1.26) – define *the microstructure*, which represent in the implicit way the semantics of the field that is an object of the concrete L-system.

The establishment and determination of the lexicographic systems microstructures let formalize and in many cases automate the process of creating the structures for the relevant dictionary databases. It gives significant advantages to the structural approach when designing the elements of the linguistic support for the information systems.

Below we introduce some concepts that help formally define the structures, which are induced on the lexicographic systems by the mappings of the special type. These mappings not only induce the natural structures on the definite lexicographic systems, but also give a set of tools for generating the structural classification of the lexicographic systems and also the methodology for the dictionary classification. Using such mappings enables to formulate the concept of closeness on the set of words (or the units of any level) – so called *pseudotopology* – and even the distances (*pseudodistances*) between words.

Let's see an elementary  $V(I^W(L)) = \{V(x)\}$  L-system. Any its  $V_i$  dictionary entry can be presented as:

$$(2.2) \quad V_i = x_0^i \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i,$$

where

$$(2.3) \quad \bigcup_{i,j} \xi_i^j = I^W(L),$$

where  $I^W(L)$  is a set of words of the  $L$  language, that are contained in all the entries of the  $V(I^W(L))$  L-system;  $\pi_i$  – the delimiters between the words (punctuation marks, auxiliary marks, wildcard characters, abridgements etc.; so called mark-up symbols). So any  $V_i$  dictionary entry is represented as a union:

$$(2.4) \quad V_i = \partial V_i \cup M_i,$$

where the following marks are used:  $\partial V_i \equiv x_0^i$  is a boundary element of the dictionary entry;  $M_i$  – the "internal" part of the entry:

$$(2.5) \quad M_i \equiv \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i.$$

Thus, any entry is a union of the internal part and the bound, and the whole dictionary is a union of set of the internal parts and the bounds. Let's mark:

$$(2.6) \quad \begin{aligned} \partial V &= \cup \partial V_i - \text{the bound of the elementary L-system } V(I^W(L)); \\ M &= \cup M_i - \text{the internal part of the elementary L-system } V(I^W(L)). \end{aligned}$$

So  $V = \partial V \cup M$ .

**Definition.** The lexicographic system is called *closed*, if for  $\forall \xi \in I^W(L) \exists V(x_0^\alpha) \equiv V_\alpha \in V$ , that  $x_0^\alpha \equiv x_\xi$ , where  $x_\xi$  is an initial form of the  $\xi$  word.

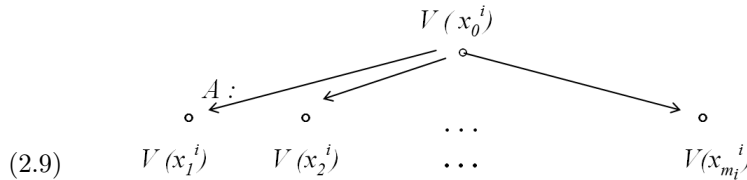
Let us define the  $A$  automorphism of the special type as the following. Let's call a set of the words:

$$(2.7) \quad (x_1^i, x_2^i, \dots, x_{m_i}^i); m_i \leq n_i$$

a *tuple* of the  $V_i$  entry, if it is a set of canonical forms to the words  $\xi_1^i, \xi_2^i, \dots, \xi_{m_i}^i$ , respectively. Not all words from  $M_i$  are included to the tuple. When examining concrete examples, in particular studying the structure of the lexicographic systems of the explanatory type, the lexicographic expediency prompts to the limitation of the tuple with the lexemes that are the part of the dictionary definitions excluding so called stop-words (the words that do not give "substantial" contribution to the semantics). Let's define the automorphism  $A \in A$  using the formula:

$$(2.8) \quad A : V(x_0^i) \equiv V_i \rightarrow \{V(x_k^i), k = 1, 2, \dots, m_i\}$$

The graphic view of the (2.8) formula looks as follows:



The (2.8)–(2.9) formulas mean that  $V(x_0^i)$  dictionary entry with  $x_0^i$  register word ("bound") corresponds to the  $V(x_k^i)$  dictionary entry,  $k = 1, 2, \dots, m_i$  with  $x_k^i$  register words,  $k = 1, 2, \dots, m_i$  respectively, if they are present in  $V(I^W(L))$ . Let's define recursively the action of the  $A$  operator on  $V(x_1^i), V(x_2^i), \dots, V(x_{m_i}^i)$ , etc. – on the results of its application to the  $V(x_1^i), V(x_2^i), \dots, V(x_{m_i}^i)$  etc. until the objects  $V(x_j^i), i, j = 1, 2, \dots$  will not repeat. Using  $V^A[x_0^i] (V^A[x_0^i] \subseteq V)$  let's mark a set of the entries  $\{V(x_0^i), V(x_1^i), V(x_2^i), \dots, V(x_{m_i}^i), \dots\}$  received as a result of the  $A$  operator action defined above.

**Definition.** A set  $V^A[x_0^i]$  is called  $A[x_0^i]$ -subdictionary of the  $V(I^W(L))$  dictionary, if  $V^A[x_0^i] = V^A[x_0^i]$ .

Thus  $A$ -subdictionary  $V^A[x_0^i]$  is an invariant set in  $V(I^W(L))$  concerning to the action of the  $A$ .

**Definition.** The  $V(x_0^i), V(x_1^i), V(x_2^i), \dots, V(x_{m_i}^i), \dots$ , elements of the  $A$ -subdictionary  $V^A[x_0^i]$  is called  $A$ -equivalent elements.

The latter definition becomes understandable when taking into consideration that a set of representations of  $A$  that generate an  $A$ -subdictionary  $V^A[x_0^i]$  induce the relation of equivalence on the set of its entries; let's mark it with  $EV^A[x_0^i]$ . Let us mark the factorset in  $V$  concerning to  $EV^A[x_0^i]$  as

$$(2.10) \quad W = V \setminus EV^A[x_0^i].$$

Hence:  $V = W \cup V'$ . According to the definition  $V' : AV' = V'A^M W = V'$  for certain  $M \geq 0$ .

Then  $V$  is a *semidirect sum* of the  $W$  and  $V'$  dictionaries:

$$(2.11) \quad V = W \triangleright V'.$$

The  $V$  dictionary with such structure is called  $A$ -indecomposable.

**Definition.** The  $V(I^W(L))$  dictionary is called  $A$ -irreducible (fully irreducible), if there are no own  $A$ -subdictionaries.

From the latter definition it follows that if  $V$  is an  $A$ -irreducible dictionary, then for any  $x, y \in I^W(L) \exists N \geq 0$ , that  $V(y) \subseteq A^N V(x)$ .

Definition. The  $V$  dictionary is called  $A$ -reducible, if it can be presented as follows:  
 $V = \bigcup_i V^i$ , and  $V^i \cap V^j = \emptyset$  when  $i \neq j$

where  $V^i$  is an  $A$ -irreducible dictionary. In this case  $V$  is expanded into the direct sum of the  $A$ -dictionaries  $V^i$ :

$$(2.12) \quad V = \sum_i \oplus V^i.$$

Let us consider an  $A$ -irreducible dictionary  $V$ :

$$V = \bigcup_{x_0^i \in S_0} V(x_0^i),$$

where  $V(x_0^i)$  is a dictionary entry with a register word  $x_0^i$ . A automorphism induces a  $S_0 \rightarrow S_0$  representation, i.e. it defines a representation of the register words sets *to itself*. This representation is defined as follows:

if  $A : V(x_0^i) \equiv V_i \rightarrow \{V(x_k^i), k = 1, 2, \dots, m_i\}$ , then  
 $A : x_0^i \rightarrow \{x_k^i, k = 1, 2, \dots, m_i\}$ .

Theorem 1. There are no invariant subsets in  $S_0$  for an  $A$ -irreducible dictionary  $V$ . So for any  $x, y \in S_0 \exists N(x, y) \geq 0$  that  $A^N x = y$ .

In other words, the  $A^N x$  path with rather large  $N$  passes through every point of the  $S_0$  subset.

Theorem 2. The full  $A$ -path on the  $G^A(S_0)$  graph is always closed.

Let's mark:  $\inf N(x, y) = \varrho(x, y)$ .

Definition. The number  $\varrho(x, y)$  is called  $A$ -pseudodistance from the word  $x$  till the word  $y$ .

The number  $\varrho(x, y)$  shows the minimum number of steps to go from the word  $x$  to the word  $y$  using algorithm  $A$ .

Let the word set:

$$(x_1^i, x_2^i, \dots, x_{m_i}^i; m_i \leq n_i) = \tau^i \equiv \tau(x_0^i)$$

be a tuple of the  $V_i$  dictionary entry, i.e. be a set of the initial forms to the words  $\xi_1^i, \xi_2^i, \dots, \xi_{m_i}^i$ , respectively. It is obvious that  $x_0^i$  also  $\in \tau(x_0^i)$ .

Definition. Let's call  $\tau(x_0^i)$  the *closed neighbourhood of the  $x_0^i$  point*.

Definition. A set

$$(2.13) \quad \{\emptyset, \tau(x_0^i), i = 1, 2, \dots, \text{Card } S_0\}$$

is called  $(V, A)$ -pseudotopology of the ELS lexicographic system.

The concept of pseudotopology can be of primary importance in the theory of the lexicographic systems, as using this concept we have a possibility to formalize the concept of closeness of the elementary information units (words). Namely, the intersection of the neighbourhoods  $\tau(x_0^i) \cap \tau(x_0^j)$  defines the closeness of the lexemes  $x_0^i$  and  $x_0^j$ .

## 4 Lexicographic environments

In reality the language objects function in its integrity, not divided into separate components of the conceptual presentation. In the lexicographic system it appears during the modeling of the language objects by means of the theory of L-systems when there is a task for integrating various types of the lexicographic effects, and also combination and coordination of heterogeneous lexicographic structures. In turn, it requires the coordination between all the elements of the L-system architecture that are subjected to the integration process.

Numerous experiments on creating the concrete computer realizations of the integrated linguistic objects let draw a conclusion on the necessity of creating special lingual information environment. Such an environment could be adopted ab origin to the processes of various lexicographic systems integration and could contain the necessary means and constructives for performing the specified processes and for fixing their results as integrated lexicographic systems that have different lexicographic structures. As a result a concept of *lexicographic environment* was proposed (Рабулець, 2002).

**Definition.** The *lexicographic environment* (*L-environment*)  $\mathbf{ML}$  is set if:

1. An  $\text{Ob } \mathbf{ML}$  class of the elements is set, each of which is a diagram like (1.35) and represents certain L-system (it may be not elementary). The elements from  $\text{Ob } \mathbf{ML}$  are called the objects of  $\mathbf{ML}$  L-environment – let's mark them with capital Latin letters:  $A, B, C, \dots$ .

2 For every pair of the objects  $A, B$  from  $\mathbf{ML}$  a set  $\text{Hom}_{ML}(A, B)$  is defined, which is called a set of morphisms from  $A$  to  $B$ , is set; instead of  $f \in \text{Hom}_{ML}(A, B)$  they also write:

$$f: A \rightarrow B \text{ or } A \xrightarrow{f} B. \text{ Where } f: CM_A \rightarrow CM_B; f: INM_A \rightarrow INM_B; \\ f: EXM_A \rightarrow EXM_B; f(\varphi_A) = \varphi_B; f(\psi_A) = \psi_B; f(\zeta_A) = \zeta_B \text{ ma } f(\zeta_A) \circ f(\psi_A) = f(\varphi_A).$$

3 For every triplet of the objects  $(A, B, C)$  from  $\mathbf{ML}$  a mapping is set

$$\mu: \text{Hom}_{ML}(A, B) \times \text{HOM}_{ML}(B, C) \rightarrow \text{HOM}_{ML}(A, C)$$

( $\mu(f, g)$  image of the  $(f, g)$  pair, where  $f \in \text{Hom}_{ML}(A, B), g \in \text{HOM}_{ML}(B, C)$ ), will be marked as  $f \circ g$  or  $f g$  and will be called an  $f$  and  $g$  morphisms composition).

4  $\text{Hom}_{ML}(A, B)$  sets and morphisms composition satisfy the following axioms:

(a) Associativity: for every triplet of  $f, g, h$  morphisms:

$$\begin{array}{c} f \quad g \quad h \\ A \rightarrow B \rightarrow C \rightarrow D \end{array} \Rightarrow f(g h) = (f g) h.$$

(b) Unit existence: for every  $A \in \text{Ob } \mathbf{ML}$  there exists a morphism  $1_A: A \rightarrow A$ , ( $1_A \in \text{Hom}_M(A, A)$ ), where  $1_A f = f$  and  $g 1_A = g$  for any morphisms  $f \in \text{Hom}_{ML}(B, A)$  and  $g \in \text{Hom}_{ML}(A, B)$ .

(c) If  $(A, B)$  and  $(A', B')$  pairs are different, the intersection of  $\text{Hom}_{ML}(A, B)$  and  $\text{Hom}_{ML}(A', B')$  is empty.

Let two lexicographic environments  $\mathbf{ML}_1$  and  $\mathbf{ML}_2$  be set. Covariant (respectively contravariant)  $F$  functor from  $\mathbf{ML}_1$  to  $\mathbf{ML}_2$  consists of:

- (a)  $A \rightarrow F(A)$  mapping that compares each object  $A \in \text{Ob } \mathbf{ML}_1$  with object  $F(A) \in \text{Ob } \mathbf{ML}_2$ ;
- (b) mappings  $F(A, B) : \text{Hom}_{ML_1}(A, B) \rightarrow \text{HOM}_{ML_2}(F(A), F(B))$  — for a covariant functor and  $F(A, B) : \text{Hom}_{ML_1}(A, B) \rightarrow \text{HOM}_{ML_2}(F(B), F(A))$  — for a contravariant functor defined for every pair  $(A, B)$  of the objects from  $\mathbf{ML}_1$  and for those where (if write  $F(u)$  instead of  $F(A, B)(u)$ )  $F(1_A) = 1_{F(A)}$  and  $F(vu) = F(v)F(u)$  (respectively  $F(vu) = F(u)F(v)$ ).

## 5 Integrated L-systems and the methods for their creation

The L-environment structure introduced above is a convenient formal object for forming complex lexicographic constructions that combine a lot of separate heterogeneous L-systems in a whole. The heterogeneity of the L-systems subjected to the integration is a multiaspect concept. Let us see the L-systems, which are heterogeneous on all the levels of the architecture – conceptual, internal and external. Such approach foresees the development of the methods for the conceptual models integration, the ways of representing the data and operational software-based platforms, and coordinating the external mappings of the relevant conceptual schemes and their internal mappings.

The integration of L-systems as information systems is an achievement of the possibility for simultaneous and common use of several information systems as a whole by the application. The Ukraine's State Standard 2941–94 proposes similar definitions: integrated system – an aggregate of two or several interrelated systems where the functioning of one of them depends on the results of the functioning of other (others) so that this aggregate can be considered as one system; integration of systems – combination of several systems for various purposes into a multifunctional system.

So for an application the integrated aggregate of various lexicographic databases (LDB) should be of the form of one LDB. Such conceptions may be transferred on the traditional dictionaries that can also be the integrated L-systems. In particular the explanatory dictionary is one of the best examples of such L-system.

In this section the maximum accent is made on L-systems integration on the conceptual level. The basic objects that should be integrated are the diagrams like (1.35) that are considered as objects of certain L-environment.

Let two  $A$  and  $B$  objects be, and  $f : A \rightarrow B, f \in Hom_{ML}(A, B)$ . Let us construct  $f$  as special morphism like three-component vector:  $f = (f_c, f_i, f_e)$ , where the diagram:

$$(2.14) \quad \begin{array}{ccc} & CM_B & \xrightarrow{\psi_B} \text{imm}_B \\ & \swarrow f_k \quad \searrow \varphi_B & \nearrow f_i \quad \searrow \zeta_B \\ CM_A & \xrightarrow{\psi_A} \text{imm}_A & \downarrow \zeta_B \\ & \swarrow \varphi_A & \searrow \zeta_A \\ & & \text{exm}_B \\ & \searrow f_e & \\ & & \text{exm}_A \end{array}$$

is commutative – this is equivalent to the following equalities:

$$(2.15) \quad \psi_B \circ f_k = f_i \circ \psi_A; \quad \varphi_B \circ f_k = f_e \circ \varphi_A; \quad \zeta_B \circ f_i = f_e \circ \zeta_A$$

The (2.14) diagram and the (2.15) equalities give the formal means, by the language of which the processes of creating the integrated architecture can be formulated.

The morphism  $f = (f_k, f_i, f_e) : A \rightarrow B$  is called regular if it satisfies the conditions: — of the full definiteness, according to which any  $A$  state corresponds to one and only one  $B$ state;

— of interpretability, according to which any element belonging to the set  $\{\sigma_B[\beta_B], RR \downarrow [V(B)]\}$ ;

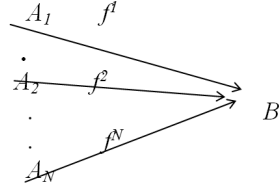
— of reproducibility, according to which the change of any  $A$  state that is performed by certain operator with  $\{\sigma_A[\beta_A], RR \downarrow [V(A)]\}$  corresponds to identical change of the relevant state that is performed by certain operator with  $\{\sigma_B[\beta_B], RR \downarrow [V(B)]\}$ .

Except for the possibility for integration of the conceptually heterogeneous lexicographic systems that represent various language phenomena, the important aspect of the integration architecture is a possibility for reaching the high level of independence for the application from DBMS and for providing their mobility concerning to DBMS of various types. The problem of program mobility is formulated as providing possibility for performing certain program on various computer platforms without its changing. The problem of programs mobility concerning to DBMS is defined similarly.

Theoretically there are the following possibilities for this: 1) universal computer language creation and the requirement for its universal use; 2) providing each computer with compilers for all the computer languages in case of its proper standardization; 3) introduction of the platform independent intermediate language with its instrumental realization on the level of virtual machine with the embedded interfaces for any platforms; 4) applying the emulation methods; 5) use of computer networks if at least one of the network computers is provided with the necessary compiler.

The analysis of these propositions leads to conclusion on the reasonableness of the third of them. In different variations they try to realize it in the systems like JAVA, standardization of the machine independent languages SQL, SQL2 etc. The specified propositions outwardly are similar to the designing common conceptual model that confirms the statement on the centrality of the conceptual model in the information system architecture. So let us examine the L-system integration processes development on the conceptual level (or L-systems conceptual mappings integration).

The content of the integration procedure consists in the following. Let us assume that instead of one object  $A$  present in the (2.14) diagram there is a set of objects  $A_1, A_2, \dots, A_N$ . Let us construct a fan-shaped mapping with morphisms  $f = (f^1, f^2, \dots, f^N)$ :



$$(2.16) \quad f^p: A_p \rightarrow B, p = 1, 2, \dots, N,$$

where every  $f^P$  is set by three-component vector  $f^P = f_e^P, f_i^P, f_e^P$  with characteristics (2.15).

In the process of using these morphisms the integration procedure is used that has the following stages. First an auxiliary L-system  $B$  is constructed from the following elements: sign systems  $\mathbf{A}(A_p)$ ; structures  $\beta_p, \sigma_p[\beta_p]$  and processes  $RR_p \downarrow [V(I^Q(S))]$  of all  $A_p$ . On this stage the mapping  $f^p: A_p \rightarrow B, p = 1, 2, \dots, N$  is interpreted as mapping of the enclosure. So certain not elementary L-system  $B$  with independent components  $A_p, p = 1, 2, \dots, N$  is come out.

The sign system for L-system  $B$  is constructed as a unification:  $\mathbf{B}(B) = \cup \mathbf{A}(A_p)$ .

Then a special semantic procedure SEM is entered, which provides the identification of the structure elements that coincide at least for two systems from the set  $A_p, p = 1, 2, \dots, N$ . Let us assume that this procedure performs the identification of not only the names of semantically identical attributes present in various L-systems  $A_p, p = 1, 2, \dots, N$ , but also the relevant fields of their values (domains). Different cases for crossing the structures that belong to different  $A_p, p = 1, 2, \dots, N$ , are possible. Let us examine them in detail.

Let us mark with  $\varepsilon_{p_1 p_2 \dots p_k}[\beta]$  a set of structure elements (with their domains) that belong to each from  $A_{p_1}, A_{p_2}, \dots, A_{p_k}, 1 \leq p_1 < p_2 < \dots < p_k \leq N$  at the same time. Let us apply a semantic operation to it:

$$(2.17) \quad \text{SEM}(\varepsilon_{p_1 p_2 \dots p_k}[\beta]) = \varepsilon_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta]$$

and thus receive semantically identical structure elements of the L-systems with numbers  $p_1, p_2, \dots, p_k$ . The availability of the elements (1.39) gives a possibility for constructing L-system with a structure:

$$(2.18) \quad \sigma_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta] = (\varepsilon_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta], [b_{p_1}^{\text{SEM}}], [\beta_{p_2}^{\text{SEM}}], \dots, [\beta_{p_k}^{\text{SEM}}], R^{\text{SEM}}),$$

where  $[\beta_i^{\text{SEM}}], i = p_1, p_2, \dots, p_k$  — structure elements of the L-systems  $A_i$ , in which the identification of common elements is performed;  $R^{\text{SEM}}$  — a unification of operations that act in each L-system. The other structure elements of the L-systems  $A_i, i = p_1, p_2, \dots, p_k$ , — let us mark them with  $S_i, i = p_1, p_2, \dots, p_k$ , remain without changes.

Thus L-system  $B$  with a structure:

$$(2.19) \quad \{ \sigma_{p_1 p_2 \dots p_k}^{\text{SEM}}[\beta]; S_i, i = p_1, p_2, \dots, p_k \}$$

integrates the L-system  $A_i, i = p_1, p_2, \dots, p_k$ .

Let us mark with  $\varepsilon[\beta]$  a set of structure elements (with their domains, that belong to all  $A_p, p = 1, 2, \dots, N$ ):

$$(2.20) \quad \varepsilon[\beta] = \bigcap_p^N [\beta] A_p.$$

As a result the L-system SBS with a sign system  $\mathbf{A}(B) = \cup \mathbf{A}(A_p)$  and structure  $\beta(B)\sigma[\beta](B), RR \downarrow [V(B)]$  is obtained. There is an interesting case when certain element or some elements of the structure are present in all the L-systems that are subjected to integration, i.e. when  $\varepsilon[\beta] \neq \emptyset$ . Then it's reasonable to hold the entire indexing of all  $A_p$  by the meanings of the relevant domains that belong to the elements  $\varepsilon[\beta]$ . These structure elements acquire the status of input to each of  $A_p$ . As an experience shows, the elements of the structure of the specified type for the natural language systems are the lexical arrays (wordforms sets) of all the languages that take part in  $A_p$ . Thus the task for constructing the procedure of natural language indexing appears.

The L-system SBS constructed in such way is called an integration of L-system  $A_p, p = 1, 2, \dots, N$ .



The lexicographic environment structure contains a set of elements necessary for presenting grammatical and lexical semantics. In particular it concerns to presenting certain relations and language phenomena that abstract away from the language continuum – word change, word building, orthoepy, phraseology, synonymy, antonymy etc. Moreover the lexicographic environment construction allows the modeling of each from the specified phenomena and relations apart with the following their integration to one lexicographic complex. The examples of applying this theory when constructing concrete lexicographic environments for explication of the properties of the Ukrainian lexics are set in the following sections.

## 6 Systemologic aspects of lexicographic effect and Löwenheim-Skolem theorem as its formal correlate

The concept of lexicographic environment gives means for developing the methods and technique of L-system integration and creating the complexes similar to the dictionaries of the unlimited complexity. So any dictionary, any dictionary system necessary gets to the lexicographic environment class and thus can be presented as a lexicographic environment with a certain structure. In this sense the question of classification and typology of the dictionaries obtains a wide field for development and various generalizations.

But the concepts of lexicographic system, lexicographic environment and everything connected to it should not necessary concern only to dictionaries and the complexes of them – they can be applied for the description of larger range of the lingual information processes. Or even more: specified concepts are compulsory everywhere where the lexicographic effects are developed. Due to their universal nature it opens the way for comparing the structures of the lexicographic systems with other examples of the formally defined structures.

As lexicographic effect is expressed in the presentation of certain (continuous) universe using discrete sets, a good possibility for its formalization through so called “Scolel paradox” appears. In the theory of models it is known as Löwenheim-Skolem (Berry, 1953; Chang, Keisler, 1990) theorem. It consists in the following: within some limitations there is certain isomorphism between the innumerable and numerable sets; in a certain sense, the potential infinity may be interpreted in a finite way. Per se, a finite model of infinity is constructed. The Löwenheim-Skolem theorem affirms that any solvable theory of the first order, which has an innumerable model, also has a numerable model. It means that if certain structured set is set by the numerable set of rules, then there is a numerable set (i.e. a subset of the natural numbers set), on which one can construct the exact model of this structured set where all the initial axioms will be performed. So this is the same presentation of the infinite object through the finite object that contains all the information on the infinite object. In this way the Löwenheim-Skolem theorem really acts the part of formal correlate for the lexicographic effect in information systems which play the role of some phenomenological principle.

## 7 The deneralized linguistic variable

We have already used the ideas of the fuzzy sets theory while expounding the concept of semantic states of language units. The purpose of this paragraph is to show that the fuzziness is naturally contained in the structure of the lexicographic systems and that on the basis of the developed theory of the lexicographic structures the generalization of the well known concept of linguistic variable can be formulated. This concept can be applied to any system of analysis, understanding, decision-making, etc., where the information is presented primarily in its unstructured, natural language form.

Let us specify and detail these ideas following the works (Заде, 1976; Борисов, 1987), where the information instrument used for processing the fuzzy information in the decision-making systems has been developed. The concept of linguistic variable is a basis of this instrument. Applying this concept allows to formalize and automate the decision-making process in the difficult situations of management. Within the linguistic approach not only numbers, but also the words,

syntagmas, collocations and sentences of the natural language are allowed as the variable values. The instrument for their formalization is the theory of fuzzy sets.

Let us give the basic definitions of the theory of fuzzy sets and of the linguistic variables.

Let  $U = \{u\}$  is a universal set. A set of pairs will be called as a fuzzy set on the  $U$ -set:

$$(2.21) \quad A = \{ \langle \mu_A(u), u \rangle \}$$

where  $\mu_A: U \rightarrow [0,1]$  is a mapping of  $U$  set into a single segment  $[0,1]$ , which is called a function of belonging. Let us use the following marks:

$$A = \bigcup_{u \in U} \mu_A(u) / u = \bigcup_{u \in U} \mu_u / u$$

The  $u$  variable is called basic. The possible interpretation of the function of belonging:  $\mu_A(u)$  is a subjective measure of how the  $u \in U$  element corresponds to the concept or idea, the sense of which is formalized with the  $A$  fuzzy set. The following set is called an  $A$  fuzzy set medium:

$$(2.22) \quad S_A = \{ u \in U : \mu_A(u) > 0 \}$$

Let two universal sets  $U = \{u\}$  and  $V = \{v\}$  are given. The following set of pairs is called a fuzzy binary R relation on the  $U \times V$  set:

$$(2.23) \quad R = \bigcup_{(u,v) \in U \times V} \mu_R(u,v) / (u,v)$$

where  $\mu_R(u,v)U \times V \rightarrow [0,1]$ : is a function of belonging the fuzzy R relation, which has the same sense as  $\mu_A(u)$ . The generalization on the n-measurable case is obvious.

The fuzzy variable is defined by the triplet  $\langle A, U, \tilde{A} \rangle$ , where:

$A$  is a fuzzy variable name;

$U$  is a universal set, the domain for the defining the fuzzy variable;

$\tilde{A} = \bigcup_{u \in U} \mu_u / u$  is a fuzzy set on  $U$ , which describes the restrictions on possible

numeric values of the  $A$  fuzzy variable.

The linguistic variable is defined by the quintet:

$$(2.24) \quad \langle N, T, U, G, M \rangle$$

where  $N$  is a linguistic variable name;  $T$  is a set of its values or terms that in turn play a role fuzzy variable names with the  $U$  domain (the basic term-set of the linguistic variable);  $G$  is a “syntactic” procedure, which describes the process of forming new, meaningful values of the linguistic variable from the  $T$  set; let us mark the result of applying  $G$  to  $T$  as  $G(T)$ ; a set  $T^* = T \cup G(T)$  will be called the extended term-set of the linguistic variable; the “semantic” procedure, which allows to assign certain semantics to each new value of  $T^*$  by forming the relevant fuzzy set, will be marked as  $M$ .

Depending on the  $U$  set nature the linguistic variables are divided into numeric and nonnumeric ones. The numeric fuzzy variable is a variable, for which  $U \in \mathbf{R}^1$  and which has a measurable basic variable. The fuzzy variables that correspond to the values of the numeric linguistic variable will be called fuzzy numbers.

The nonnumeric linguistic variables have domains  $U$  which consist of nonnumeric objects, in particular, words and other constructions of the natural language.

The structural theory of the lexicographic systems provides the means for generalizing the concept of linguistic variable and introducing the concept of generalized linguistic system, the basis for which is a recursive reduction procedure, the lexicographic structure of L-systems and the concept of semantic state of language units.

While analyzing the structure of the elementary semantic states in the lexicographic system BELS, we find that the lower floors end with illustrations (“microcontexts”) that play the role of the terminal elements of the explanatory part. So the overall presentation of the elementary semantic state in the lexicographic system can be given with a formula:

$$(2.25) \quad \Psi_{MN}^Q(X) = X \rightarrow C_M^Q \rightarrow J_{MN}^Q$$

or, in more detail:

$$(2.26) \quad \Psi_{MN}^Q(x) = x \rightarrow C_M^Q \begin{array}{l} \nearrow J_{M1}^Q \\ \rightarrow J_{M2}^Q \\ \vdots \\ \searrow J_{Mn(N)}^Q \end{array}$$

where  $Q, M, N$  are group indices that specify the types of ways that appear in the definition of structural graphs of the relevant semantic states.

We have already noticed that not only text segments, but also “fragments of reality” like video, audio, numeric data, images, etc. can be used in the dictionaries as illustrations. We will consider that interpretation of the lower (terminal) elements as illustrations to the meanings of the relevant elementary information units is also fair in the lexicographic systems of general type. Naturally, the contribution of various illustrations to the interpretation of the relevant elementary information units (the interpretation of word meaning) is not the same, although there are no universal methods of evaluating “the validity” of such contribution.

While formulating (2.25)–(2.26) definitions, we implicitly assumed that the lexicographic structure  $\varrho$  has a hierarchical composition and is presented with a graph. In general, we assume that the  $C_{\{M\}}^{\{Q\}}$  values have a common set of  $J_{\{M\}\{N\}}^{\{Q\}}$  terminals for the whole permissible set of values of the indices  $\{\{M\}\{N\}\}^{\{Q\}}$ .

Now let us formulate the concept of generalized linguistic variable. Its mechanism is quite simple and reduces to the generalization of the elements that appear in (2.24) definition by interpreting them as elements of the structure of certain lexicographic systems. In other words, we will consider that the quantities represented in (2.24) formula, are defined in accordance with the structural theory of lexicographic systems. This approach is justified with the fact that for the traditional definition of the linguistic variable (2.24) it is easy to construct a basic lexicographic system, the structure forming elements of which match the elements given in this definition.

As a result of the process we receive new, generalized definitions of the linguistic variable elements.

(1). *A set of  $N$  names* of the generalized linguistic variable is a set of the elementary information units of  $D$  system (their names) concerning to the lexicographic effect  $Q$ :  $N = \{x, x \in I_0^Q(D)\}$ . Since  $x \in I_0^Q(D)$  identify the elements of the appropriate ELS:  $x \Leftrightarrow V(x) \in V(I_0^Q(D))$ , we will consider this correspondence to be established in defining the generalized linguistic variable and will identify a set of  $N$  names of the generalized linguistic variables with a set of pairs  $\{x, V(x)\}$  if necessary. For the names of the generalized linguistic variables we set an interpretation as terms (dictionary entries) of the relevant lexicographic system  $ELS[I_0^Q(D)] = \{I_0^Q(D); V(I_0^Q(D)) \equiv (A(I_0^Q(D)); P(I_0^Q(D)); H; A; \lambda; \varrho; \Sigma)\}$ .

(2). *A set of  $T(x)$  terms* of the generalized linguistic variable  $x \in I_0^Q(D)$  is a set of its elementary semantic states  $J_{\{M\}\{N\}}^{\{Q\}}(x)$  that are defined with formulas (2.25)–(2.26).

(3). *A basic  $U(x)$  term-set of the set  $T(x)$*   $J_{\{M\}\{N\}}^{\{Q\}}(x)$  for the generalized linguistic variable  $x \in I_0^Q(D)$  is a set of illustrations  $J_{\{M\}\{k\}}^{\{Q\}}, k = 1, 2, \dots, n(N)$ , – the terminal elements of the system of the semantic states  $J_{\{M\}\{N\}}^{\{Q\}}(x)$ . From the context of defining the quantity  $J_{\{M\}\{k\}}^{\{Q\}}$  it is clear that they can take both numeric and nonnumeric values, i. e. they have a combined nature.

(4). *The function of belonging the basic  $U(x)$  term-set* is a mapping:

$$\mu^J(x): J_{\{M\}\{k\}}^{\{Q\}} \rightarrow [0, 1],$$

the introduction of which transforms *the basic  $U(x)$  term-set* and the whole generalized linguistic variable into the fuzzy object.

(5). For the generalized linguistic variable it is reasonable to define a certain auxiliary function – *the function of belonging the set  $T(x)$*

$$\mu(x): J_{\{M\}\{N\}}^{\{Q\}}(x) \rightarrow [0, 1],$$

the introduction of which transforms *the set of  $T(x)$  terms* into the fuzzy object.

(6). *A set of procedures for extending the  $I_0^Q(D)$  system:*  $W = \{w_i, i=1, 2, \dots\} = \{A; \pi; \dots\}$ , where  $A$  is a set of automorphisms ELS;  $\pi$  – generating function –  $\pi: I_0^Q(D) \rightarrow I^Q(D)$ .  
 $w_i: I_0^Q(D) \rightarrow \hat{I}^Q(D)$ ;  $I_0^Q(D) \subseteq I_0^Q(D) \subseteq \hat{I}^Q(D)$ .

(7). *A set  $M$  – semantic procedure – is providing content of the linguistic variables to the elements of the extended  $\hat{I}^Q(D)$  system.*

Definition 2. The (1) – (7) statements define the *deneralized linguistic variable*.

The concept of *linguistic system* is naturally formulated on the basis of the concept of deneralized linguistic variable. This is done in the following way.

Definition 3. *The linguistic system* is defined with an octet:

$$(2.27) \quad \langle N; V; T; U; \mu^J; \mu; W; M \rangle,$$

the elements of which are interpreted as:

$N = \{x, x \in I_0^Q(D)\}$  is a set of names of the generalized linguistic variables, that is a set of elementary information units of a certain (basic) lexicographic system;

$V \equiv \text{ELS}[L] = \{I_0^Q(D); V(I_0^Q(D)) \equiv (\Lambda(I_0^Q(D)); P(I_0^Q(D)))\}$ ;  $H; A; \lambda; \varrho; \Sigma\}$  is a basic elementary lexicographic system;

$T = \cup T(x)$  is an aggregate of term sets of the generalized linguistic variables,  $x \in I_0^Q(D)$ ;

$U = \cup U(x)$  is an aggregate of basic term-sets for  $T$ ;

$x \in I_0^Q(D)$

$\mu^J = \{\mu^J(x), \forall x \in I_0^Q(D)\}$ ;

$\mu = \{\mu(x), \forall x \in I_0^Q(D)\}$ ;

$W$  is a set of procedures for  $I_0^Q(D)$  system extension;

$M$  is a set of procedures for giving the content of the linguistic variables to the elements of the extended  $\hat{I}^Q(D)$  system.

The linguistic system is a certain type of the fuzzy information system, because it is really an information system (as a lexicographic system that is its substratum), which develops a range of fuzzy relations.

Here are examples of the generalized linguistic variables and the generalized linguistic systems.

In the lexicographic system of the Ukrainian Language Dictionary, the structure of which is used here in the form constructed in the fifth section of the book (Широков, 2004), the linguistic system structure is induced in the following way.

The register units of the ULD, which form the  $I_0(U)$  set, serve as generalized linguistic variables here.

For each  $x \in I_0(U)$  the relevant  $T(x)$  term-set is defined with formulas:

$$\psi_i^J = (x) \rightarrow C_i, \quad i = 1, 2, \dots, n(x)$$

$$\psi_{ik}^V = (x) \rightarrow C_i \rightarrow V \rightarrow V_{ik}, \quad k = 1, 2, \dots, n(i)$$

$$\psi_{ip}^F = (x) \rightarrow C_i \rightarrow FC \rightarrow FC_{ip}, \quad p = 1, 2, \dots, n(k)$$

$$(2.28) \quad \psi_{ipr}^{FV} = (x) \rightarrow C_i \rightarrow FC \rightarrow FC_{ip} \rightarrow FC_{ipr}^V, \quad r = 1, 2, \dots, n(p)$$

that are got from the (2.25)–(2.26) formulas with a reduction of the latter ones by the relevant  $J$  elements.

A set association of the relevant illustrations serve as the basic  $U(x)$  term-set of the  $T(x)$  term-set:

$$(2.29) \quad \{J_{ij}\} \cup \{J_{ikl}^V\} \cup \{J_{ipq}^{FC}\} \cup \{J_{iprs}^{FCV}\}$$

The definition of  $W$ , which is a set of procedures for extending the  $I_0(U)$  system for ULD, may include the following elements:

- the relations of paradigmization;
- the thesaurus on  $I_0(U)$  [“genus-species”, “part-whole”, “complex-element”, “cause-effect”];
- the relations of synonymy;
- government and agreement;
- word equivalents;

– idioms, etc.

The function of belonging  $\mu$  is defined with the (2.22)–(2.23) formulas. The procedures for defining the specific  $\mu$  values can use the mechanisms of expert appraisals at the forming the linguistic system or other mechanisms, but we should take into account that the  $\mu$  values may vary in the process of the semantic analysis depending on the following factors: introduction of the additional information, identification of some elementary semantic structures from the analyzed text that immediately leads to the reduction of distribution (2.24)–(2.26), etc. The definition and introduction of these mechanisms to the semantic analyzer gives the necessary flexibility to the system, which brings together its behavior with the behavior of the human analytical systems. Using this formalism, the concepts, statements and results relating to the linguistic criteria, fuzzy expressions, linguistic lotteries, information granules, etc., are easily transferred and generalized.

## 8 «Quantum» linguistics

The fundamental role of discretization, the peculiar «quantification» of reality demonstrated by the lexicographic effect and Löwenheim-Skolem theorem, induces to more intent examination of the natural theories, for which the discretization is typical. Here we mean quantum mechanics. Just there first in the science history of the new time it was ascertained that all natural processes are quantized due to existence of the fundamental quantity of minimal interaction – Planc constant. It is necessary to notice that according to the classical science the nature is continuous and interaction between certain physical objects may be unrestrictedly small. The limitation of interaction intensity from below was put with introducing Planc constant. Not less fundamental consequences result from this fundamental fact.

At first, as space and time remain continuous and so there are unrestrictedly small areas of space and time, then the energy density with the transition to smaller cells of space and time should increase.

In connection with the stated above it is necessary to give certain general scientific considerations as to the conception of system states. The specified concept used in many natural, social and technical disciplines is deeply theoretically and practically worked out in the quantum mechanics where it is basic.

According to the canonic doctrine of the quantum mechanics each system at certain period of time is at a certain state (with a certain probability). The system state is formalized as a solution of Schrödinger equation for the system. As Schrödinger equation is a certain type of differential equation in the partial derivatives, a set of its solutions which is identified with the states of the examined system, forms the infinite-dimensional Hilbert space. So the number of states for the quantum-mechanical system is theoretically infinite.

The system state represents the most full its description in theory and defines the probabilistic interpretation. But the state itself is not a directly observable. The observables are represented in the quantum mechanics by the Hermitian operators acting in the Hilbert space of the states, and possible meanings of the observables are calculated as matrix elements of these operators in the space of the states. But in some other theories the system states are really observable. For example, in classic mechanics the state of material point is set with a pair coordinate – impulse at the present moment:  $(x(t), p(t))$ , which are observable – both separately and together. In the quantum mechanics there is a fundamental limitation for simultaneous measuring the coordinates and impulses that is defined by the Heisenberg uncertainty principle.

So the concept and status of the observables is not invariant and is defined differently in different natural-science (and other) theories. It adds some piquancy for using the concept of state in the theory of calculi, which in its present view ignores the phenomenon of observability.

The requirement for the theory to operate only the observable quantities could be put. But this question is not so easy, because both observable quantities and directly not observable ones are used for the states characterization. They should have different logical and ontological status. So an interpretation of correlation between the observable and directly not observable quantities of certain theory can be presented: *they represent the “formal” and “substantial” sides of the*

*investigated object respectively and can be formalized as register part and interpretation part of certain hypothetical L-system respectively.*

Applying to the language objects this interpretation can be detailed so that any language unit state could be distributed on formal and substantial parts. The formal part is available for direct perception by the object – it can be sound or graphic presentation. The substantial part is represented by the aggregate of “all contexts” where this language unit can function – this circumstance makes the specified part of the state not fully observable.

In the scientific discussion on logical and psychological foundations of the phenomenon of observability it is important to say about Mach’s principle. According to it the sensual effects are ordered in human mentality so that these effects are grouped to stable complexes economically. A. Einstein considered this principle too banal for being the universal gnoseological law, but he pointed out the particular role of language in the ontological-logical-psychological expansion of the cognition process (Гейзенберг, 1989). He thought that language structures were not only the way for sensual complexes fixation, but also the reflection of what existed (or what even could exist) behind the measures of that complexes and without connection between them. Einstein’s comment concerning to the language role confirms our thesis on the universality of the lingual information processes on all levels of the reality.

The following comments concern the criterion of simplicity for the scientific theory. Let us notice that the quantum theory was formulated the concepts of simplicity and complexity were general; the theory of complexity was formulated later – in the 1950s. The connection of such characteristics as complexity of the objects and their descriptions with information was not yet found out. The concept of complexity developed by A. Kolmogorov and other scientists, its connection with information aspects of the reality description and with the concept of information and its quantitative measure, has a deep connection with the criterion of simplicity and beauty of the scientific theory. The minimality of the investigated object description, which is an objective measure of the quantity of information on this object, induces the scientists to find descriptions of such type, but does not point the ways for it. However the absence of these ways is not a disproof of the objectivity for the minimum description existence – it is only the evidence that there are no formulas or algorithms for obtaining new truths of science. And when such a description is found, it should be the simplest. So the principle of simplicity (or beauty) of the scientific theory is not as much the result of the principle for mentality saving, but it rather follows the general nature of the information and corresponds to the formal determinations of information measure by A. Kolmogorov.

When there is the most adequate description of the object (process, system etc.) investigated, this description should be minimal, as it gives only the substantial information on the nature of the object investigated and does not contain the descriptions of the casual, unimportant details, which add unnecessary elements to the description. The scientist instinctively strives for obtaining the description of the investigated things that will be agreed with determination of information quantity with Kolmogorov’s measure. So when the researcher receives a formula, equation etc., he feels himself confident.

The formalism of the theory of complexity is both clear and deep; it should be perceived ontologically as an objective property of things. One of nontrivial displays of the mentioned feature is that the complexity of the composite formation is not equal to the sum of complexities for the entities that form it. That is, complexity is not an additive function of the system. In other words, if there is a system composed of other, “smaller” subsystems that are its constituents, that is if:

$$(2.30) \quad D = \bigcup_i D_i,$$

where the investigated system is marked with  $D$  symbol, and its constituents are marked with  $D_i$ , then

$$(2.31) \quad K(D) \neq \Sigma K(D_i)$$

where  $K(D)$  is a quantitative measure of system  $D$  complexity, and  $K(D_i)$  – quantitative measures of its constituents  $D_i$  complexity respectively (usually  $K(D) < \Sigma K(D_i)$ ); the mentioned concepts are spread on separate  $D_i$ , and on their constituents.

In the process of formation, functioning and interaction of the composite systems, the phenomenon of “*complexity self-compensation*” is taken place. The content of this phenomenon is the following. The constituents that are identified as a composite object display in the “connected” state only a part of their total, “immanent” complexity. Such behavior provides the principal possibility of cognizing “the shown” being and maybe even its existence. Otherwise, the complexity of any object would be actually endless. But the complexities of separate components are “self-compensated” in the formation of the whole. So we can affirm that potentially the complexity of any thing is infinite, because today we see no limits for matter divisibility and each lower structural level has a nonzero complexity. But simultaneously all kinds of the component complexity are not shown in general, they are revealed only by the level. So, the complexity in each case is renormalized if you go by analogy with quantum electrodynamics, where you should apply the procedure of “subtracting the infinities” to remove the differences. The language gives us a demonstrative example of complexity self-compensation. For example, the complexity measure of the specific word can be considered as a length of the relevant dictionary entry of the explanatory dictionary, which considers the effects of grammatical and lexical semantics, including a set of grammatical meanings, lexical polysemy, phraseological word structure etc. Meanwhile, a word functions in a sentence only in the certain sense – in one or in a “mixture” of several possible meanings for polysemantic lexemes. Thus only part of the dictionary entry is the word complexity measure in the concrete context. In some cases it may be a tenth and even one hundredth of the total complexity of a lexeme. Thus, the whole sentence complexity may be less than the full complexity of a separate word, which is its integral part.

The construction of being appears paradoxical! It turns out that complex things actually consist of more sophisticated ones. In this sense “more” is lower for “less”. The nontrivial confirmation to this thesis is well-known effect that has ontological and psychological dimension – it concerns the complexity of scientific theories. The theory of atoms, for example, does not seem simpler than the theory of molecules, the theory of nuclei does not seem simpler than the theory of atoms, the theory of elementary particles is not simpler than the theory of nuclei, etc. In linguistics, for example, the theory of word (“lexicology”) is also not simpler than the theory of sentence (“syntax”). In light of the said the reductionism principle, according to which the complex things should be made of more simple ones, seems to be not obvious and even doubtful. This leads to reviewing the foundations for the standard systems analysis taking into account the effects that can be described by the theory of complexity. At this level the theory acquires the features and status of natural-science, not only mathematical doctrine.

It seems that the concept of state is applied to language units of any level. Since the main feature of language objects is a meaning property, then we can assume that any linguistic unit in the real context is in a certain semantic state, which is an element of the set for its “permissive” states. It would be nice if for the units of a certain level of the language system we could find a simple, transparent and formal mechanism for generating the states like the Schrödinger equation, which generates the states of the quantum-mechanical system. So far we have only solitary examples of constructing the formal mechanisms for generating the states of certain partial language subsystems. In particular, the algorithms for generating the inflective paradigms (for the Ukrainian language they are developed in the Ukrainian Language-Information Fund (Шевченко, 2000)), or the algorithms for generating the lexicographic structures of the ULD (Широков, 1998). The experience we have at the moment let us affirm that the representative of a semantic state of a certain language unit is its description (the relevant dictionary entry text) in the properly constructed lexicographic system. The information on the relevant language unit state is “encrypted” in the dictionary entry text as an element of L-system. The process of abstracting and explicating this information – it is contained in the L-system structure – is “grammatical” by its nature. This adds additional arguments to consider the lexicographic system as an object, which combines both the properties of dictionary and grammar.

Let us continue to say about the concept of calculus. Like the algorithm that sets the *algorithmic*, or *computational*, process (i.e. the process of algorithm work), each calculus sets the *calculated*, or *generating*, process, i.e. the process of calculus work. This process is divided into separate steps (or stages). Each step is getting a new object (state) from the objects (states) that

have been already received before the beginning of this step. Getting a new object is performed by applying a “permissive” rule contained in the calculus. The objects, to which the rule is applied, are called its *premises*. Note that applying the same rule to the same premises may lead to different results. But if you fix the rule and premises, then the number of different results is always finite. For each rule the number of premises is fixed. If all these numbers are limited to a certain number  $n$ , then the calculus is called *n-premised*.

The concepts of *permissible object* or *permissible state* (for this calculus) are the reflection of the intuitive understanding of the objects that are got in the process of calculus. The definition of this concept is inductive.

If  $b$  object is got from  $a_i, \dots, a_k$  by applying one of the “permissive” rules of calculus and if  $a_i, \dots, a_k$  are the permissible objects, then  $b$  is also permissible. This is a step of its inductive definition. The beginning of induction is provided with *zero-premised* rules: if  $b$  satisfies (i.e. is got by applying this rule “from nothing”), then  $b$  is permissible. If there are no zero-premised rules, then a set of objects is empty. In particular, due to zero-premised rules in the logistic calculi the axioms are declared probative.

Any calculus works with the objects of a certain  $W$  group, which is called a run-time environment of the calculus. The work of the calculus is in forming new permissible elements of the run-time environment, or permissible states. The main distinction between the algorithmic and calculated processes is in the following. In the algorithmic process, each state appeared is definitely determined with the preceding process. In the calculated process, the state appeared is only one of many possible that are permitted by the preceding process. If the concept of time to associate with the event alternation (the event is the appearance of new states), we can say that in the algorithmic process the time flows linearly (in this sense the algorithm simulates the physical time that flows in the selected reference system), and in the calculated process the time flows branchily (and its “physical” interpretation is not as transparent).

The history of appearance of a concrete permissible state in the calculated process may be fixed as a single object called an output. We can give it the following spatial interpretation. The *output* of the permissible state  $u$  is a tree, at the tops of which there are certain permissible states and the rules of calculus are in compliance with the tops:

1.  $x$  is in roots;
2. for any  $v$  top of the tree, if  $y$  is a state that is in this top, and  $y_i, \dots, y_k$  are the states that are in those tops, where the ribs follow from  $v$ , then  $y$  is obtained from  $y_i, \dots, y_k$  by the rule that is compared with  $v$  top.

Thus marks on the leaves of the tree are got due to the zero-premised rules. For one-premised calculi the output obtained in this manner are the chains.

Specifying our understanding of calculus, we reach a conclusion that the finite list of “permissive” rules is the most important but only one component of the calculus. It makes the “core” of the calculus like the direct processing operator makes the “core” of the algorithm.

The second component is an instruction for dividing elements into the *basic* and *auxiliary* ones. Let us call this instruction as *a rule for selecting the basic states*. The necessity for this rule is caused by the fact that for various reasons we are interested not in all states, but only in the states of a special type (let us call them basic), while other states are considered only as an auxiliary material for getting the basic states. The role of the rule for selecting the basic states for calculi is similar to that of the signal on getting solution for the algorithm.

Finally the third component of the calculus is *a rule of extracting the result, or the output procedure*. It is similar to the relevant procedure for the algorithm. This procedure transforms each basic state to a certain object. The result of applying the initial procedure to any permissible basic state is called a *result* or *output* of the calculus. About each such object we will also say that it is *generated by the calculus*. About the set of all objects that are generated by the calculus we will also say that it is generated by the calculus. The two calculi are considered equivalent if they generate the same sets of outputs (results). It is considered that a set generated by the calculus is in some *ensemble of outputs* of the examined calculus.



Let us see the example of realization of a certain calculus as a logistic system, otherwise – formal axiomatic theory. Let  $\mathbf{A}$  be the alphabet of the system. We consider that the standard formation rules (construction rules) and the transformation rules (output rules) are known. According to them the variables ( $\mathfrak{c}$ ), terms ( $\mathfrak{T}$ ), formulas ( $\mathfrak{F}$ ) and probative formulas ( $\mathfrak{D}$ ) are selected from all words over  $\mathbf{A}$ . A chart of constructing the logistic calculus is the following. The states in it look like  $\langle a, b \rangle$ , where  $a$  is one of the letters “3”, “T”, “F”, or “D”, and  $b$  is a word from  $\mathbf{A}$ . All these states are put in the relevant run-time environment. The basic states are the states like  $\langle \mathfrak{D}, b \rangle$ . The rule of extracting the result is a transition from  $\langle \mathfrak{D}, b \rangle$  to  $b$ . The permissive rules are got by the obvious modification of the formation and transformation rules mentioned above. Thus, a one-premised rule, which allows the transition from any state like  $\langle \mathfrak{D}, b \rangle$  to the state like  $\langle \mathfrak{T}, b \rangle$  with the same  $b$ , corresponds to the rule, which affirms that any variable is a term. The two-premised rule, which allows the transition from  $\langle \mathfrak{D}, b_i \rangle$  and  $\langle \mathfrak{D}, (b_i \rightarrow b_2) \rangle$  to  $\langle \mathfrak{D}, b_2 \rangle$ , corresponds to the affirmation rule. The zero-premised rules correspond to the axioms: a generated object  $\langle \mathfrak{D}, b \rangle$  can be constructed “from nothing” for any  $b$  axiom. The term superposition rule, which gives a new term  $y$  due by the terms  $t_i, t_2$ , leads to the two-premised rule, which allows the transition from  $\langle \mathfrak{T}, t_i \rangle$  and  $\langle \mathfrak{T}, t_2 \rangle$  to  $\langle \mathfrak{T}, u \rangle$ . If, for example, the substitution rule, which gives a new probative formula  $g$  (which is the result of substituting  $t$  in  $f$  instead of all free entries of  $x$  element) by the probative  $f$  formula,  $x$  variable and  $t$  term, was among the transformation rules, then the three-premised rule should be introduced to our calculus, which allows the transition from the states  $\langle \mathfrak{D}, u \rangle, \langle \mathfrak{D}, x \rangle, \langle \mathfrak{T}, t \rangle$  to the state  $\langle \mathfrak{D}, g \rangle$ .

The last example shows that there is no need of compulsory putting the generated objects to any type, because the information about the object type may be “hidden” inside the state selected properly. So instead of generating the terms and formulas (using terms), we have generated the objects like  $\langle \mathfrak{T}, b \rangle$  and  $\langle \mathfrak{F}, b \rangle$ .

Some of the rules that generate the calculi set various transformations (or they *are* transformations). These are the output rules and the rule of extracting the result. Other rules define the properties (or shorter, they are properties) – for example, the rules for selecting the basic states.

In the first section the concept of lexicographic calculus (the calculus on lexicographic structures) was defined. Its basic concept is a concept of the state, the linguistic content of which was just considered. We now will combine the concept of semantic state with the concept of lexicographic calculus. The idea of this combination is to use the lexicographic structures as a source for parametrization of the semantic states. Let us consider the semantic state representative, introduced with the (2.4)–(2.6) formula, but somewhat modified:

$$(2.32) \quad \psi(x_0^i) = \langle x_0^i; [M(x_0^i) \equiv \tau_1 \xi_1^i \tau_2 \xi_2^i \tau_3 \dots \tau_n \xi_n^i] \rangle,$$

where  $M(x_0^i) \equiv \tau_1 \xi_1^i \tau_2 \xi_2^i \tau_3 \dots \tau_n \xi_n^i$  is the “internal” part of the dictionary entry  $V(x_0^i)$  of the explanatory ULD with a structure defined in the third section. A set of the states defined in this way forms the run-time environment of the calculus. We will consider the text  $M(x_0^i) \equiv \tau_1 \xi_1^i \tau_2 \xi_2^i \tau_3 \dots \tau_n \xi_n^i$  an adequate representative of the complete semantic state of  $x_0^i$  lexeme and that is quite natural. So, the  $M(x_0^i)$  text contains information that allows to select the fragments in it that represent the elementary semantic states, which correspond to the certain grammatical and lexical meanings that in turn allow to present  $(x_0^i)$  in the form:

$$(2.33) \quad \begin{aligned} \psi(x_0^i) &= \mu_1(\psi_1) \psi_1(x_0^i) + \mu_2(\psi_2) \psi_2(x_0^i) + \dots + \mu_n(\psi_n) \psi_n(x_0^i) \equiv \\ &\equiv \sum_k \mu_k(\psi_k) \psi_k(x_0^i) \end{aligned}$$

where  $(x_0^i)$  is a *complete semantic state* of  $x_0^i$  language unit as a fuzzy superposition of the elementary semantic states  $\psi_k(x_0^i)$ . Thus:

$$(2.34) \quad \sum_k \mu_k(\psi_k) = M$$

is a normalization condition, which has a simple linguistic sense as a fuzziness measure in determining the contribution of each elementary semantic state to the complete semantics of  $x_0^i$  language unit. The  $\mu_k(\psi_k)$  numbers represent a fuzziness degree in determining the contribution of each elementary semantic state to the complete semantics of the lexeme; with  $M=1$  the probabilistic

interpretation can be given to them. The statement about the fuzziness follows from the idea that the lexeme complete semantics, a set of basic semantic states and their content represent the view of lexicographers of the basic explanatory lexicographic system (BELS). And we do not guarantee that all meanings of  $x_0^i$  language unit are included and that it is done in perfect way.

Further semantic structures can be introduced and calculated by using various types of  $A$  operators from the set of automorphisms of the BELS lexicographic system. These can be the following operators, the action of which is defined on the BELS lexicographic structure:

- the relations of semonymy (synonyms, antonyms, paronyms, homonyms)
- the relations of word formation (the related word)
- the relations of thesaurus: [“genus-species”, “part-whole”, “complex-element”, “cause-effect”]
- the relations of associations and analogies (associators and analogems).

From the states like (2.32) we can select the basic and auxiliary ones by introducing an instruction (or instructions), which represents an expediency conditioned linguistically. Moreover, the  $\tau_i$  metalanguage elements can be involved in the defining the states. Then the definition of lexicographic calculus can be expanded and generalized by introducing the new objects and states. For example, the objects may be not only separate “lexemes” – the  $I_0^W(D)$  elements, but also other elements of the relevant lexicographic system (for example, “interpretation formulas”, their constituents, the grammatical description elements, illustrations, remarks, etc.). The output rules may serve here as algorithms of testing the elements of the sign system, structure, the correct arrangement of dictionary entries, etc. The example of the lexicographic calculus oriented on the automatic modification and replenishment of the explanatory dictionary by its interaction with the “internal environment” (that is a text array), will be set on the descriptive level in the next paragraph.

## Bibliography

- Berry, G. D. W. (1953)** Symposium on the Ontological Significance of the Löwenheim-Skolem Theorem, Academic Freedom, Logic, and Religion. Philadelphia, PA: Amer. Philos. Soc., pp. 39-55.
- Chang, C. C. and Keisler, H. J (1990)** Model Theory, 3rd enl. ed. New York: Elsevier.
- Борисов А.Н. (1987)** Применение лингвистической переменной в системах принятия решений. – М., 340 с.
- Гейзенберг В. (1989)** Физика и философия. Часть и целое. Пер. с нем. – М.: Наука. – С.191–196.
- Заде Л. (1976)** Понятие лингвистической переменной и его применение к принятию приближенных решений. – М., 168 с.
- Рабулець О. Г. (2002)** Інтегровани лексикографічні системи: Автореф. дис. ... канд. техн. наук. – К. – 18 с.
- Словник (1970–1980)** Словник української мови в 11 томах. – К.: Наукова думка.
- Шевченко. И. В. (2000)** Модели та алгоритмічно-програмне забезпечення лексикографічних систем: Автореф. дис. ... канд. техн. наук. – К.
- Широков В. (1998)** А. Информационная теория лексикографических систем. – К.: Довира, 331 с.
- Широков В. А. (2004)** Феноменология лексикографических систем. – К.: Наукова думка, 328 с.

# Towards Semantic Concordances in Slovene

Darja Fišer<sup>1</sup>, Tomaž Erjavec<sup>2</sup>

<sup>1</sup> Department of Translation, Faculty of Arts, University of Ljubljana; Aškerčeva 2, Ljubljana, Slovenia  
`darja.fiser@guest.arnes.si`

<sup>2</sup> Department of Knowledge Technologies, Jožef Stefan Institute; Jamova cesta 39, Ljubljana, Slovenia  
`tomaz.erjavec@ijs.si`

**Abstract.** The paper presents the annotation of a Slovene language corpus with semantic information. Manual annotation was performed with an automatically generated wordnet for Slovene by two annotators. The cases in which they disagreed were consolidated by the third annotator acting as a referee. The analysis of the results shows that practically all polysemous can be assign a sense from wordnet but also that the task was quite challenging; in many cases, wordnet sense distinctions are too fine-grained even for human annotators to distinguish between them. This is why annotation with more coarse-grained senses could prove to be more successful.

## 1 Introduction

Two very different types of linguistic resources, textual corpora and lexical resources, can be interrelated and enhanced through *semantic concordances*, in which words from the corpus are connected with their meanings specified in a semantic lexicon. Semantic concordances are an extremely useful resource for a wide range of applications, such as automatic word sense disambiguation or for corpus-based studies of sense frequency, distribution and co-occurrence, and are also invaluable as an aid for translation as well as for vocabulary acquisition in a foreign language.

The topic of this paper is a project in which frequent nouns from a corpus of Slovene were manually annotated with wordnet senses. Polysemous nouns were extracted from wordnet and identified in the corpus. Then each occurrence of the target word in the corpus was assigned one of the wordnet senses of the target word to according to the context in which the word occurred. The result of the annotation process is a list of concordances in which each nucleus word has an assigned sense. If required, additional information about this sense of the word, such as its definition, synonyms and other related words, can be directly retrieved from wordnet. On the other hand, the annotated corpus can be seen as a companion resource to the lexicon, providing examples for and relative frequencies of word senses, additional semantic relations etc.

The paper is structured as follows: Section 2 discusses related work; Section 3 introduces the resources used in this project, namely the jos100k corpus and sloWNet; Section 4 details the annotation process; Section 5 gives an analysis of the corpus annotations; and Section 6 gives the conclusions and directions for further work.

## 2 Related work

In the past years, multi-level annotation of corpora has become common practice in order to turn them into even more useful resources for increasingly complex HLT tasks. A critical element in corpus annotation at any level is ambiguity resolution. While morpho-syntactic ambiguities can nowadays be tackled by PoS taggers and shallow parsers for many languages, word-sense disambiguation has not reached the same level of maturity (Resnik and Yarowsky 1997). Most approaches are still manual or semi-automatic and semantically annotated corpora for languages other than English have only started to emerge recently.

There are two main paradigms for semantic annotation of corpora. The first one is the labeling of semantic roles and predicate-argument structures (Baker, Fillmore and Lowe 1998) which are

used in tasks such as information extraction and question answering (Surdeanu, et al. 2003). An example of a corpus with semantic role annotations is PropBank (Palmer, Gildea and Kingsbury 2005). In the project described in this paper we follow the second paradigm, which is the annotation of polysemous words with one of its senses (Landes, Leacock and Tengi 1998).

This type of resources, such as SemCor<sup>3</sup> and MultiSemCor,<sup>4</sup> have mostly been developed within the Senseval initiative and are used for automatic word sense disambiguation and machine translation (Kilgarriff 1998).

### 3 Resources used

This section presents the two resources used in the project, namely the jos100k reference corpus of Slovene and the wordnet for Slovene, called sloWNet.

#### 3.1 The jos100k corpus

The jos100k corpus has been developed within the JOS project<sup>5</sup> that is developing annotated corpora and associated resources meant to facilitate developments in human language technologies for the Slovene language. At present, the JOS resources comprise morpho-syntactic specifications, two word-level annotated corpora, and two web services. The developed resources are available under the Creative Commons licenses.

The jos100k corpus (Erjavec and Krek, 2008) is a 100,000 word Slovene corpus containing sampled paragraphs from the Slovene reference corpus FidaPLUS.<sup>6</sup> The corpus is annotated with manually validated morphosyntactic descriptions and lemmas. The corpus has been carefully composed and checked and is meant to serve as a gold-standard reference corpus. In the scope of the JOS project we are annotating it for syntactic structures, and for lexico-semantic information, which is the topic of this paper.

#### 3.2 sloWNet

sloWNet<sup>7</sup> is a lexico-semantic resource for Slovene, in which words that have the same meaning (literals) are organized into sets of synonyms (synsets). Synsets are linked into a semantic network with various lexical and semantic relations. sloWNet was built semi-automatically from Princeton Wordnet (Fellbaum 1998) and is aligned to all wordnets for other languages that use Princeton WordNet ids for concept representation. The creation process consisted of three stages (Fišer and Sagot 2008):

1. Core wordnet: A bilingual dictionary was used to translate basic concepts into Slovene. The translations were then checked and corrected by hand.
2. Polysemous words: Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages.
3. Monosemous words: Equivalent terms for monosemous words were found in open-source resources, such as Wikipedia and Eurovoc thesaurus.

The latest version of sloWNet (2.0, August 2008) contains about 20,000 unique literals which are organized into almost 17,000 synsets. It is rich in basic concepts as well as specific ones. The former were mostly obtained from the dictionary and a parallel corpus while the latter come from Wikipedia. sloWNet mostly contains nominal synsets, although there are some verbal and

<sup>3</sup> <http://multisemcor.itc.it/semcor.php>

<sup>4</sup> <http://multisemcor.itc.it/>

<sup>5</sup> <http://nl.ijs.si/jos/>

<sup>6</sup> <http://www.fidaplus.net/>

<sup>7</sup> <http://nl.ijs.si/sloWnet/>

adjectival synsets as well. Apart from single word literals, there are also plenty of multi-word expressions. The most common relation in sloWNet is hypernymy which represents almost half of all relations in wordnet. A comparison of nouns in sloWNet and the jos100k corpus showed that sloWNet nouns cover 30% of the nouns present in jos100k, with 90% coverage of the most frequent nouns (Fišer and Erjavec 2008).

## 4 The annotation process

The main goal of our annotation process was to obtain the first semantically annotated corpus for Slovene which can be used in corpus-based linguistic research as well as a resource for HLT applications requiring training data. However, because sloWNet had been created automatically and had been based on a foreign-language resource, our secondary goal was to check coverage of the senses it contains compared to the senses represented in the corpus and thereby evaluate the developed lexicon in a practical semantic tasks and to improve it.

Because no application for automatic sense assignment exists for Slovene, the annotation had to be done completely manually. As opposed to sequential annotation, in which all the words in the corpus are annotated, we followed the *targeted semantic annotation principle* (Miller, et al. 1994) which aims at determining senses only for a selection of polysemous corpus words. The main reasons for choosing this method was limited project resources and because the results are directly applicable to automatic word-sense disambiguation tasks.

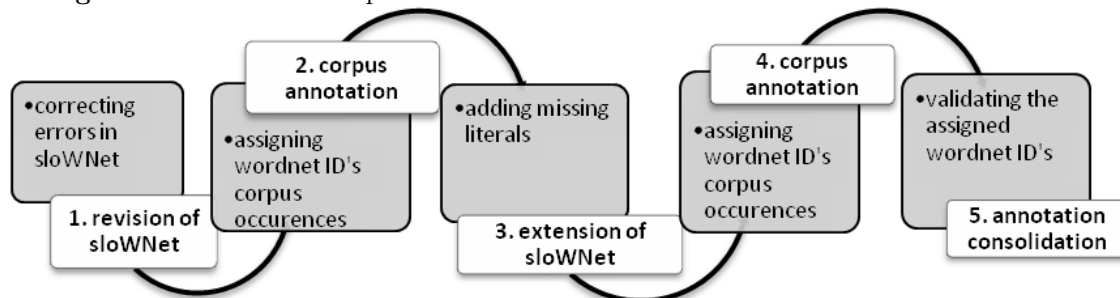
Targeted or *transversal annotation* is preferred by many researchers (see Kilgarriff 1998) because this way the semantic characteristics of each word are taken into consideration only once, and the whole corpus achieves greater consistency. In sequential or *linear annotation*, the annotator has to remember the sense structure of each word each time the word appears in the corpus, making the annotation process much more complex, thus further increasing the possibilities of low consistency and disagreement between the annotators (Navarro, et al. 2003).

In addition, we followed the *joint approach* of coordinated wordnet validation, refinement and corpus annotation as proposed by Agirre et al. (2006) because it ensures that word senses in the lexicon reflect real usage and guarantees a better fit between sense distinctions in the lexicon and the corpus, which will improve subsequent automatic word sense disambiguation.

In order to ensure more reliable annotations, the same concordances were annotated by two different annotators, after which a third annotator, acting as a referee, chose the most suitable annotation in case each of the annotators suggested a different sense.

As Figure 1 shows, the annotation procedure consisted of several stages: first, annotators started from sloWNet in which they checked all senses of a given word and correct any errors they found. This stage was necessary because sloWNet had been built automatically and had not been fully checked by hand, which is why errors in synsets were possible. In the second step, the annotators turned their attention to the concordances and tried to assign a wordnet sense to each occurrence of the given word in the corpus. If they came across a meaning of a word or a phrase they could not find in sloWNet, they added it to the wordnet. In the end, the annotations were consolidated and validated by the referee.

**Figure 1.** The annotation procedure



#### 4.1 Selection of the words to be annotated

In the first attempt of semantically annotating Slovene, we limited the task to nouns only because sense assignment for nouns is the easiest and because they are currently best covered in sloWNet. We extracted all the common nouns that exist in sloWNet and appear in the jos100k corpus with a frequency of 30 or higher. There were 103 such nouns, most of which belong to the Basic Concept Sets in wordnet. The most frequent nouns in the corpus are *leto* (Eng. *year*, freq. 348), *dan* (Eng. *day*, freq. 151) and *delo* (Eng. *work*, freq. 145). 87.4% of the extracted nouns have more than one sense in sloWNet. While the most have three senses (17.5%), the most polysemous ones are *vrsta* (Eng. *type*, 14 senses), *stvar* (Eng. *thing*, 13 senses) and *mesto* (Eng. *place*, 12 senses). Once the annotation is completed, this will yield a total of 5,592 tokens with a manually assigned sense, which means that on average there will be about 54 annotation examples for each noun included in the annotation process.

It is expected that the complexity of sense assignment to the target nouns will correspond to their level of polysemy in sloWNet. On the other hand, it is highly likely that the lexicon is still missing some senses for nouns which are frequent in the corpus but have very few senses or are even monosemous in the current version of sloWNet, which is why these nouns need to be carefully examined as well (e.g. *člen*, freq. 57 appears in sloWNet only in the sense of Eng. *link* but not in the sense of *article* in a legal document or the grammatical *article*).

Concordances for all occurrences of these 103 nouns were extracted from the jos100k corpus. If an occurrence of a target word (e.g. *delavec*, Eng. *worker*) belonged to a multi-word expression that exists in sloWNet (e.g. *kvalificiran delavec*, Eng. *skilled worker*), it was not extracted because multi-word expressions are typically monosemous and therefore do not require manual sense assignment. If an occurrence of a target word belongs to a multi-word expression which does not yet exist in sloWNet, it was extracted and will be annotated as part of a multi-word expression and the expression will be added to sloWNet by the annotator.

**Figure 2.** Revision of synsets in DEBVisDic with highlighted editing features

The screenshot shows the DEBVisDic Slovene Wordnet interface. At the top, there is a search bar with the text 'abeceda' and buttons for 'Search', 'Search in all', and 'User query'. Below the search bar, a list of synsets is displayed, with '[n] Grška abeceda:1, grški alfabet:1' highlighted. A toolbar contains buttons for 'Preview', 'Tree', 'Revtree', 'Edit', 'Query', and 'Xml'. Below the toolbar, there are buttons for 'New', 'Delete', 'Save', and 'Clear form'. The 'Part of Speech' section shows a dropdown menu with 'n' selected and a checkbox for 'Not Lexicalized'. The 'Synonyms: Literal, Sense, LNote' section contains a table with two rows of synonyms, each with a 'Remove' button. The 'Definition' section has a text input field containing the definition 'abeceda, ki jo uporablja grška pisava'.

Synonym	Sense	LNote	Action
Grška abeceda	1	0/1:enwikipedia	Remove
grški alfabet	1	0/1:enwikipedia	Remove

## 4.2 Annotation guidelines

In order to facilitate the annotation process and to ensure a greater consistency of annotations, annotation guidelines have been provided for the annotators. The annotators' first task was to revise and validate all the synsets, including all multi-word expressions, containing the target word. Wordnet revision was carried out in a multi-lingual wordnet editor called DEBVisDic (Horak, et al. 2005), as illustrated in Figure 2. If an error was found (e.g. incorrect capitalization), it was corrected at this stage. In case a literal was found in an inappropriate synset, it was deleted, and if a literal was missing in the synset, it was added to sloWNet together with a source confirming the appropriate sense and usage of the word (e.g. dictionary or corpus). Wordnet revision also entailed making sure that all the hypernyms of the target word exist in sloWNet. If a hypernym synset was empty, the annotator translated it from English at this stage.

After all the senses of the target word had been validated, the annotation of the corpus began. Because no tailor-made annotation software was available, the annotation was performed in MS Excel. Annotators received xls files with the concordances containing the target word that were extracted from the jos100k corpus. After reading an occurrence of the target word in context they determined which synset was the most appropriate for it, they annotated it with the corresponding synset id from wordnet (see Figure 3).

**Figure 3.** Annotation of the corpus in MS Excel

A	C	D	E	F
n	<i>pomen</i>	<i>Opomba</i>	<i>levi kontekst</i>	<i>beseda</i> <i>desni kontekst</i>
13			poto Triglava , je za	knjigo Triglav , Sveta gora Slo
14			i Založbi Mladinska	knjiga , izbral France Stele . F
15	ENG20-06013091-n		to iz avtorjeve nove	knjige Imena nebesnih teles .
16			stu 1986 izšla njena	knjiga How Institutions Think t

The goal of the annotation was to attempt to assign a sense to all occurrences of the target words. If more than one sense seemed appropriate, the annotators were directed to choose the most basic sense. Only if they really could not choose a single sense, were they allowed to annotate a word with multiple synset ids. If an occurrence of the target word belonged to a multi-word expression, it was annotated with that sense and a note #MWE was added. In case the target word was (part of) a proper name that does not exist in wordnet, a note !PROPER was added. If the target word in the corpus was used in the sense that was missing in sloWNet but could be found in PWN, it was added to sloWNet and annotated with the newly added sense. If the appropriate sense could not be found in either sloWNet or PWN, the word was left unannotated and a note !NO was added. It is likely that these senses are language-specific and should therefore be added as such to sloWNet at a later stage of wordnet development.

The files with annotations were then uploaded and analyzed through a web service which reported any structural errors in annotations.

## 5 Analysis of annotations

The annotation of the corpus is almost complete but some words are still missing, which is why the figures included in the analysis below are only partial and will change by the end of the project. The analysis was conducted on 77 which have been submitted so far.

### 5.1 The extent of wordnet revision

The analysis shows that a total of 852 synsets were changed in the revision process. A great majority (80.5%) were changed by a single annotator while only 19.5% of the synsets were changed by both annotators. Just over a half of these synsets were modified (54.8%) and the rest (45.2%) were added to sloWNet by the annotators.

At the level of wordnet literals, sloWNet originally contained 649 literals for the 77 target nouns included in the annotation process. Many new (1044) were added and only a few of the literals (128) that had been automatically generated in sloWNet were deleted, so that at the end of wordnet revision, we are left with 1553 target literals. Wordnet contained relatively few (181)

multi-word expressions with one of the target nouns before revision, only 47 of which were deleted whereas many more (557) were added by the annotators.

Few deletions of both single- as well as multi-word literals suggest that the precision of the automatically generated wordnet is very good. On the other hand, many literals and synsets were added at this stage, implying that the generated wordnet had a low recall and that many senses of the words processed in the project had been missing from sloWNet. A substantial share of the added synsets contain multi-word expressions, which could not be added automatically due to the limitations of wordnet generation method.

## 5.2 Comparison of annotations

In this section we turn to comparing the annotations of the corpus made by the annotators. The total number of tokens that have been annotated so far is 3520, which means there are 45.7 annotated tokens per noun on average. Because all the words were annotated by two different annotators, we have 7040 annotations in total, less than one per cent of which are ambiguous i.e. are annotated with more than one synset id. Annotators added a note to slightly over 12% of their annotations, which means that these occurrences belong either to a multi-word expression or proper name, or that a satisfactory sense for them could not be found in wordnet, and should be therefore re-examined and resolved at a later stage in the project.

Occurrences of the 77 already annotated nouns were assigned 389 different senses, 93% of which were only used once. These figures include many of the nouns that were treated as multi-word expressions by the annotators and therefore annotated with a greater number of different synset ids.

181 or 46.5% of the synsets containing the target nouns were not used by either annotator. There is a good reason for not using many of these synsets because the target nouns appeared in them only due to insufficient disambiguation during wordnet generation and were deleted by the annotators at the wordnet revision stage. An example is the word *sodišče* which appears in some synsets because the English word *court* was wrongly translated into Slovene in three synsets:

1. *a yard wholly or partly surrounded by walls or buildings* – the correct translation is *dvorišče*,
2. *the sovereign and his advisers who are the governing power of a state* – the correct translation is *dvor* and
3. *the family and retinue of a sovereign or prince* – the correct translation is *dvor*.

Other senses were not used because they did not appear in the corpus. However, they should not automatically be treated as irrelevant for Slovene because the 100.000 token corpus that was used is far too small for such conclusions and it would do more harm than good if such senses were deleted from sloWNet at this stage. One such example is the noun *stran* (Eng. *page*) which has seven senses in sloWNet, four of which do not appear in the corpus not because they are not used in Slovene at all but because they simply did not appear in our corpus:

1. *an extended outer surface of an object*,
2. *a distinct feature or element in a problem*,
3. *a sheet of any written or printed material (especially in a manuscript or book)* and
4. *one side of one leaf (of a book or magazine or newspaper or letter etc. or the written or pictorial matter it contains).*

The noun *šola* (Eng. *school*) received the highest number of new senses by the annotators. The noun initially had three senses in sloWNet, and four more were added by the annotators: one sense was added because it was missing (*an educational institution*) and the other three were part of multi-word expressions that were identified in the corpus (*glasbena šola*, Eng. *music school*, *osnovna šola*, Eng. *primary school* and *srednja šola*, Eng. *secondary school*).

On average, 4.7 senses were used for each noun and the most frequent number of senses for a noun (20.1%) is 5. This is slightly higher than the most frequent number of senses of the nouns to be annotated before the revision of sloWNet, which was 3, but because many senses were added



during the annotation process, the figures are still comparable. A single sense was assigned to all the occurrences of only two target nouns, while the two most polysemous nouns were assigned 13 different senses.

A comparison of annotations for the same target word that were submitted by two different annotators shows that their annotations vary to a great extent: they chose the same synset id for only 1848 or 52.5% of the annotated tokens. It has also been observed that target words differ substantially in the level of agreement between the annotators, which means that some words were much easier to annotate than others. Perfect agreement is reached only with the words that were assigned only one sense (e.g. *odstotek*, Eng. *percentage*), but words such as *člen* (Eng. *article*) and *oče* (Eng. *father*) have an agreement exceeding 95% as well. This is not very surprising because the number of initial as well as the number of senses used for the words with a high inter-annotator agreement was rather low (3 or 4) and much higher (11 or 12) for those with a low agreement. Also, the level of agreement decreases with the increase of target word frequency in the corpus. This confirms our initial hypothesis.

Next, we checked whether annotators agreed on the most frequent sense for a given word despite the relatively low inter-annotator agreement. The most predominant sense is very useful for many HLT applications because it has been found that the predominant sense baseline is quite hard to beat by word sense disambiguation algorithms. It turns out that the distribution of senses of the annotated words are in favour of the predominant sense, and that non-predominant senses chosen are in the minority. Also, annotators agreed on the most frequent sense in most cases.

One of the words in which the annotators disagreed even on the most frequent sense is *predstavnik* (Eng. *representative*) for which the share of the most frequent sense is similar (56.7% and 46.7%) with both annotators but the synsets they used to annotate the most occurrences of this noun in the corpus are different. One annotator most frequently chose the synset *agent: a representative who acts on behalf of other persons or organizations* while the other one preferred the synset *representative: a person who represents others* most of the time. When we study both synsets in detail, we find that they are both very similar and it is indeed hard to distinguish between them. This shows that sense distinctions in wordnet are not very clear-cut and are very fine-grained, which is a common criticism of the resource as a sense repository for practical applications.

## 6 Conclusion

Based on the results of the analysis we conclude that it is possible to annotate practically all occurrences of polysemous words in the corpus with sloWNet synsets. The most problematic words to annotate were culturally-specific expressions that are not included in the Princeton WordNet which was used as the backbone of the Slovene wordnet. A very positive finding is that most senses that were required to annotate the corpus had already been present in sloWNet whereas the same is not true for non-core senses and especially for multi-word expressions which had to be added by the annotators in many cases. This suggests that sloWNet will have to be further extended in order to ensure a thorough coverage of the sense inventory relevant for Slovene.

Semantic annotation of a corpus, be it manual or automatic, is still one of the challenging tasks that need to be done. It is very different from e.g. morpho-syntactic annotation in which all the units are annotated with the same set of categories, whereas in determining the meaning of a word, different categories have to be used for each unit we wish to annotate. This is why inter-annotator agreement are typically lower for semantic annotation than other annotation tasks. An experiment conducted within the Senseval initiative, in which a French corpus was annotated with senses from a French dictionary that contain fewer sense distinctions than wordnet, reports 75% agreement (Veronis 1998). That annotating with wordnet is harder is shown by a similar experiment in annotating English sentences with Princeton WordNet senses which shows a substantial drop in agreement, which reaches only 68% (Mihalcea, Chklovski and Kilgarrieff 2004). Our results (52.5%) are still significantly lower than that which might be due to two factors. First, we annotated only the most frequent nouns in the corpus, which are also the most highly polysemous ones and therefore harder to disambiguate. Second, due to project constraints, we used 50 undergraduate

students as annotators, where their large number also leads to lower consistency and agreement in annotations. This shortcoming is already being compensated by a third annotation cycle in which an experienced team of 4 annotators are checking and consolidating the differences between the original two annotators. Their work will hopefully provide much more reliable and consistent data that will be more useful in further research.

One way of simplifying and improving the annotation process in the future is collapsing fine-grained hard-to-distinguish senses into more general categories, called supersenses. This had already been done manually by Palmer, Dand and Fellbaum (2007) and automatically by Bruce and Wiebe (1998) who achieved a 10% improvement on the results.

Notwithstanding the difficulties of the annotation, the result of this project is the first Slovene corpus that is annotated at the semantic level. The corpus will be freely available for linguistic analysis or as a training set for applications in human language technologies.

## Bibliography

- Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K. & Quintian, M. (2006).** A methodology for the joint development of the Basque WordNet and Semcor. *Proceedings of LREC'06*. Genoa.
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998).** The Berkeley FrameNet Project. *Proceedings of ACL'98*, (pp. 86–90). Montreal.
- Bentivogli, L., Forner, P., & Pianta, E. (2004).** Evaluating cross-language annotation transfer in the MultiSemCorpus. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva.
- Bruce, R., & Wiebe, J. M. (1998).** Word sense distinguishability and inter-coder agreement. *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing* (pp. 53–60). Granada.
- Erjavec, T., & Krek, S. (2008).** The JOS Morphosyntactically Tagged Corpus of Slovene. *Proceedings of LREC'08*. Marrakech.
- Fellbaum, C. (2002).** On the Semantics of Troponymy. In R. Green, C. Bean, & S. Myaeng (Eds.), *Relations*. Dordrecht: Kluwer.
- Fellbaum, C. (Ed.). (1998).** *WordNet: An Electronic Lexical Database*. Cambridge, London: MIT.
- Fišer, D., & Erjavec, T. (2008).** Predstavitev in analiza slovenskega wordneta. *Proceedings of IS-LTC'08* (pp. 37–42). Ljubljana.
- Fišer, D., & Sagot, B. (2008).** Combining Multiple Resources to Build Reliable Wordnets. *Proceedings of TSD'08*. Brno.
- Horak, A., Pala, K., Rambousek, A., & Povolny, M. (2005).** DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool. *Proceedings of the GWA'05* (pp. 325–328). Brno.
- Kilgarriff, A. (1998).** Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language. Special Use on Evaluation*, 12 (4), 453–472.
- Kilgarriff, A. (1998).** SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. *Proceedings of LREC'98*. Granada.
- Landes, S., Leacock, C., & Teng, R. I. (1998).** Building Semantic Concordances. In C. Fellbaum, *WordNet* (pp. 199–216). Cambridge, Massachusetts: MIT Press.
- Mihalcea, R., Chklovski, T., & Kilgarriff, A. (2004).** The Senseval-3 English lexical sample task. *Proceedings of ACL/SIGLEX Senseval-3*.
- Miller, G. A., Chodorow, M., Landes, S., Leacock, C., & Thomas, R. G. (1994).** Using a semantic concordance for sense identification. *Proceedings of the workshop on Human Language Technology*. Plainsboro, NJ.
- Navarro, B., Civit, M., Martí, M., Marcos, R., & Fernández, B. (2003).** Syntactic, Semantic and Pragmatic Annotation in Cast3LB. *Proceedings of CL'03, Workshop on Shallow Processing of Large Corpora*. Lancaster.

- Palmer, M., Dand, H. T., & Fellbaum, C. (2007).** Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* (13), 137–163.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005).** The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31 (1), 71–105.
- Resnik, P., & Yarowsky, D. (1997).** A perspective on word sense disambiguation methods and their evaluation. *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* (pp. 79-86). Washington, DC.
- Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003).** Using Predicate-Argument Structures for Information Extraction. *Proceedings of ACL'03*. Sapporo.
- Veronis, J. (1998).** A study of polysemy judgements and inter-annotator agreement. *Programme and advanced papers of the Senseval workshop*. Herstmonceux Castle.

# A Syntactic-Semantic Treebank for Domain Ontology Creation

Kiril Simov, Petya Osenova

Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria  
kivs@bultreebank.org, petya@bultreebank.org

**Abstract.** This paper presents a methodology for the creation of a domain ontology which exploits the idea of a manually created treebank as a mechanism for extraction of an initial set of domain concepts and relations. The creation of a treebank is an expensive task. In order to justify this effort, the set of text for the treebank was selected in such a way that it increased the number of the extracted concepts and relations. Also, we wanted the extracted information to be reliable. To satisfy this requirement, industrial standards in the domain of interest as well as a specialised vocabulary were selected for processing.

## 1 Introduction

Domain ontologies are crucial element in many applications in information technology, ranging from semantic annotation and semantic search to semantic integration of information. Most of the methodologies for creation of domain ontologies include participation of domain experts in their life cycle as a source of conceptual information. They provide information for the central concepts in the domain and the relations between these concepts. The process requires interaction between the domain experts and knowledge engineers. In order to be fruitful, the interaction needs some common understanding. For that reason, the domain experts are usually asked to fill questionnaires about the domain from which the conceptual information is extracted. However, this process is expensive and slow because the questions to be answered are not natural for them in many ways. As solutions to this problem, the methodologies developed in the direction of analysis of domain text. The aim is the extraction of an initial set of concepts and relations, or the creation of technology for extraction of the conceptual information from the results of the everyday work of these experts, such as models, documents, etc. The main problems in these new approaches are the identification of the conceptual information and its reliability. In our work we applied our experience from the treebank construction to an artifact from the work of the domain experts — namely, the standards and terminological lexicons in the domain.

First we present an overview of the methodologies for building of ontologies. Then we select the methodology we applied to construct AsIsKnown Home Textile Ontology and we discuss our choices. There we also discuss the choice of an upper ontology which was employed during the ontology construction. Then we present the processing of the standards in home textile domain which resulted in the identification of the concepts and the relations included in the first version of the ontology.

## 2 Methodology for Ontology Construction

### 2.1 Overview of the Ontology Construction Methodologies

Several surveys on ontology construction methodology already exist and this is why we only report of their findings here. Then in the next section we present our choices for the different stages of the construction of the ontology. The methodologies for creation of ontologies described in the literature have two sides – organizational and technological. The first one considers the steps in the creation of ontology and the second – the tools necessary for realization of each step. Here we focus mainly on the organizational side of the methodology. The overview in this section is based

on: [3], [11], [7], [4], [5]. First we outline some of the most popular methodologies for ontology development, and then we conclude this section with general points about the steps in ontology creation.

#### *Uschold and King Methodology*

This methodology was published first in [11]. They identify the following stages of the process of ontology creation: (1) Identify Purpose; (2) Building the Ontology; (3) Evaluation; (4) Documentation. The building of the ontology (step 2) is divided into three steps: (2.1) Ontology capture; (2.2) Ontology coding; (2.3) Integrating Existing Ontologies.

The purpose identification is important in order to clarify why the ontology is built and what its intended uses are. Some purposes found in literature are that the purpose of the ontology is to be a shared vocabulary for a domain, a meta-level specification of a logical theory, a way of structuring of a knowledge base. Identification of the purpose of an ontology is in a clear relation to the scope of the ontology – which concepts and relations in the domain to be formalized, on what level of granularity this formalization to be done. One important requirement mentioned in the paper is the reuse of the ontology within the group developing it, but also in a broader context. In our view the considered purposes are not mutually incompatible and thus when the ontology is developed the group developing it could try to achieve as many as possible of them.

As it was mentioned, the actual building of the ontology comprises three steps: (2.1) Ontology capture; (2.2) Ontology coding; (2.3) Integrating Existing Ontologies. The ontology capture means: (1) identification of the key concepts and relationships in the domain of interest; (2) production of precise unambiguous text definitions for such concepts and relationships; (3) identification of terms to refer to such concepts and relationships; and (4) agreeing on all of the results from previous steps. Important question with respect to ontology capturing is how the key concepts and relations are identified. According to [11] there are three approaches: top-down – the most general concepts and relations are selected first and then they are specialised to the necessary level of domain dependence; bottom-up – first, the most specific concepts and relations in the domain are represented and then generalizations over them are defined; middle-out – the most fundamental concepts and relations in the domain are selected, and later their specializations and generalizations are added. The ontology coding requires formalization of the selected concepts and relations in a formal language. Also the process requires encoding of the axioms that the concepts and relations have to satisfy. In our view, at this stage, the following question needs to be answered: what the trade-off between the expressivity of the formal language for the ontology and the inference in this formal language will be. Sophisticated ontologies require encoding in very expressive languages like KIF, FOL, Modal Logic, but these languages are not fully supported by ontological tools. The implemented languages like Description Logics (OWL-DL, for instance) allow only a part of the necessary axioms to be represented. This question is important, because its solution could lead to a language dependant ontology which is not desirable. One option here is to have two versions of the ontology – one represented in an expressive language and one derived from the first version in an implemented language. Such an approach was applied by the developers of the foundational ontology DOLCE – see [9]. The integrating existing ontologies is the basis of the reuse of existing ontologies. The task requires special attention with respect to the effort necessary to do such an integration. It depends on the quality of the existing ontology, the documentation, whether it corresponds to the purposes of the new ontology. On the other hand it is very appealing with respect to the potential benefits.

The evaluation and the documentation stages are not described in great detail. The first is based on a definition given in [6]: “to make a technical judgement of the ontologies their associated software environment and documentation with respect to a frame of reference ... The frame of reference may be requirements specifications competency questions and/or the real world.” The documentation stage needs established guidelines for documenting ontologies.

#### *Grueninger and Fox Methodology*

The methodology was developed under the TOVE (Toronto Virtual Enterprise) project – see [7]. It comprises the following stages of ontology development: (1) motivating scenarios; (2) informal competency questions; (3) terminology specification; (4) formal competency questions; (5) axiom specification; and (6) completeness theorems.

The motivating scenarios are stories about the different usages of the ontology in a given application. They highlight problems encountered in an application and their possible solutions. The solutions provide an informal intended semantics for the concepts and the relations to be included in the ontology. The informal competency questions arise from the motivating scenarios and place requirements of the ontology, based on the motivating scenarios. The ontology must be able to provide answers to each of these informal questions. These questions act as an evaluation on the ontological commitments made in the previous stage. The questions have to reflect the specificity of the concepts and relations to be encoded in the ontology. Thus, they can be used when one of already existing ontology is incorporated in the new one. The transferred knowledge has to give answers to these questions. The terminology specification comprises two steps - determination of the terms used in the informal competency questions; and encoding of these terms in a formal language. The formal competency questions are formalised requirements on the ontology. They are based on the informal competency questions. The axiom specification encodes the axioms that specify the definition of terms and constraints on their interpretations. They are given in first-order logic, guided by the formal competency questions as the axioms must be necessary and sufficient to express the competency questions and their solutions. The completeness theorems define the conditions under which the solutions to the competency questions are complete.

#### *Methodology Methodology*

The methodology is described in [4] and [5]. It presents the construction of ontology on knowledge level identifying the following activities: (1) specification; (2) knowledge acquisition; (3) conceptualisation; (4) integration; (5) implementation; (6) evaluation; (7) documentation.

The specification defines the purpose of the ontology, including the intended users, scenarios of use, the degree of formality required, etc., and the scope of the ontology including the set of terms to be represented, their characteristics and the required granularity. The knowledge acquisition acquires knowledge about the domain of the ontology. Many different knowledge sources are analysed in order to achieve the task. The two main sources are expert interviews and analyses of domain texts. The conceptualisation structures the domain terms as concepts, relations, properties and instances. The integration of ontologies is required when ontologies or definitions from other ontologies should be incorporated in the new one. During the implementation the ontology is formally represented in an ontological language (KIF, for example). The evaluation stage comprises checks for incompleteness, inconsistency and redundancy cases in the ontology. The documentation reflects the results from the previous activities in natural language.

The methodology accepts that the life cycle of ontology is based on the refinement of a prototype. The ontology goes through the following states: specification, conceptualisation, formalisation, integration, and implementation. Knowledge acquisition, evaluation and documentation are carried during the entire life cycle.

On the basis of the overview we do the following conclusions:

- The life cycle of ontology is similar to this of other software products: design phase, prototyping, implementation, exploitation, support, documentation.
- There is a difference between a domain and an application in the domain. The ontology has to reflect the domain in as much as possible application independent way, but satisfying the needs of the application.
- Ontology development is an iterative process. Usually the iteration is from less expressive to more expressive version and from informal to formal representation.
- An evaluation of the sources of information is essential to the development of ontology. The evaluation has to be done with respect to the availability of the source, the effort to use the source and how reliable are these sources.

All these conclusions are taken into account during the definition of our methodology for the creation of the commonsense ontology in the domain of the home textile.

## **2.2 AsIsKnown Methodology**

In this section we define the methodology for creation of the commonsense ontology in the domain of the home textile within the AsIsKnown project. The resulting ontology needs to cover the

domain of the home textile in details necessary for representation of the product, product sets, trends, designs and user profiles. Also the ontology has to help the interaction to the external world in terms of formal translation of producer's and ordering data, human language interaction with human users of the system and analysis of multimedia documents. Having in mind the differences in the usage of the ontology it follows that different views over the domain have to be incorporated in the ontology. First the customers view the products with respect to their usage in interior designs. Thus features like size, type of materials, colours, etc are promoted, while other features like flammability, resistance to chemical agents, etc are demoted in the background. For the people responsible for the safety issues of the interior designs only the last types of features are of importance. Another axis of characteristics of the home textile is in terms of styles of interior designs like rococo, post-modern, etc. Such features are important for retailing people who are trying to offer complete decisions to their customers and for the authors of articles in fashion magazines who describe realized interior designs. In order to ensure this coverage and granularity of the ontology we envisage the following steps of creation of the ontology:

*Processing of the standards and vocabularies in the domain*

We consider standards in the domain as reliable sources of conceptual information. Being created by leading experts in the domain with the goal to facilitate the whole process of production and usage of the home textile, the standards can be viewed as "expert questionnaires" usually used in the process of knowledge acquisition. Thus, we expect to find definitions of the most important concepts and relations in the domain of textile. The definitions also help us to establish the main relationships between the extracted concepts. As a means for the extraction of the concepts and the relations we are using partial analysis of the definitions (see Section 4 for more details). Then we inspect manually the analysis in order to identify the relevant knowledge. The result of this step is a list of (concept) terms (in English), a list of relations (relational terms), a list of triples – (term1 relation term2), and additional information like the number of some objects that are related to an instance of a given concept. These lists are the backbone of the ontology. Some of the relations are general ontological relations like *is-a*, *part-of*, etc. The rest of the relations are domain specific. Each term and each relation are connected to a natural language definition. These definitions have to reflect the triples for the term and the features of the relations (like whether it is an instance of more general relation, whether the reverse relation is also a relation in the domain which are encoded in the ontology, etc.). Later the definitions became part of the documentation of the ontology.

We consider this step as bottom-up approach to the creation of the ontology.

*Formalization of the terms*

The next step is to define formal definitions of the extracted concepts and relations in OWL-DL. We have selected OWL-DL, because there are implemented reasoners for it. For each term in the term list we construct a class definition in OWL-DL. We do the same for each relational term. We also encode the additional information in the definitions of the terms and the relations. The result of this step is an initial formal version of the ontology. It is possible that the level of errors in the concepts and the definitions is high. The possible errors are of two kinds – ontological and domain. The first ones are connected with the evaluation of the defined concepts and the relations between them with respect to meta-ontological properties. For the purpose to clean these errors we used the OntoClean approach – see [8]. It is discussed below. The second kind of errors was repaired with the help of the partners who are experts in the area of textile – see also below.

*Link to an upper ontology*

The establishing of the connection between the upper and the domain ontology helps us to check the consistency of the domain ontology with respect to the ontology construction methodology behind the upper ontology and to inherit the knowledge encoded in the upper ontology to the domain ontology. After we finish this process we have at our disposal the first version of the ontology which can be incorporated within the architecture of AsIsKnown system in order to be used and debugged.

There are several choices we can select in order to determine which upper ontology to be used as a basis of the development of the domain ontology. The initial list of ontologies included: DOLCE Ontology, SUMO Ontology, OpenCyc Ontology, Omega Ontology, Basic Formal Ontology,

PROTON Ontology, SmartWeb Integrated Ontology. For the selection of the upper ontology with respect to our purposes we consider the following criteria: (1) The ontology to be constructed on rigorous basis which reflects the OntoClean (or similar methodology) and suites our domain; (2) It to be easy to be represented in some of the ontological languages (OWL-DL preferably); (3) There to be domain ontologies constructed with respect to it (in order to facilitate the links with our domain ontology); (4) Support provided to us by the authors of the upper ontology. After some initial evaluation of the candidate ontologies and consultations with other evaluations of upper ontologies we have selected DOLCE Ontology as our upper ontology.

DOLCE ontology is descriptive, multiplicative ontology which adopts possibilism as an approach to existence and perdurantism as an approach to change. Selecting DOLCE as an upper ontology for our domain ontology we also adopt the formal background on which this ontology is constructed – OntoClean approach. Having in mind all these properties of DOLCE we decided that it is the most appropriate upper ontology for our purpose.

The linking from the domain ontology to DOLCE ontology is done manually. Each concept is connected to one or more concepts in DOLCE. Similarly each relation is attached to a relation in DOLCE. For relations could be the case that there is no appropriate relations in DOLCE, such relations have only local definitions in the domain ontology. The meaning of the links between the two ontologies are *is-a* (concept and relation specialization). Thus, the concepts and the relations inherit the definitions of the corresponding concepts (relations) in DOLCE. Also the OntoClean meta-properties are inherited. The next step is to check the consistency of the inherited information. If there are conflicts, then we examine the involved concepts and relations and the domain ontology are redefined locally. When the concept/relation definitions are inherited from DOLCE, it could be the case that there is a need to create new domain concepts/relations in order to have better inheritance. The result of this process is the first version of the ontology.

We consider this step as top-down approach to the creation of the ontology. Comparing to the definitions of top-down, bottom-up and middle-out approaches of [11] (see above), our approach is strictly top-down with respect to the inheritance from DOLCE to the domain ontology. With respect to the links from the domain ontology to DOLCE we can assume that it is bottom-up, because at some places it could be necessary we to introduce new concepts between the domain concepts and the concepts in DOLCE. We call these new concepts middle level concepts. Among the concepts in the domain ontology it is hard to predict the level of specificity of the extracted concepts; it depends on the information represented in the standards in the domain.

#### *Evaluation by domain experts*

The evaluation of the first version of the ontology was done in two ways:

#### *Practical evaluation*

The ontology is incorporated within AsIsKnown system. In order it to be used within the system, the information external to the AsIsKnown system was translated in the terms of the ontology. This information is: the producer data, describing their products; the ordering data, which needs to reflect the actual realised designs; interior design information, it is about combinations of materials and products in a design; user profiles; multimedia document annotation and description of trends. If all this is possible then we can assume that the ontology passes the practical evaluation. If some information can not be represented within the ontology, then we extended the ontology in order to cover these cases.

#### *Expert evaluation*

Then the ontology was reviewed by the partners in the project. The review was done on the basis of manual examination of portion of the ontology and by filling of specially developed questionnaires. The questionnaires are orientated to examine problematic places in the ontology. Such places are where there was no enough information in the textual materials on which the ontology is based; where there were most of the reconstructions made, because of the inherited from DOLCE information.

The result of the evaluation is in the form of requirements for introducing of new concepts, new relations or restructuring of the ontology. In the second turn of the development of the ontology we incorporated these requirements. The resulting ontology is the second version of the ontology.

#### *Documentation*



In the process of construction of the ontology we kept track on the sources of each concept, relation or axiom. The track records point to the context from where the concept originated. Also, we kept information about all changes in the definition of the concept or the relation. Important part of the documentation are the definitions in natural language created during the initial stage of concept definitions.

#### *Lexicons*

In parallel to the ontology construction we have done also lexicon creation in English, German, French, Hungarian and Bulgarian. These lexicons provide the vocabulary for the ontology concepts and relations in the corresponding language. In this way we facilitate the usage of the ontology for interaction to the human users of the system and the analysis of the multimedia documents. These lexicons could be considered as ordinary terminological lexicons in the domain, but the difference is that each term in them has a formal definition represented in the ontology. The English lexicon was constructed during the creation of the ontology, because we mainly processed English standards and vocabularies. English also was used as lingua franca within the ontology creation. For the creation of the other lexicons we used parallel sources (terminological lexicons in several languages) and terms provided by the partners.

### 3 Analysis of Domain Standards

The first step towards the construction of the domain ontology is the availability of an adequate terminological lexicon. We decided to use twofold approach, i.e. to process standards as well as other specialized domain lexicons. Both of them deliver at least the following text information: the term and the definition. The difference usually is in the structuring of the information. The standards usually give a hierarchically structured view of the terms (from more general to more specific ones), while the other specialized domain lexicons preserve a flat presentation of the notions. The definitions per se encode different relations, such as the prototypical **is-a** relation (hypernym-hyponym), **part-of**, **used-for**, **made-of** etc. For that reason we decided to use the definitions as an underlying corpus of data for extraction of ontologically significant relations. A similar data-driven method was explored by [1]. For the pre-ontological processing of the data we use the CLaRK System tool, described in [10].

The definitions from both types of resources have an established and therefore, easily predictable phrase structure. For that reason we assumed that the relations could be accessed through syntactic patterns. Hence, the definitions were syntactically analyzed at a (more or less) shallow level. Our analysis is neither pure chunking, nor full syntactic analysis. Its detailness was affected by the expected information from the various patterns. Around 1000 terms from domain lexicons and standards with definitions were manually processed. This linguistically created corpus was the base for the development of an automatic tool for the extraction of ontology relations.

The lexicon consists of entries. Each entry has a term and a definition. The text is further divided into sentences and phrases.

Our procedure includes two processing steps. First, manual annotation of the data and second, semi-automatic deriving of semantic patterns from the syntactic ones. Let us consider them in order.

#### *Processing of the data*

The definitions were tokenized automatically. Then a syntactic analysis was applied to tokens. We preferred manual annotation in order to have a reliable data at further stages of our work. The elements we annotated are the following: sentences (s), nominal groups (NP), verbs (V), prepositional groups (PP) and clauses (CL). Our rule is to annotate the longest sequences. Later on we can provide deeper analysis within these tags. It is assumed that the initial NPs are most likely hypernyms of the terms. They can come in a direct way like in:

```
<term>Garnetting<term>
  <definition><NP>a technique</NP> ...</definition>
```

Also, some more specific relation might be present in the hypernym:

```
<term>Gingham</term>
  <definition><NP>plain-weave lightweight fabric</NP> ...</definition>
```

or

```
<term>Isotactic</term>
  <definition><NP>a type of polymer</NP> ...</definition>
```

The verbal tag includes all the morphological forms of the verb (tense, voice etc.) and reflects different ontological relations too. For example, it can specify how the product is made (obtain), what it consists of (have, include, involve, contain), what it is used for or its function (use, create, dissolve, maintain, produce, resist, foster) etc. The PPs usually add to the semantics of the verb (used for, applied by, achieved by, measured in). When within NPs, prepositional phrases can indicate possession relation, part-of and others. In Fig. 2 three prepositional phrases within NPs with the preposition of are given, and one verbal PP with the preposition on, which has locational meaning. The three of-phrases are not homogeneous, which is expected to be observed by the human expert (the first prepositional phrase introduces a subject to a deverbal noun, the second indicates the form of the material and the third one equals the part-of relation).

The clauses are usually truncated relative clauses that modify some nominal group and give additional information. When no other predicate is present outside the clause, then the clause should be considered more closely. If, however, there is a main verb, then the clausal information might be temporarily omitted.

#### *Semi-automatic deriving of initial semantic relations from the syntactic patterns*

Our intention is first to derive semi-automatically the semantic relations, present in the more straightforward and easily predictable syntactic patterns. For the tricky and complex ones we rely predominantly on human inspection.

In order to excerpt an initial set of relations we have been applying several NLP techniques over the annotated data. One of these techniques is the so called concordance. It shows the selected data in context use. Applied on verbs, prepositional phrases, nominal phrases, it can give some idea on the typology of syntactic types and relations that can later on be matched to the semantic ones. The sort tool gives a powerful vision to the expert on both: the aggregation and the variety of the collocational phenomena (both syntactic and semantic). Another technique is the extraction of syntactic patterns. For example, the patterns NP V NP and NP V PP can give some insights about the relation type between two entities, entity and direction, etc.

Needless to say, also patterns of the type NP PP, NP NP, NP CL are to be considered.

Having derived different types of patterns, we further apply some statistics to them. The aim is to cross-check the relation of the natural language elements to the semantic abstraction patterns and to estimate the impact of the relations on their frequency within the definitions. Also, the ambiguity of the single elements are reduced, when combined in pairs or triples. For example, the preposition by has several latent meanings, but in the passive pair developed by it has the meaning of introducing the agent of the action. Another example is the preposition 'under'. Its trivial meaning is locational, but in the pair specified under it introduces some kind of regulation document. For each pattern the corresponding instances are stored and later transferred into ontological expressions.

All the information received through the NLP techniques was systematically repeated every time we expand our set of annotated data.

## 4 Conclusion

In this paper we presented an ontology development methodology which incorporate linguistic processing of standards within the domain of interest (in this case home textile). Using standards as an initial source of conceptual information helps us to overcome the communication gap between the knowledge engineers and domain experts. In this way the domain experts check the consistency of the already created ontology. In comparison to other domain texts standards are very reliable

sources of conceptual information. Being created by leading experts in the domain with the goal to facilitate the whole process of production and usage of the home textile, the standards can be viewed as “expert questionnaires” usually used in the process of knowledge acquisition. Thus, we expected to find definitions of the most important concepts and relations in the domain. The definitions also helped us to establish the main relationships between the extracted concepts. As a means for the extraction of the concepts and the relations we have been using a treebank constructed semi-automatically over the text of the standards. Then we inspected manually the analysis in order to identify the relevant knowledge.

## 5 Acknowledgements

This work has been supported by the FP6 European project AsIsKnown (A Semantic-Based Knowledge Flow System for the European Home Textiles Industry) (FP6-028044).

## Bibliography

- [1] Buitelaar, Paul. (2006). Knowledge Markup and Ontology Learning for Semantic Metadata Extraction. *An invited talk at the First Workshop on Natural Language Processing for Metadata Extraction - NLP4ME 2006.*, Varna, Bulgaria.
- [2] Fellbaum Chr. (1998). Editor. *WORDNET: an electronic lexical database.* MIT Press.
- [3] Mariano Fernandez-Lopez, Asun Gomez-Perez, Jerome Euzenat, Aldo Gangemi, Y. Kalfoglou, D. Pisanelli, M. Schorlemmer, G. Steve, Ljiljana Stojanovic, Gerd Stumme, York Sure. (2002). A survey on methodologies for developing, maintaining, integrating, evaluating and reengineering ontologies. *OntoWeb deliverable 1.4.* Universidad Politecnica de Madrid. 2002.
- [4] Mariano Fernandez, Asuncion Gomez-Perez and N. Juristo. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In: *Proceedings of AAAI97 Spring Symposium Series, Workshop on Ontological Engineering.* pp. 33-40.
- [5] Mariano Fernandez, Asuncion Gomez-Perez, Alexandro Pazos Sierra and Juan Pazos Sierra. (1999). Building a Chemical Ontology Using METHONTOLOGY and the Ontology Design Environment. *IEEE Expert (Intelligent Systems and Their Applications).* 14(1), pp. 37-46.
- [6] Asuncion Gomez-Perez, Natalia Juristo and Juan Pazos. (1995). Evaluation and Assessment of the Knowledge Sharing Technology. Towards Very Large Knowledge Bases. N.J.I. Mars. Ed. IOS Press. pp. 289-296.
- [7] Michael Grueninger and Mark Fox. (1995). Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of IJCAI95's Workshop on Basic Ontological Issues in Knowledge Sharing.*
- [8] Guarino, N., and Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean.* Communications of the ACM, 45(2): 61-65.
- [9] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). *Ontology Library (final).* WonderWeb Deliverable D18, December 2003. <http://www.loa-cnr.it/Publications.html>.
- [10] Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In: *Proceedings of the Corpus Linguistics 2001 Conference.* Lancaster, UK.
- [11] Mike Uschold and Michael Gruninger. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review.* 11(2).

# Design of a Multilingual Terminology Database Prototype\*

Mária Šimková<sup>1</sup>, Radovan Garabík<sup>1</sup>, Ludmila Dimitrova<sup>2</sup>

<sup>1</sup> Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

<sup>2</sup> Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia

**Abstract.** A prototype of a multilingual terminology database has been designed and implemented, with the intention to facilitate collaboration among MONDILEX member institutes, where either missing or incompatible Slavic languages terminology of modern aspects of linguistics can be a hindrance of mutual communication. The database is intended to contain entries with specialized corpus linguistics terms, and the prototype is filled with terms in Bulgarian and Slovak, with relevant English equivalents. The plan is to add terms in all the MONDILEX languages, and eventually release the database with the hope that its content will grow beyond the very narrow terminology of corpus linguistics.

## 1 Introduction

As the corpus linguistics is relatively new in Slavic languages – the development began only after the personal computer boom – there is no unified terminology of this field. The terminology started to develop uncontrollably, either by directly adopting English terms or by calquing the English expressions, or by embracing and extending existing linguistic terminology in each country. This development lead to widely varied terminology in different countries, and even to different terminology used by different institution in the same country, while sometimes the English terms are considered to be just a part of an informal slang.

The key issue is to harmonise the definitions and thus ensure consistency and clarity of information across the languages, especially when communicating with experts from various countries, where the use of bridge language is often not sufficient, or when dealing with bi- or multilingual resources, with the consequent need of multilingual documentation.

Since “the ultimate purpose of any terminological resource is to facilitate and enhance knowledge acquisition” [2], the database has been designed in a way to function as a quick reference source of terms in different languages, which has influenced its overall design.

The database, once finished, could be also used to compare the usage and acceptance of English terms in various languages.

Extensive and theoretical study on definitions and formalism is beyond the scope of this paper – we describe only the technical implementation and general features of our database.

## 2 Implementation

Multilingual terminology database (MLTD) is developed using the MoinMoin wiki engine as a backend. The data is kept in plain text files, with one file (MoinMoin page) corresponding to one terminology entry. The technical implementation, and to an extent a terminology entry structure has been inspired by the Slovak Terminology Database design [4, 5].

As a minimum, a terminology entry in MLTD should contain a term, its definition (explanation) and a source of the definition. Intentionally, MLTD tries to keep the minimalistic approach and therefore adds no additional data.

Compared with the simplicity of MLTD, Slovak Terminology Database entry has 13 fields, 5 of them are obligatory (*term*, *field*, *definition*, *biblio*, *acceptability*). *Field* is simulated by the page category, and *acceptability* (pragmatic term character, one of *normalised*, *legislative*, *recommended*,

---

\* The study and preparation of these results have been partly supported by the EC's Seventh Framework Programme [FP7/2007-2013] under the grant agreement 211938 MONDILEX.

*suggested, incorrect, archaic, neologism*) is mostly relevant for national terminology systems dealing with terminology standardisation, and as such has no place in MLTD – it is implicitly included in the information about definition source.

This design allows the internal format of the database entry to be kept very simple, nothing more than a plain text file with a minimal layout, without any special formatting markup. By a design decision, internal page format does not use any immediately visible markup language. The motivation stems from our empirical observation regarding usability – the presence of any, even the most inopious markup distracts the editors, unless they are reasonably well trained in the markup (and discourages them to learn to use the system). Our markup is hidden in the overall text structure, using nothing more than strategically placed paragraph breaks, colons and parentheses used in a relatively (hopefully) intuitive way.

Each page consists of several entries (one for each language), separated by an empty line. Each entry starts with a term name, prefixed with an ISO 639-1 language identifier separated by a colon (:), followed by an empty line, followed by a definition, followed (immediately) by a source of the definition. Each page can belong to one or more categories – these are expressed by using the usual category mechanism (adding `Category*` link to the end of the page). For the prototype described, there is just one category used, `CategoryCorpusLinguistics`.

Terms in corpus linguistics entered Slavic languages mostly from the English language. The origin of a significant number of them (mostly purely linguistic terms), however, was known long before the differentiation of corpus linguistics as an independent branch of linguistics. These terms have originated either in Greek or Latin: for example, *corpus* and *segment* came from Latin, *lemma* and *lexeme* from Greek. It is even possible that the same term entered different target languages through different intermediaries.

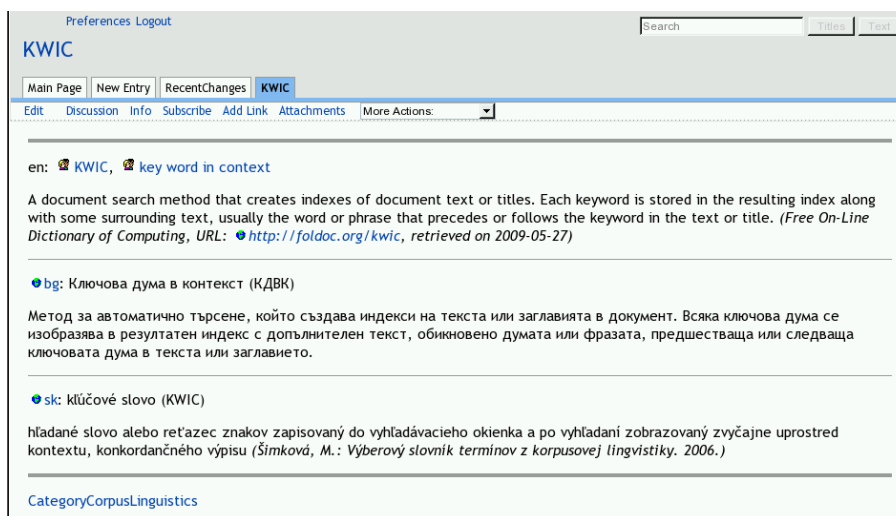


Fig. 1. Example of an entry

A special parser for MoinMoin has been written to display the entries in a distinct graphical way. Main features of the parser are:

- language entries are separated by a horizontal ruler
- ISO 639-1 language identifiers point to an external URL with more information about the language used
- English term is hyperlinked with the corresponding English Wikipedia entry
- definition source is emphasized
- URLs in definitions or sources are automatically recognized

en:KWIC, key word in context

A document search method that creates indexes of document text or titles. Each keyword is stored in the resulting index along with some surrounding text, usually the word or phrase that precedes or follows the keyword in the text or title.  
(Free On-Line Dictionary of Computing, URL:  
<http://foldoc.org/kwic>, retrieved on 2009-05-27)

bg:Ключова дума в контекст (КДВК)

Метод за автоматично търсене, който създава индекси на текста или заглавията в документ. Всяка ключова дума се изобразява в резултатен индекс с допълнителен текст, обикновено думата или фразата, предшестваща или следваща ключовата дума в текста или заглавието.

sk:kľúčové slovo (KWIC)

hľadané slovo alebo reťazec znakov zapisovaný do vyhľadávacieho okienka a po vyhľadaní zobrazovaný zvyčajne uprostred kontextu, konkordančného výpisu (Šimková, M.: Výberový slovník termínov z korpusovej lingvistiky. 2006.)

----  
CategoryCorpusLinguistics

**Fig. 2.** Internal representation of an entry

```

<entry> ::= <language entry> {<p> <language entry>} \n ---- \n <category>
<language entry> ::= bg | cs | en | pl | ru | sk | sl | uk : <terms>
                \n {\n} <definition> ( <bibliography> )
<definition> ::= ? characters ?
<bibliography> ::= ? characters ?
<terms> ::= <term> | <term> , <terms>
<term> ::= ? characters ?
<p> ::= \n \n {\n}
<category> ::= Category ? characters ?

```

**Fig. 3.** Formal description of an entry syntax

The points outlined are implemented in order to make the navigation around the database more efficient – they should be thought of as a visual and formatting aid to the database representation, not as a part of the database itself. In fact, the parser can be very easily modified to accommodate different visual styles and different formatting representations.

The database can use all the usual MoinMoin features concerning efficient collaborative editing. The most relevant ones, emphasised by the database design are:

- efficient indexing and searching, using the built-in Xapian search engine (even if for the database of the intended size – hundreds of entries at most, any search engine is more than sufficient)
- full Unicode support, with only some limitations concerning right-to-left scripts (irrelevant for Slavic languages)
- full editing history with backup of page revisions, allowing to see the complete history of previous entry versions
- review of differences between arbitrary page versions, using diff-like output with coloured differences
- multiuser support with full access control list – however, our database does not use complicated permission schemes, relying on the ease of reverting unwanted changes instead
- warnings to avoid editing conflicts, in case when two users intend to edit the same entry simultaneously

As a prototype, the database has been filled with corpus linguistics entries from [3], which has been compiled as a concise list of term (cf. the needs of colleagues from Czech Republic, where two different lists have been compiled: [1], [7], including data from other areas of linguistics).

We faced following problems when converting the data into MLTD:

- Homonymy:
  - Corpus linguistics is an intradisciplinary research field, where two different areas meet – computer science and linguistics, and these two areas sometimes use the same word to denote (often a little) different objects. Traditional lexicography deals with this polysemy using numbered entries for each meaning (e.g. **corpus 1.** database of digital texts..., **2.** collection of texts for a specific kind of research). The Slovak Terminology Database separates the meanings into different entries, with headwords marked by the numeral.
  - Often encountered problem is a dichotomy of meaning of verbal derived nouns, where a noun can mean both a process and its result (e.g. *annotation* can be both the process of annotating and the resulting data). In the area of terminology, these two meanings are considered to be strictly separate.
- Traditionally, synonymy in dictionaries is reflected in a lexical entry either in the heading (as two or more equal headwords) or after a definition, while they can form a reference to a relevant entry (e.g. **anotácia – tagovanie – značkovanie**). In the Slovak Terminology Database, synonyms are stored in a separate input field (and are automatically hyperlinked). In the MLTD, different terms have to be kept separately. There is no provision in MLTD for entering synonyms.
- Terminology entries have been often described using encyclopedic style and format – under the general headword there are often specified other, narrow meanings (e.g. **korpus** — **korpus hovorených textov**: elektronická databáza hovorenej formy jazyka; – **korpus písaných textov**: elektronická databáza písanej formy jazyka; — **národný korpus**: jednojazyčný korpus textov konkrétneho národného (jazykového) spoločenstva; — **synchronný korpus**: korpus jazyka v jeho súčasnej vývinovej fáze; — **všeobecný korpus**: nešpecifický, základný korpus zahŕňajúci široké spektrum jazykových štýlov a žánrov, vecných oblastí (domén), autorských generácií, vydavateľských úzov, regiónov a pod.). However, in the MLTD, each of the meanings has to be entered separately.
- In the Slovak Terminology Database, each term has a facultative field for storing (arbitrary) foreign language equivalents; in the MLTD, the only equivalents are those given in the other languages present.

### 3 Conclusion

The database is envisaged to contain entries in following languages: Bulgarian, Czech, English, Polish, Russian, Slovak, Slovene and Ukrainian. The English has been added as a semi-bridge language, unifying the entries (and taking into account that most of the terminology originates in the English language).

As a prototype, the database has been filled with corpus linguistics entries from the Slovak from the Slovak Terminology Database, together with their English equivalents (but missing English definitions), and with Bulgarian terms added later. Overall, considering the abovementioned discrepancies in database designs, 45 corpus linguistics terms were imported, out of about 150 terms present in the Slovak Terminology Database.

### Bibliography

- [1] Čermák, F. (1997). Slovník lingvistických termínů. In *Jazyk a jazykověda*, Prague. Pražská imaginace.
- [2] Faber, P. & Sánchez, M. T. (2004). Codifying conceptual information in descriptive terminology management. *Meta*, 46(1), 192–204.
- [3] Šimková, M. (2006). Výberový slovník termínov z korpusovej lingvistiky. In M. Sokolová, M. I. (Ed.), *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli.*, Prešov. Filozofická fakulta Prešovskej univerzity.
- [4] Levická, J. (2007). Terminology and Terminological Activities in the Present-Day Slovakia. In J. Levická, R. G. (Ed.), *Computer Treatment of Slavic and East European Languages*, Brno. Tribun. Proceedings of the conference Slovko 2007.
- [5] Levická, J. (2008). Analysis of “classical” and legislative definitions for the term records of the Slovak terminology database. Proceedings of the Third Conference on Translation, Interpreting and Comparative Legi-Linguistics. Poznań, Poland. In print.
- [6] Popov, D. (1994). *Bulgarian Explanatory Dictionary*. Sofia: Nauka i Izkuvtvo Publishing House. In Bulgarian.
- [7] Šulc, M. (1999). Výběrový slovníček pojmů z lingvistiky. In *Korpusová lingvistika*, Prague. Karolinum.



# Dictionary of Slovak Collocations\*

Peter Ďurčo<sup>1</sup>, Radovan Garabík<sup>2</sup>, Daniela Majchráková<sup>2</sup>, Matej Ďurčo<sup>3</sup>

<sup>1</sup> St. St. Cyril and Methodius University, Trnava

<sup>2</sup> Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

<sup>3</sup> Austrian Academy Corpus, Austrian Academy of Sciences, Vienna

**Abstract.** Presented lexical database of Slovak language collocation should cover collocation profiles of several hundred words of different parts of speech (nouns in the first phase of the project) and will be a base of a modern collocation dictionary. The database is built using MediaWiki engine, which offers excellent remote collaboration features along with automatized processing possibilities.

## 1 Introduction

The standard use of corpora for linguistic research and lexicography is aimed predominantly at the examination of occurrences and co-occurrences of word forms and lemmata. The main goal is to acquire data about semantic, grammatical and combinatorial behavior of words.

For the Slovak language, the only one existing collocation dictionary has been published in 1931, with a revised edition in 1933 (the author called this book ‘a dictionary of phrasemes’, but in fact it has been a dictionary of not only phrasemes, but also of common word collocations) [15, 16]. Clearly, since then the whole language underwent immense changes in almost all of its parts, starting with the whole sociolinguistic situation and ending with substantial changes in the vocabulary and orthography. By today, the dictionary is mostly of diachronic importance, and there is a notable gap in Slovak language lexicography concerning a database of collocations – modern approaches in lexicography, especially the use of large language corpora fill the gap somewhat, but they still cannot replace a well documented, systematically built dictionary.

Presented electronic dictionary of Slovak collocations is being compiled at the University of St. Cyril and Methodius, Trnava in cooperation with the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava. The project on Slovak collocations that started in 2007 is the first of its kind in Slovakia and is aimed at the registration and description of selected multiword lexemes and phrasemes as well as typical collocations with restricted collocability. The dictionary provides an overview of the combinatorial behaviour of words, in the first phase the most frequent nouns extracted from the Slovak National Corpus database, with the intention to include also verbs, adjectives, adverbs and particles. Currently, the database contains information about nouns and (as a separate subproject) particles. The combinatorial potentials of word forms of a word are the basis for the creation of so-called collocational templates which the patterns of collocations are based on [17]. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partial collocations.

## 2 Obtaining collocation profiles

An efficient tool for modelling semantic proximity of words and their collocation profiles in large lemmatized corpora is the sketch engine<sup>4</sup> [12] – a corpus tool which generates word sketches, i. e. corpus based summaries of a word’s grammatical and collocational behaviour. Disadvantages of the

\* The lexical database has been supported by the grant agreement VEGA 1/0006/08 *Konfrontačný výskum kolokácií v slovenčine a v nemčine*. The study and preparation of these results have been partly supported by the EC’s Seventh Framework Programme [FP7/2007-2013] under the grant agreement 211938 MONDILEX.

<sup>4</sup> <http://www.sketchengine.co.uk/>

sketch engine are long lists of isolated lemmata and too many automatically generated redundant data in the results, obtained through fixed set of unary, dual, symmetric and trinary rules, which do not always correspond to natural collocational clusters in the language.

The basic tool for searching collocations for each entry is the corpus manager client Bonito which provides searching, sorting and statistical evaluation of collocations. By using this tool we can observe each given word, extract concordances for each word to get an overview of its behaviour in a context, get statistical information like absolute frequency, MI-score, t-score, MI3, log likelihood, min. sensitivity and salience to recognize word co-occurrences [13].

Despite these new language technological analysis, scepticism still prevails regarding the possibility of seizing and of describing examined data completely. This scepticism results particularly from two problems. Word co-occurrences represent a diffuse continuum of semantically differently strong connected elements. The borders between “free” and “firm” can not be specified clearly. On the other hand, the main problem of the statistical approach is that the frequency and semantic firmness of word combinations do not correlate directly. Not all high frequent word combinations are also firm. One finds typical collocations in all ranks of the frequency distribution [9, 10].

In our lexical database, the (meaningful) collocations are manually selected from the first 500 occurrences of each grammatical structure listed by The Sketch Engine and cross-checked against the Slovak National Corpus concordances.

The statistical results vary, they depend both on the used statistical method and the quality and accuracy of taggers and lemmatisers, the precision rates whereof are different. It means that we have to compare very long lists of indexes from different scores. The table 1 shows the identity of collocation candidates between scores within the first 50 rows.

	T-score	MI	MI3	log likelihood	min. sensitivity	salience
Abs. freq	73.9	5.4	5.4	54.9	34.4	36.6
T-score		5.0	53.6	70.1	62.0	44.4
MI			32.7	24.5	14.0	36.0
MI3				75.0	57.1	81.6
log likelihood					71.4	69.4
min. sensitivity						60.0

**Table 1.** Comparing collocations – identity between different measures, in percents.

### 3 Technical implementation of the lexical database

Since the dictionary has been conceived from the beginning as a collaborative project involving several contributors, the choice of the working environment has been driven by several requirements – easy remote editing, access control list, revision history, communication between editors. These requirements can be easily met by deploying a wiki based software, we have chosen MediaWiki software system, with MySQL as a relational database backend.

MediaWiki is written in the PHP programming language and has many attractive options for the intended purposes, among them the possibility to use templates (a kind of macro) for better handling of repeating text parts. Templates are basically predefined text snippets in wiki-format with additional specialized markup for accommodating passing of arguments which are dynamically loaded inside another page. More on this in section 6.2.

While a wiki system has proved as highly suitable for the task of creating the dictionary, the way of representing the dictionary information to the end user is still an open question, the layout provided by the wiki-entries being probably not the most appealing and useful one.

lemma	MI-score	lemma	T-score
anakreontov	13.03	ľudový	50.70
hiawathowa	13.03	text	23.91
kancionálový	12.99	táto	20.87
paraliturgický	12.96	tá	19.91
brelov	12.77	duchovný	19.54
rózsov	12.77	spievať	19.41
švihrovsky	12.77	nový	19.29
<b>hymnický</b>	12.74	<b>labutí</b>	17.57
dylanový	12.55	populárny	16.65
pestúnkina	12.31	vianočný	15.85
slávikoví	12.19	pieseň	15.55
<b>labutí</b>	12.07	známy	15.07
schubertový	12.03	zaspievať	14.04
legendický	12.01	<b>nábožný</b>	13.63
švihrovský	11.90	titulný	13.17
pijácky	11.86	náboženský	12.78
podkovitý	11.77	<b>hymnický</b>	12.61
cherubínsky	11.77	oblúbený	12.61
povaľačský	11.77	ľúbostný	11.54
regrútsky	11.70	rómsky	11.14
symfonia	11.36	večný	10.78
lennonový	11.36	smutný	9.04
silvánov	11.27	mariánsky	8.51
mický	11.12	svadobný	8.49
zaspievanie	11.09	oslavný	7.99
carlina	10.96	rusínsky	7.99
barnabášov	10.96	tanečný	7.90
kramársky	10.82	skladať	7.83
trampský	10.79	lyrický	7.11
<b>nábožný</b>	10.78	pohrebný	7.10

**Table 2.** Differences in the lists of collocation candidates extracted by MI-score and T-score, lemma *pieseň*. Words in boldface are shared between top 30 occurrences of both scores.

## 4 Prerequisites

In the initial phase of the project, the collocations were obtained from Slovak National Corpus (SNK), version *prim-3.0* containing about 330 million tokens. Halfway during the work on the database, a new version of the SNK has been released (*prim-4.0*), bringing the number of tokens up to 530 million, which faced us with a dilemma: as the new version had not only substantially increased in the volume, but also improved lemmatization and morphology annotation, it would be advantageous to use this new information, but on the other hand, changing the input data would require to go through and redo all the entries already done. At the end, we decided to use the new version for new entries and analyse the collocational profiles with respect to changed statistical measures in order to evaluate the changes brought by a new corpus.

There is also the question of which variant of the corpus to use – there are three main flavours of the Slovak National Corpus database, *prim-4.0-public-all* contains all the texts, *prim-4.0-public-sane* contains only texts that satisfy certain requirements (correct diacritics, non linguistic texts, no texts written by Slovak minorities outside of Slovakia proper, some controversial writers removed), *prim-4.0-public-vyv* is a balanced corpus, containing 1/3 newspaper texts, 1/3 fiction, 1/3 scientific texts (see Tab. 3 for comparison).

Version	number of tokens		
	-all	-sane	-vyv
prim-3.0-public-	339 063 215	319 644 966	199 822 572
prim-4.0-public-	526 082 640	507 101 251	254 236 903

**Table 3.** Comparing versions 3.0 and 4.0 of the Slovak National Corpus.

Corpus	prim-4.0-public-all	prim-4.0-public-sane	prim-4.0-public-vyv
Identity: sane+vyv		75.5%	
Identity: all+sane	93.5%		
Identity: all+vyv		74.5%	
Identity: all+sane+vyv	73.4%		
Isolated occurrences	9.7%		

**Table 4.** Comparing identity of collocation candidates of the word *pieseň* (song) in three different versions of the Slovak National Corpus, version 4.0.

## 5 Basic structure of the database

The database serves two different purposes – the first is to build a Slovak language collocation dictionary, the second one to build a (semi)bilingual dictionary of German collocations with Slovak equivalents [18, 20]. These two projects share the same database and the same MediaWiki instance, and (to an extent) use the same methods and guidelines regarding the collocation profiles. However, logically these are two separate projects. In this paper we deal exclusively with the Slovak dictionary.

The database macrostructure is simple – all the entries are equal, each entry corresponds to one MediaWiki page, we are using neither subpages<sup>5</sup> nor redirects<sup>6</sup>. A page is named by an entry lemma, in case of clash between German and Slovak (e.g. Internet, System), the Slovak page adds the string ‘(sk)’ to the page name, so that the pages will be named ‘Internet (sk)’, ‘System (sk)’. Unfortunately, MediaWiki automatically converts the names to titlecase, otherwise the compulsory capitalization of German nouns would distinguish between German and Slovak entries. Slovak lexical entries are differentiated from other pages (system pages, German entries, user discussions) by the category they belong to (one of Slovak Nouns, Slovak Adjectives, Slovak Verbs, Slovak Particles).

## 6 Structure of an entry

An entry page consists of three main sections: *Významy* (Meanings), *Kolokácie* (Collocations), *Externé odkazy* (External links). While the structure of *Významy* and *Externé odkazy* is the same for all the parts of speech and these sections do not have any substructure, the structure of *Kolokácie*, the most important section, is more complicated [19].

### 6.1 Významy

This section (“meanings”) contains a bullet list of descriptions of different definitions of the lexeme. We do not split the collocations according to polysemy (or homonymy) of the base noun inside one part of speech category at all, neither we distinguish between homonyms in collocations. This was a deliberate design decision, based on two observations: First, often a collocation is not clearly attributable to a specific meaning; let alone trying to define and distinguish meanings, which is traditionally a very cumbersome task, where no general consent could be achieved. This was not seen as a task for this project and would unnecessarily considerably slow down the dictionary constructions and open door to endless discussions inside and outside the project team about the distinction of individual meanings.

### 6.2 Kolokácie

All the collocation data are contained in this section. The detailed structure is differentiated according to part of speech the entry stands for. For nouns, it is divided into two subsections for the singular and plural, reflecting the fact that collocates often exhibit different phenomena according to the grammatical number of the base noun. Each of these subsections is further divided into many subsubsections, each for a specific collocation combination (see Fig. 1, 2).

The subsubsections’ naming scheme encodes some human readable information about the collocations, with the base noun marked by the string *Sub1Xxx*, where *Xxx* is the abbreviation of the noun’s case (so the whole string will be one of *Sub1Nom*, *Sub1Gen*, *Sub1Dat*, *Sub1Aku*, *Sub1Lok*, *Sub1Ins*). We are ignoring the Slovak vocative controversy by conflating (semantic) vocatives with the nominative case – fortunately, it just happened that none of the nouns chosen for the collocation dictionary is from the set of those few Slovak words that have a morphological vocative (either having retained their Old Slavic vocative forms<sup>7</sup>, or developing a ‘new vocative’, common for some proper nouns and family relationships<sup>8</sup>).

The other part of the subsubsection name reflects describes the neighbouring word part of speech, so it can be one of *Sub2*, *Verb*, *Atr* (another noun, verb, attribute). *Atr* subsumes adjectives,

<sup>5</sup> A subpage is a page that is subordinate to its parent page. Subpages can be used to implement a whole hierarchy (tree structure) of entries, which – considering lexicographic use – can be used to distinguish between homonymy and polysemy [11]. However, given the small number of entries, we decided to refrain from this.

<sup>6</sup> A redirect is a page that has no data by itself; it just refers to another page.

<sup>7</sup> e.g. *otče*, *pane*, *bože*

<sup>8</sup> e.g. *babi*, *mami*, *Zuzi*, *Feri*

pronouns, particles or numerals. This string is positioned either to the left or to the right of the previous base noun string, depending on the predominant position of the word in collocations (but including also the collocations with a different word order). The strings are concatenated with a plus sign, so e.g. the whole subsection name *Verb + Sub1Aku* indicates that the subsection contains collocation of verb and base noun in acusative (not necessarily in this order).

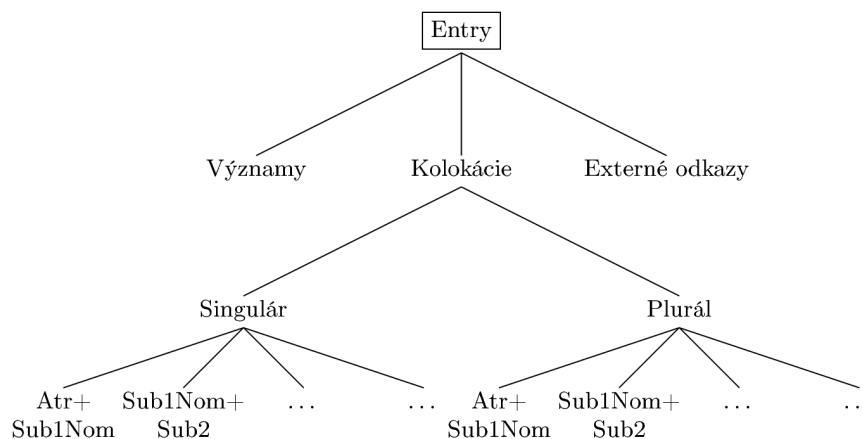


Fig. 1. Entry structure diagram for nouns

Sub1	Sub2	Verb	Atr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Atr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Atr
...	...	...	...
Sg Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Atr
Pl Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Atr
Pl Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Atr
...	...	...	...
Pl Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Atr

Fig. 2. Matrix for the entry structure of a noun

### 6.3 Externé odkazy

This section is populated by several macros (templates), providing links to external resources. Each macro has one parameter, equal to the identification of given word in the target database – mostly the same as the lemma, different only in case of homonyms (differentiated at the target). The macros construct an URL pointing to an external resource and insert it as an http hyperlink into the rendered page. The macros in use are `{{ma|...}}` to link to morphologic database (this macro is intended to record relations between full word paradigms and the collocation dictionary entries, both for the end user and for eventual computer processing), `{{slovník|...}}` to link to dictionaries[7] published at the Ľ. Štúr of Linguistics WWW page, `{{linky|...}}` to point to several search engines, such as Google[1], Ask[2], Yahoo[3], Cuil[4], as well as the Slovak National Corpus[6]. The latter two templates are meant for human consumption, not for computer parsing (due to somewhat unpredictable nature of the target data). In case we need to either add or remove

an external data source (e.g. a search engine), or if the form of URL parameters changes, we need to modify just the template, and the change will be automatically reflected across all the database entries.

## 7 Automated database processing

There are several options for automated data modification. First and most obvious is to access the SQL backend directly, reading and modifying the tables. However, this method requires detailed knowledge of internal MediaWiki database structure, and modifying would have to be done with a great care, in order not to disrupt the database and introduce structural inconsistencies.

Much better way is to use a MediaWiki API, designed for a remote access. As the MediaWiki is probably the most widely used Wiki framework, there is a plethora of tools available[5] for automated processing in various programming languages. However, we settled on using a slightly different approach – WikipediaFS[8], a fuse-based[14] filesystem that presents remote WikiMedia installation as a fake filesystem, so that the pages can be read and written as simple text files, either for automated scripted processing or to be edited with an ordinary text editor. The advantage of WikipediaFS over using MediaWiki API is the availability of plain text, filesystem like view of the data, which makes it easy to use standard UNIX command line tools for text processing (`sed`, `awk`, `grep`, ...). We used WikipediaFS and some simple scripts to add automatically the abovementioned links to external resources to all the entries in the database.

## 8 Collocation entry microlanguage

The lexical database has been designed with a goal of a human readable collocation dictionary in mind, published both online and in printed form. However, the importance of the need to keep the data in computer readable format cannot be stressed enough – if nothing else, to automatise the typographic formatting process for the printed version, and indexing for the online version. Therefore the entry microformat is designed to be computer readable, except of some minor exceptions, where the (complete) readability stands in the way of human interaction.

Each collocation can be thought of as consisting of two units: the base noun and the collocate. The collocates are normalised (lemmatised), and the collocation is written with the base in its corresponding case/number. The exception is only for the combination *Atr + Sub1Nom*, which is so frequent that we omit the base in nominative, if it follows the attribute. Auxiliary particles/pronouns are sometimes rearranged, to fit the syntactical requirements of the base (this applies mainly to the reflective pronouns *sa*, *si* in combination with infinitives). From this follows that the parser must include the morphology generator in order to recognise the base noun in other forms than nominative singular, and a complete automatised parsing is difficult without including some sort of syntactical rules into the parser. Collocate is terminated by the | (U+007C VERTICAL LINE) character surrounded by whitespace. The vertical line has to terminate also the ultimate collocate in the subsection. If there are no collocates for a given collocation pattern, the entry consists of a single vertical line character in a separate line. Optional words (which are sometimes present in a given collocation) are enclosed in parentheses, separated by the rest of collocation by a whitespace or punctuation. Parentheses adjoined to a word specify optional prefixes or suffixes (mostly verb negation or aspect modifier). Variants in words (two or more words that do not change the collocation meaning and are approximately equally frequent) are separated by a slash, three dots (ellipsis, ...) denote incomplete variant enumeration (signalling that there are more variants occurring in the corpus than given, usually these variant components belong to a specific lexico-semantic group). Special indefinite pronouns (*niekto*, *niečo*, ...) serve as wildcard valency markers which stand for a general class of animate/inanimate nouns (and thus signal that the collocation is too broad to be automatically parsed).

There are on average 173 collocations per entry – the distribution of entry sizes is depicted on Fig. 4. We see that the symmetry is slightly skewed in favour of small number of bigger sized entries (the median is 157). The entry with fewest number of collocations is *kára* (cart, barrow), with 40

collocations, the highest number has the word *svet* (world) – 584 collocations. However, we have to realise that the exact number of collocations per entry is subject to several arbitrary conditions, among them the level of detail in describing collocation variants, inclusion of otherwise optional ellipsis and indefinite pronouns, and in general subjective evaluation of collocation candidates by a lexicographer compiling the entry.

==Atr + Sub1Gen==

neznalý pomerov | z chudobných pomerov | znalý pomerov |

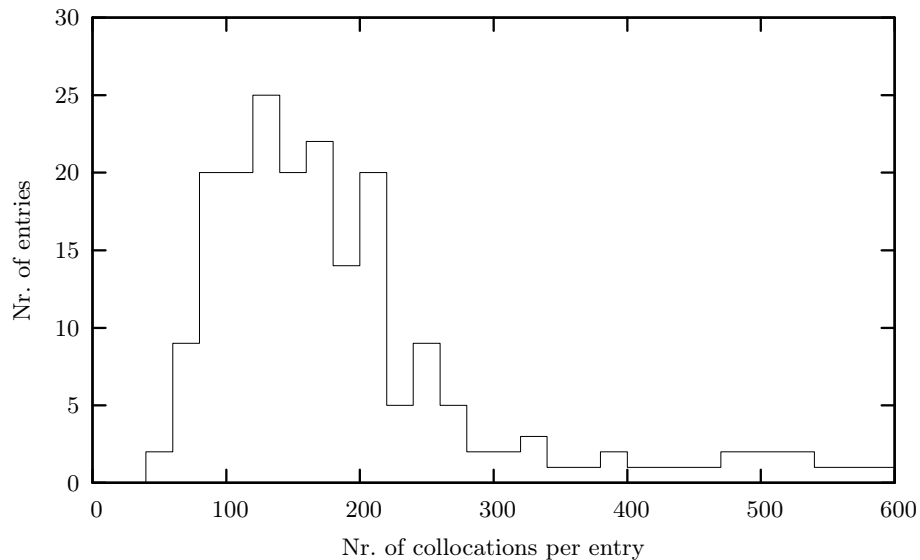
==Sub2 + Sub1Gen==

demokratizácia pomerov | konsolidácia pomerov | kritika pomerov |  
 neznalosť pomerov | obraz (politických / reálnych / ... ) pomerov |  
 stabilizácia pomerov | úprava pomerov | usporiadanie pomerov |  
 zlepšenie pomerov | zmena spoločenských / vlastníckych pomerov |  
 znalec (našich domácich) pomerov | znalosť tunajších pomerov |

==Verb + Sub1Gen==

pochádzať z (dosť) chudobných / skromných pomerov |

**Fig. 3.** Fragment of a collocation entry, word *pomer*



**Fig. 4.** Distribution of number of collocations per noun, bin size = 20



## 9 Further plans & Conclusion

The plan for the first phase of the project is to create a dictionary of noun collocations, with the number of entries exceeding 500. Currently, the database contains collocation profiles of 190 nouns and 38 particles.

After the first phase, a new methodology for a dictionary of other parts of speech will be delineated and the dictionary will be extended. It is expected that by that time a new version of the Slovak National Corpus database will be available, and already existing entries could be cross validated against these new data. The dictionary will be a valuable contribution to modern Slovak language lexicography, reflecting real language usage by being based on the real data from the Slovak National corpus.

From the theoretical point of view, research of collocations will add to our knowledge about the collocability of words, presented collocation database can serve as a base for confrontational Slovak language research. Collocations per se form an inseparable part of many different kinds of dictionaries, and they are especially important in language teaching, giving examples of real language usage. We believe that the collocation dictionary will be used in teaching Slovak as a foreign language, since the mastery of idioms is a sign of a true language competency.

## Bibliography

- [1] <http://www.google.com>.
- [2] <http://www.ask.com>.
- [3] <http://www.yahoo.com>.
- [4] <http://www.cuil.com>.
- [5] Botwiki. <http://botwiki.sno.cc/wiki/Manual:Frameworks>. A wiki for documenting and testing bots. Retrieved 2009-06-08.
- [6] Slovak National Corpus. <http://korpus.juls.savba.sk>.
- [7] Slovenské slovníky. <http://slovník.juls.savba.sk>.
- [8] Blondel, M. WikipediaFS. <http://wikipediafs.sourceforge.net/>. Retrieved 2009-06-08.
- [9] Čermák, F. (2006a). Kolokace v lingvistice. In Čermák, F. & Šulc, M. (Eds.), *Kolokace. Studie z korpusové lingvistiky. Sv. 2.*, (pp. 9–16), Prague, Czech Republic. Nakladatelství Lidové noviny, Ašstáv Českého národního korpusu.
- [10] Čermák, F. (2006b). Statistical Methods for Searching Idioms in Text Corpora. In Häcki-Buhofer, A. & Burger, H. (Eds.), *Phraseology in Motion 1. Methoden und Kritik*, (pp. 33–42). Baltmannsweiler (Schneider Verlag Hohengehren).
- [11] Garabík, R. & Špirudová, J. (2009). Design of a New Slovak-Czech Lexical Database. In Garabík, R. (Ed.), *Metalanguage and Encoding Scheme Design for Digital Lexicography*, (pp. 71–76), Bratislava, Slovakia. Tribun.
- [12] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The sketch engine. *Information Technology*, 105.
- [13] Majchráková, D. & Ďurčo, P. (2009). Compiling the First Electronic Dictionary of Slovak Collocations. To be published.
- [14] Szeredi, M. Filesystem in Userspace. <http://fuse.sourceforge.net/>. Retrieved 2009-06-08.
- [15] Tvrđý, P. (1931). *Slovenský frazeologický slovník*. Trnava: Spolok sv. Vojtecha.
- [16] Tvrđý, P. (1933). *Slovenský frazeologický slovník. Druhé doplnené vydanie*. Praha and Prešov: Nákladom Československej grafickej unie, úč. spol.
- [17] Ďurčo, P. (2007a). Collocations in Slovak (Based on the Slovak National Corpus). In Garabík, R. & Levická, J. (Eds.), *Computer Treatment of Slavic and East European Languages*, (pp. 43–50), Bratislava, Slovakia. Tribun.
- [18] Ďurčo, P. (2007b). O projekte nemecko-slovenského slovníka kolokácií. In Baláková, D. & Ďurčo, P. (Eds.), *Frazeologické štúdie V. Princípy lingvistickej analýzy vo frazeológii*, (pp. 70–93), Ružomberok, Slovakia. Katolícka univerzita v Ružomberku.

- [19] Ďurčo, P. (2007c). Zásady spracovania slovníka kolokácií slovenského jazyka. <http://www.vronk.net/wicol/images/Zasady.pdf>. Online documentation.
- [20] Ďurčo, P. (2008). Zum Konzept eines zweisprachigen Kollokationswörterbuchs. Prinzipien der Erstellung am Beispiel Deutsch-Slowakisch. In Hausmann, F. J. (Ed.), *Collocations in European lexicography and dictionary research. Lexicographica*, volume 24, (pp. 69–89)., Tübingen, Germany. Max Niemeyer Verlag.

# A Comparison of Two Morphosyntactic Tagsets of Polish<sup>\*</sup>

Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences

**Abstract.** The aim of this paper is to present the main differences between the IPI PAN Tagset, used for the morphosyntactic annotation of the IPI PAN Corpus of Polish, and the NKJP Tagset, employed in the National Corpus of Polish.

## 1 Introduction

Morphosyntactic tagsets, i.e., formal specifications of morphosyntactic interpretations assigned to words in a given language, are usually developed for the purpose of the morphosyntactic annotation of corpora. While presentations of morphosyntactic systems of various languages found in textbooks and grammars may be sufficient for many linguistic purposes, the task of assigning a morphosyntactic tag (in short: tag) to each word in a large corpus requires a codification of such a system. The resulting tagset must exhaustively specify the repertoire of grammatical classes (parts of speech) assumed for the language, morphosyntactic categories appropriate for particular classes, and possible values of these categories.

A tagset of Polish called the IPI PAN Tagset was proposed in a series of papers (in English: Przepiórkowski and Woliński 15, 16; in Polish: Woliński 19 and Przepiórkowski 11; summarised in the bilingual publication Przepiórkowski 12, 13) within the IPI PAN Corpus (<http://korpus.pl/>) project.<sup>1</sup> Since then, the tagset has been used in a number of projects, including various projects carried out by the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences, as well as, e.g., in the Polish WordNet project (Piasecki et al. 10; <http://plwordnet.pwr.wroc.pl/>), it inspired the tagset used in the Morfologik dictionary (<http://morfologik.blogspot.com/>), and it influenced the common tagset for a Polish-Ukrainian Parallel Corpus [8].<sup>2</sup> A relatively conservative extension of the tagset is proposed in Broda et al. 2.

At the time of its creation in 2004, the IPI PAN Corpus was the largest corpus of Polish, the only one that was linguistically annotated. However, there were two other independently developed corpora in public existence, namely, the PELCRA Corpus of Polish (<http://korpus.ia.uni.lodz.pl/>) and the PWN Corpus of Polish (<http://korpus.pwn.pl/>), as well as a non-public corpus developed at the Institute of Polish Language, Polish Academy of Sciences, and a small corpus of Polish developed in the 1960s (<http://www.mimuw.edu.pl/polszczyzna/pl196x/>). In 2007, the stakeholders in all large corpus efforts decided to combine their forces and a project was launched with the aim of merging the existing corpora and extending them to a 1-billion word *National Corpus of Polish* (henceforth, NCP or, in Polish, NKJP for *Narodowy Korpus Języka Polskiego*); see <http://nkjp.pl/>.

NCP is being annotated at various linguistic levels, including morphosyntax, named entities, syntax and limited word sense disambiguation. At the morphosyntactic level, NCP adopts the main assumptions of the IPI PAN Tagset, including the morphosyntactic definition of grammatical classes (e.g., a numeral is defined on the basis of its morphosyntactic behaviour, not in the

<sup>\*</sup> The work described in this article was carried out within the project *Narodowy Korpus Języka Polskiego* funded by the National Ministry of Science and Higher Education (grant number R1700303). This publication is supported by the European FP7 project *MONDILEX*.

<sup>1</sup> IPI PAN is the Polish acronym of *Instytut Podstaw Informatyki Polskiej Akademii Nauk* ‘Institute of Computer Science, Polish Academy of Sciences’, where the IPI PAN Corpus project was carried out.

<sup>2</sup> Some comparisons of the IPI PAN Tagset to MULTEXT-East [6, 7] tagsets may be found in Dimitrova et al. 5 and Derzhanski and Kotsyba 4.

traditional semantic terms) inspired by works of Zygmunt Saloni and his colleagues (see, e.g., Saloni and Świdziński 18 for a summary) and the detailed flexemic approach to the delimitation of grammatical classes (e.g., infinitive and finite verb are two separate classes, as they have different inflectional characteristics) following work by Janusz S. Bień (1991).

Nevertheless, some modifications of the IPI PAN Tagset were necessary, both for theoretical and for practical reasons. The tagset resulting from these modifications and used in the NCP annotation is called the NKJP Tagset. The aim of this paper is to describe and — where necessary — justify the differences between the IPI PAN Tagset ( $T_{\text{IPI}}$  in brief) and the NKJP Tagset (henceforth,  $T_{\text{NKJP}}$ ).<sup>3</sup>

## 2 Differences

Within NCP, a 1-million word corpus is being annotated manually. Manual annotation is one of the most expensive corpus building tasks, and one way to reduce the cost is to annotate the corpus automatically and only correct or disambiguate the automatic annotation manually. For the morphosyntactic annotation, a new version of the morphological analyser Morfeusz [20] is used in NCP, which is based on the linguistic data described in Saloni et al. 17. Some of the differences between  $T_{\text{IPI}}$  and  $T_{\text{NKJP}}$  stem from the availability of new linguistic information in this version of Morfeusz.

### 2.1 New non-inflecting classes

The main criterion for distinguishing grammatical classes in  $T_{\text{IPI}}$  is morphosyntactic, i.e., inflection and agreement. According to this criterion, all non-inflecting (f)lexemes fall into one bag, so an additional — distributional — criterion must be applied to distinguish, e.g., prepositions from conjunctions, and only a few traditional non-inflecting categories are posited in  $T_{\text{IPI}}$ . With the benefit of hindsight it seems that these classes are too coarse-grained, so four additional non-inflecting classes are carved out in  $T_{\text{NKJP}}$  from those present in  $T_{\text{IPI}}$ .

**Interjection** In principle any word may be used as an interjection, but for the purpose of  $T_{\text{NKJP}}$  interjection (*interj*) is understood rather narrowly. A segment (i.e., a word-level token receiving a morphosyntactic interpretation) is marked as an interjection, if one of the following holds:

- it may only be used as an interjection, e.g., segments such as *ach*, *och*, *oj*,
- if the same form has other interpretations, they are not related to the interjection use of that form, e.g., *a* (which may also be a conjunction or an abbreviation),
- it is onomatopoeic, e.g., *mu* or *kukuryku*.

Examples of segments which may be used interjectively but are not marked as interjections are *tak* ‘yes’ and *kurwa* ‘whore’.

**Subordinate conjunction** Where  $T_{\text{IPI}}$  only recognised conjunctions (Pol. *spójniki*),  $T_{\text{NKJP}}$  differentiates between coordinate conjunctions (Pol. *spójniki równorzędne*; *conj*), e.g., *i*, *lub* and *oraz*, and subordinate conjunctions (Pol. *spójniki podrzędne*), sometimes called complementisers (*comp*), e.g., *że*, *aby*, *bowiem*. It is clear that these two non-inflecting classes have very different syntactic behaviour.

**Predicative adjective** There are three adjectival classes in  $T_{\text{IPI}}$ : the usual inflecting adjectives (*adj*), ad-adjectival adjectives (*adja*), e.g., *polsko* ‘Polish’ in *polsko-niemiecki* ‘Polish-German’, and post-prepositional adjectives (*adjp*), e.g., *polsku* in *po polsku* ‘in Polish’. To these,  $T_{\text{NKJP}}$  adds

<sup>3</sup> A detailed presentation of  $T_{\text{NKJP}}$  may be found in the guidelines for annotators [14]; a stable version of these guidelines will be made available at <http://nkjp.pl/>.

another non-inflecting adjectival class, namely, the class of one-form lexemes consisting of forms which may only be used in predicative contexts (*adjc*)<sup>4</sup>, e.g., *zdrow* ‘healthy’ (cf. *On wydaje się zdrow* ‘He seems healthy’, but not \**zdrow człowiek* ‘healthy man’) or *ciekaw* ‘curious’ (e.g., *Jestem ciekaw* ‘I am curious’, but not \**ciekaw człowiek* ‘curious man’).

**Bound word** The segmentation principles of  $T_{\text{IPI}}$ , adopted in  $T_{\text{NKJP}}$ , rule that there are no segments containing spaces, so, e.g., *po trochu* ‘little by little’ cannot be treated as one segment. But the form *trochu* in contemporary Polish is a bound word, occurring in this construction only, so there is no reason to treat it as a noun or an adjective — any decision would have to be arbitrary. In  $T_{\text{NKJP}}$ , such indeterminate bound words are marked as **burk**, with the name of the class inspired by Derwojedowa and Rudolf 3.

## 2.2 Abbreviations

Abbreviations play an important role in the task of automatic segmentation of text into sentences: a full stop after an abbreviation may, but need not, also signal the end of a sentence, so each abbreviation should be marked for whether it requires a full stop or not.

Unlike in  $T_{\text{IPI}}$ , there is a separate abbreviation class (**brev**) in  $T_{\text{NKJP}}$ . There is a technical category associated with this class, “fullstoppedness”, which may take one of two values: **pun** (the abbreviation segment should be followed by a full stop) and **npun** (the segment does not have to be followed by a full stop).

The lemma for a segment marked as **brev** is the full dictionary form of the abbreviation, e.g., for *np* (*na przykład* ‘for example’), the tag should be **brev:pun** (*np* should be followed by a full stop) and its lemma should be **NA PRZYKŁAD**. For the segment *dr*, on the other hand, the lemma will always be **DOKTOR**, but the tag should be — in accordance with Polish orthographic rules — either **brev:pun** (e.g., in masculine accusative) or **brev:npun** (e.g., in nominative).

## 2.3 Adverbs and particles

In  $T_{\text{IPI}}$ , the class of particle-adverbs (**qub**), separate from the class of adverbs (**adv**), is considered an “else” class: if a segment does not fit any other class, it is annotated as **qub**. With the addition of several non-inflectional classes (see above), the need for such an “else” class diminishes, so in  $T_{\text{NKJP}}$  this class is defined in a constructive way. It may contain various particles (described in more detail in Przepiórkowski 14), the reflexive marker **SIĘ**, ad-numeral operators such as **OKOŁO** ‘around’ and **BLISKO** ‘almost’, and intensifiers such as **JEDYNIE** ‘only’ and **NAWET** ‘even’.

On the other hand, the class of adverbs is larger in  $T_{\text{NKJP}}$  than in  $T_{\text{IPI}}$ ; adverbs in  $T_{\text{NKJP}}$  are implicitly split into two subclasses:

1. de-adjectival or gradable adverbs, e.g., **DLUGO** ‘long’ and **BARDZO** ‘very’, which are always specified for degree (positive, **pos**, in case of de-adjectival adverbs which are not synthetically gradable); this subclass in  $T_{\text{NKJP}}$  corresponds closely to the whole **adv** class in  $T_{\text{IPI}}$ ;
2. traditional adverbs which are neither de-adjectival nor gradable, e.g., **GDZIE** ‘where’ and **WCZORAJ** ‘yesterday’; they are not marked for degree in  $T_{\text{NKJP}}$ ; in  $T_{\text{IPI}}$  they belong to the class **qub**.

## 2.4 Other differences

Apart from the substantial differences listed above, there is a number of minor differences between the two tagsets, mentioned below.

<sup>4</sup> The mnemotechnics of *adjc* is ‘adjective after the copula’, although such forms may occur in various predicative environments, not only copular, also as secondary predicates.

**Alien elements** There are two technical classes in  $T_{IPI}$  corresponding to various “alien” elements in texts, mostly foreign language expressions and passages: *xxs* for those segments which occupy a nominal position and, hence, may be assigned case, number and gender, and *xxx* for other foreign expressions. In  $T_{NKJP}$  there is only one “alien” class, *xxx*, for those segments which do not enter into relations with other (non-alien) segments in the sentence. This class is used mostly for annotating longer foreign expressions or whole passages in a foreign language. Other foreign segments, which enter into relations with other elements of the sentence, i.e., also those occupying nominal positions, should be marked in the usual way, as nouns, adverbs, etc.

**Collective numerals** Although some of the  $T_{IPI}$  publications listed above mention the class of collective numerals, *numcol*, that class was absent from the tagset actually used for the annotation of the IPI PAN Corpus and it is reintroduced in  $T_{NKJP}$ .

**Comparative degree** Since *comp* is used in  $T_{NKJP}$  as the name for the class of complementisers, the comparative degree is marked as *com* in this tagset, in contradistinction to *comp* used for that purpose in  $T_{IPI}$ .

### 3 Conclusion

Since  $T_{IPI}$  is relatively widely used, the modifications in  $T_{NKJP}$  were kept to the minimum and consist mainly in adding a few classes for non-inflecting elements and the removal of a hardly ever used class *xxs*. Both tagsets are well-documented, so we hope that an adaptation of existing tools to the new  $T_{NKJP}$  will turn out to be a manageable task. To further facilitate that task, the appendix below contains a specification of  $T_{NKJP}$ .

### Appendix: NKJP Tagset

In the following specification of  $T_{NKJP}$ , section [ATTR] lists all morphosyntactic categories and their possible values, while section [POS] specifies grammatical classes and categories appropriate for these classes. For example, any noun must be marked as *subst:number:case:gender*, where, e.g., *number* must be replaced by one of the possible values of this category, i.e., *sg* or *pl*. Hence, a full tag for the form *lampę* ‘lamp’ should be *subst:sg:acc:f*.

Some categories are optional for some classes, e.g., only some prepositions (such as *w*) have a vocalic (*we*) and a non-vocalic (*w*) form, so the segment *we* could be marked as *prep:acc:wok*, while the tag of, e.g., *na* could be *prep:acc*.

At the end of the specification some constraints are listed which should be respected by any tools used for the processing of this tagset.

All grammatical classes and categories not mentioned above are described in  $T_{IPI}$  publications listed in section 1.

## NKJP Tagset (version 1.0 of 23 June 2009)

[ATTR]

<i>number</i>	= <i>sg pl</i>
<i>case</i>	= <i>nom gen dat acc inst loc voc</i>
<i>gender</i>	= <i>m1 m2 m3 f n</i>
<i>person</i>	= <i>pri sec ter</i>
<i>degree</i>	= <i>pos com sup</i>
<i>aspect</i>	= <i>imperf perf</i>
<i>negation</i>	= <i>aff neg</i>
<i>accommodability</i>	= <i>congr rec</i>

accentability = akc nakc  
 post-prepositionality = npraep praep  
 agglutination = agl nagl  
 vocalicity = nwok wok

fullstoppedness = pun npun

## [POS]

adja =  
 adjp =  
 adjc =  
 conj =  
 comp =  
 interp =  
 pred =  
 xxx =  
 adv = [degree]  
 imps = aspect  
 inf = aspect  
 pant = aspect  
 pcon = aspect  
 qub = [vocalicity]  
 prep = case [vocalicity]  
 siebie = case  
 subst = number case gender  
 depr = number case gender  
 ger = number case gender aspect negation  
 ppron12 = number case gender person [accentability]  
 ppron3 = number case gender person [accentability] [post-prepositionality]  
 num = number case gender accommodability  
 numcol = number case gender accommodability  
 adj = number case gender degree  
 pact = number case gender aspect negation  
 ppas = number case gender aspect negation  
 winien = number gender aspect  
 praet = number gender aspect [agglutination]  
 bedzie = number person aspect  
 fin = number person aspect  
 impt = number person aspect  
 aglt = number person aspect vocalicity  
  
 brev = fullstoppedness  
 burk =  
 interj =

## This class should not appear in the results of manual annotation:

ign =

## Non-defeasible constraints:

##

```

## siebie --> base = siebie
## siebie --> case IN gen dat acc inst loc
## pant --> aspect = perf
## pcon --> aspect = imperf
## pact --> aspect = imperf
## ger --> gender = n
## depr --> number = pl
## depr --> gender = m2
## depr --> case IN nom voc acc
## numcol --> gender IN n m1
## aglt --> aspect = imperf
## bedzie --> aspect = imperf
## impt --> number:person IN sg:sec pl:pri pl:sec
## prep --> case IN nom gen dat acc inst loc

## Defeasible constraints:
##
## ger --> number = sg
## num --> number = pl

```

## Bibliography

- [1] Bień, J. S. (1991). *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, volume 383 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw. <http://bc.klf.uw.edu.pl/12/>.
- [2] Broda, B., Piasecki, M., and Radziszewski, A. (2008). Towards a set of general purpose morphosyntactic tools for Polish. In [9], pages 445–454.
- [3] Derwojedowa, M. and Rudolf, M. (2003). Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu. *Poradnik Językowy*, 5:39–49.
- [4] Derzhanski, I. and Kotsyba, N. (2009). Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In Garabík, R., editor, *Metalanguage and Encoding Scheme Design for Digital Lexicography: Proceedings of MONDILEX Third Open Workshop*, pages 9–26, Bratislava.
- [5] Dimitrova, L., Koseska-Toszewa, V., Derzhanski, I., and Roszko, R. (2009). Annotation of parallel corpora (on the example of the Bulgarian-Polish parallel corpus). In Shyrovkov, V. and Dimitrova, L., editors, *Organization and Development of Digital Lexical Resources: Proceedings of MONDILEX Second Open Workshop*, pages 47–54, Kiev. National Academy of Sciences of Ukraine, Ukrainian Lingua-Information Fund.
- [6] Erjavec, T., editor (2001). *Specifications and Notation for MULTEXT-East Lexicon Encoding*. Ljubljana.
- [7] Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1535–1538, Lisbon. ELRA.
- [8] Kotsyba, N., Shypnivska, O., and Turska, M. (2008). Principles of organising a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). In [9], pages 475–484.
- [9] Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., and Trojanowski, K., editors (2008). *Intelligent Information Systems*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.
- [10] Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej. To appear.
- [11] Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.
- [12] Przepiórkowski, A. (2004a). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.



- [13] Przepiórkowski, A. (2004b). *Korpus IPI PAN. Wersja wstępna*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- [14] Przepiórkowski, A. (2009). *Zasady znakowania morfosyntaktycznego w NKJP*. Version 1.03 of 29 June 2009.
- [15] Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pages 33–40, Budapest.
- [16] Przepiórkowski, A. and Woliński, M. (2003b). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116.
- [17] Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007). *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw.
- [18] Saloni, Z. and Świdziński, M. (2001). *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 5th edition.
- [19] Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII–XXIII:39–55.
- [20] Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Kłopotek, M. A., Wierzchoń, S. T., and Trojanowski, K., editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin.

# Morphosyntactic Specifications for Polish and Lithuanian.

## Description of Morphosyntactic Markers for Polish and Lithuanian Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)<sup>\*</sup>

Danuta Roszko, Roman Roszko

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw  
danuta.rozsko@ispan.waw.pl    roman.rozsko@ispan.waw.pl

**Abstract.** This is a follow-up of the earlier paper where Roman Roszko [18] presented the foundations for a scientifically-rigorous classification of lexemes into classes (parts of speech). Then he presented and analysed a portion of a new and already widespread classification into parts of speech authored by Zygmunt Saloni [19]. The first stage involved development of morphosyntactic specifications for Polish nouns. Given the innovative subdivision into parts of speech, differing from traditional grammatical descriptions, and the existence of morphological, semantic and syntactic subcategories not found in other languages, the author expanded the number of markers for Polish nouns. The following categories were taken as the new morphosyntactic specifications: human, animate, post-prepositionality, stressability, depreciativeness. The category of gender has been rearranged. The author did not follow the elaborate gender system proposed by Saloni [19] and retained the subdivision into masculine, feminine and neutral gender, as used in MULTEXT-East [5]. Instead, he proposed new characteristics, human and animate, as independent, stand-alone attributes. In this paper the authors continue to discuss the questions addressed in [18]. They expand the description of morphosyntactic characteristics for nouns in Polish and Lithuanian, in accordance with the conventions adopted in MULTEXT-East for morphosyntactic specifications. The aim of the analysis is to create a joint morphosyntactic description for three languages: Polish, Bulgarian and Lithuanian, which are to be used in the parallel Polish-Bulgarian-Lithuanian corpora, now under development<sup>1</sup>, and in electronic Bulgarian-Lithuanian and Bulgarian-Polish-Lithuanian dictionaries [3].

The next step in the process will be to develop morphosyntactic specifications for the remaining parts of speech in Polish and Lithuanian.

**Keywords:** POS : parts of speech, nomen, annotation of parallel corpus, Polish, Lithuanian.

## 1 Introduction

The problem involving the degree of morphologisation of various meanings in natural language has a significant bearing on the grammatical description of that language. A high number of morphological categories, their transparency and absence of exceptions greatly facilitate such a description. However, Polish is not one of the languages where the degree of formalisation of meanings would facilitate grammatical description. On the other hand, Lithuanian, particularly in the nomen class, retains a number of archaic characteristics and shuns innovations, which facilitates a morphosyntactic description. The categories of Lithuanian nouns seem to have remained unchanged for centuries, keeping clear-cut formal characteristics.

---

<sup>\*</sup> The study and preparation of these results have been supported by the EC's Seventh Framework Programme [FP7/2007–2013] under the grant agreement 211938 MONDILEX.

<sup>1</sup> The first Bulgarian-Polish-Lithuanian (BG-PL-LT) corpus contains more than 3 million words and comprises two corpora: parallel and comparable. The BG-PL-LT parallel corpus contains more than 1 million words. A small part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of official documents of the European Union available through the Internet. The texts (fiction) in other languages translated into Bulgarian, Polish, and Lithuanian form the main part of the parallel corpus [4].

**Diachronic aspects** The Lithuanian nomen forms are typical in their archaic nature. Let us just note that all Lithuanian nomen forms do retain endings whereas even classic Latin shows reduction of inflection, cf. Lit. *vyr-a-s* [vīrās] and Lat. *vīr-(ø)* [vīr]. When compared with Slavic languages, Lithuanian notably retains old forms of noun inflection whereas those inflection patterns have been considerably changed in Polish, as in the old inflection with consonants: Lit. *ses-uo* (nom.sg.), *ses-er-s* (gen.sg.) and Pol. *siostr-a* (nom.sg.), *siostr-y* (gen.sg.). Contemporary Lithuanian still retains the old inflection with consonants whereas contemporary Polish no longer has it. Another example is the development of old declension patterns in Polish and Lithuanian. Tables 1.-2. present two old inflection patterns with a short *-o-* [ǒ] and a short *-u-* [ũ].

Table 1. Noun declension. Pattern with short *-o-* [ǒ] (so-called Declension I). Proto-Slavic, contemporary Polish and contemporary Lithuanian *вѣлкѹ* – *wilk* – *vilkas* ‘wolf’

Case	Proto-Slavic[10, p. 223]	Polish	Lithuanian
	Singular		
nom.	вѣлк-ѣ	wilk-	vilk-as
gen.	вѣлк-а	wilk-a	vilk-o
dat.	вѣлк-у	wilk-owi	vilk-ui
acc.	вѣлк-ѣ	wilk-a	vilk-ą
instr.	вѣлк-омѣ	wilk-iem	vilk-u
loc.	вѣлк-ѣ	wilk-u	vilk-e
voc.	вѣлк-е	wilk-u	vilk-e
Plural			
nom.	вѣлк-и	wilk-i	vilk-ai
gen.	вѣлк-ѣ	wilk-ów	vilk-ų
dat.	вѣлк-омѣ	wilk-om	vilk-ams
acc.	вѣлк-ы	wilk-i	vilk-us
instr.	вѣлк-ы	wilk-ami	vilk-ais
loc.	вѣлк-ѣхѣ	wilk-ach	vilk-uose
voc.	вѣлк-и	wilk-i	vilk-ūs

Table 2. Noun declension. Pattern with short *-u-* [ũ] (so-called Declension II). Proto-Slavic, contemporary Lithuanian and contemporary Polish *сынѹ* – *syn* – *sūnus* ‘son’

Case	Proto-Slavic[10, p. 226]	Polish	Lithuanian
	Singular		
nom.	сын-ѣ	syn-	sūn-us
gen.	сын-у	syn-a	sūn-aus
dat.	сын-ови	syn-owi	sūn-ui
acc.	сын-ѣ	syn-a	sūn-ų
instr.	сын-ѣмѣ	syn-em	sūn-umi
loc.	сын-у	syn-u	sūn-nuje
voc.	сын-у	syn-u	sūn-au
Plural			
nom.	сын-ове	syn-owie (/ -y)	sūn-ūs
gen.	сын-овѣ	syn-ów	sūn-ų
dat.	сын-ѣмѣ	syn-om	sūn-ums
acc.	сын-ы	syn-ów	sūn-us
instr.	сын-ѣми	syn-ami	sūn-umis
loc.	сын-ѣхѣ	syn-ach	sūn-uose
voc.	сын-ове	syn-owie (/ -y)	sūn-ūs

As shown in Tables 1.-2., Lithuanian continues to have old inflection patterns with short *-u-* and short *-o-*, whereas Polish has lost the clarity of the former inflection with short *-u-* (which merged with the pattern with short *-o-*).

**Morphosyntactic descriptions for Polish** The system of morphosyntactic markers developed for the Polish language at the Institute of Computer Science, Polish Academy of Sciences (Pol. IPI PAN) (A. Przepiórkowski, M. Woliński: [15], [16], [22], [12], [13]/[14]), is based on a sound methodological foundation comprising linguistic work by authors such as Z. Saloni, M. Świdziński, J. S. Bień. It is thanks to this foundation that the IPI PAN's tagset goes beyond the fossilised traditional framework dating back to Aristotle. On the other hand, the MULTEXT-East tagset, which serves as a point of reference here, is based on the traditional subdivision into parts of speech (this is why, among others, pronouns have been singled out as a part of speech). MULTEXT-East also replicates the deeply-rooted traditional stereotypes regarding the subdivision into parts of speech as well as linguistic categories or morphological subcategories. Moreover, spaced versus unspaced spelling may decide whether or not a category is identified (e.g. articles, interpreted differently for various languages).

The IPI PAN tagset is used not only in the IPI PAN Corpus ([14]/[13]) but also, in a somewhat modified version, in the National Corpus of Polish (<http://nkjp.pl/>) and a few other projects, some of which are conducted outside IPI PAN, e.g. in the Polish WordNet project (M. Piasecki: [2], [11]) developed in Wrocław or in Morfologik (M. Miłkowski: <http://morfologik.blogspot.com/>; <http://nlp.ipipan.waw.pl/NLP-SEMINAR/061016.pdf>) and other. These facts clearly indicate that the IPI PAN tagset has become a benchmark for the Polish language.

Consequently, the aim of this series of papers is neither to revise the good and highly refined IPI PAN tagset nor to replace it with a new tagset for Polish. The issue in question is what kind of compromise should be sought when developing a joint tagset to be used for simultaneous description of three languages in the BG-PL-LT parallel corpus. For some reasons the MULTEXT-East tagset (developed previously for many languages) has been selected as the leading one for this corpus [4]. Therefore, the aim of this series of papers is to provide a theoretical study of various categories of Polish and Lithuanian, to set priorities (e.g. morphological, semantic, syntactic) in identifying various meanings and to provide a classification of morphosyntactic phenomena which does not contradict the MULTEXT-East standard and does not deviate too strongly from the IPI PAN tagset. In a sense, we seek to establish correspondence/consistency between the two tagsets. If such correspondence proved impracticable, we would confine ourselves to specifying differences (not only significant ones) between the MULTEXT-East tagset and the IPI PAN tagset.

It cannot be excluded that due to the obvious difficulties in achieving consistency of the intertagset the BG-PL-LT corpus will use the IPI PAN tagset for Polish and its modification for Lithuanian. This solution would certainly necessitate a list of more or less close equivalents for the two tagsets: a tagset for Bulgarian on the one hand, and the IPI PAN tagset on the other (for Polish and an extended version for Lithuanian). For some reasons a review of the tagset for Bulgarian is not planned at this stage of work.

It is important to emphasise that only a coherent tagset for a parallel multilingual corpus 1) allows complete linguistic confrontation, 2) enables identification of linguistic facts, 3) enables a search based on pre-defined unambiguous morphosyntactic characteristics.

**Morphosyntactic descriptions for Lithuanian:** as a basis for morphosyntactic descriptions of Lithuanian serve the Academic grammar of the Lithuanian language [1] and the Functional grammar of Lithuanian [21]. A tool for morphosyntactic annotation for Lithuanian — MorfoLema — has been created by Vytautas Zinkevičius in Centre of Computational Linguistics of Vytautas Magnus University (Lithuania) [23]. The program MorfoLema can perform a morphosyntactic analysis and generate forms of Lithuanian words based on user's morphosyntactic characteristic. For disambiguation the MorfoLema uses "Two-level morphology" method of Kimmo Koskeniemi [8]. The next step of the development of a system for morphological annotation (*Morfologinis anotatorius* = tagger for Lithuanian: [http://donelaitis.vdu.lt/main.php?id=4&nr=7\\_1](http://donelaitis.vdu.lt/main.php?id=4&nr=7_1)) has been realised by Vidas Daudaravičius and Erika Rimkutė. Vidas Daudaravičius has created disambiguation tools for the *Morfologinis anotatorius*. More information about the *Morfologinis anotatorius* and used set of tags we can find on [http://donelaitis.vdu.lt/main.php?id=4&nr=7\\_1](http://donelaitis.vdu.lt/main.php?id=4&nr=7_1) (in Lithuanian). (The names of tags are in Lithuanian, because the authors of the *Morfologinis anotatorius*

didn't use English terms.) It is possible to perform online a morphosyntactic analysis through the web-page [http://donelaitis.vdu.lt/main.php?id=4&nr=7\\_2](http://donelaitis.vdu.lt/main.php?id=4&nr=7_2). The results are visualized on the screen, and it is possible to receive the result as a file.

**Our aim here is** to adjust the grammatical description of Polish and Lithuanian to the existing description within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004), developed for a larger group of languages (eleven, to date). Consequently, it must be immediately noted that one cannot talk about classes or parts of speech as a universal phenomenon, common to all natural languages [7]. This suggests that a description of morphosyntactic characteristics for multiple languages is difficult and calls for some satisfactory compromise. Let us point out that when making a subdivision into parts of speech, we must make the following important realisations: 1. What exactly is it that we are subdividing? and 2. What is the goal of this subdivision? A classification into parts of speech which is to be created should meet the criteria of scientific rigour. Therefore, a dichotomous subdivision (into two) is required at each stage. Also, clear, non-contradictory and uniform subdivision criteria are required. A criterion that has been already used at one level should not be used again at a lower level for a narrower set of lexemes. Moreover, the resulting subdivision should be easily verifiable, which means that, above all, it should cover the entire vocabulary.

**Orthographical word vs lexeme** One of the problems for building a description of morphosyntactic characteristics is an unclear notion of 'word' which, as apposed to morpheme, is neither stable nor fixed. 'Word' continues to have arbitrary definitions. As a result, if a definition of 'word' is adopted, this is likely to exert significant influence on the final shape of such classification. In [18] we present a few meanings of 'word': phonological word, orthographical word, textual word, grammatical word (dictionary word, or lexeme). Again, for well-known reasons, we will discuss the problem of an orthographic word here.

There remain significant differences between Polish and Lithuanian spelling rules. Differing rules may have marked consequences for the development of a comparable morphosyntactic description for the two languages. For this reason, we decided to show certain differences which are important for the morphosyntactic description and which occur between orthographic norms for the two languages. Spaced or unspaced spelling is the most important factor which may result in different morphosyntactic characteristics being ascribed to otherwise equivalent Polish and Lithuanian forms. To prevent the adverse impact of spaced vs. unspaced spelling on the equivalence between the morphosyntactic descriptions of the two languages, we allow, in justified cases, the following: 1) splitting single orthographic words, or 2) merging two (or more) orthographic words. If a lexeme is not to be interpreted as a form which is congruent with the orthographic word, the meaning must be referred to as the semantic basis for identifying lexemes.

As a reminder: orthographical word — a string of written text, delimited by spaces; it is an artificial creation and, as such, cannot represent the basis of classification into parts of speech. Let us consider the two functionally close examples quoted in [18]: *\_na\_pewno\_* and *\_naprawde\_* or the recent spelling reform which recommends that *nie* with the so-called adjectival participles should be spelt as a single word, confirm the high degree of arbitrariness in spelling rules.

In Lithuanian, *ne* (equivalent of Polish *nie*) is spelt together with participles, gerunds and verbs, as in the example below:

Table 3

Lithuanian	Polish
<i>nedirbu</i> (present)	<i>nie</i> pracuję (non-past)
<i>nebuvaу padaręs</i> (past perfect)	<i>nie</i> zrobięm
<i>neparašęs</i> (past participle)	ten który <i>nie</i> napisał
<i>nedirbus</i> (past gerund)	po tym jak <i>x nie</i> pracował, <i>P(y)</i>

It is important to note the migration of Lithuanian *si*. If a word is not prefixed (has no prefix), then the morpheme *si* takes a position typical of an agglutinated part: *sveikina-si*. However, if a

word is prefixed (has a prefix), then the morpheme *si* takes a position between the prefix and the root: *at-si-sveikino*. The Lithuanian *si* corresponds with the Polish *się*<sub>1</sub> (for discussion of *się*<sub>1</sub> and *się*<sub>2</sub>, see Saloni in: [19] and [18]). The Lithuanian *si* is not equivalent to the Polish *się*<sub>2</sub>. Based on J. Tokarski's a tergo dictionary [20] and Saloni's grammatical dictionary (Saloni in: [19, p. 19–21]), some cases of unspaced spelling for two words, typical of the Polish language, are given below.

Particles *ć*, *że/ż*, *li*, e.g. *pójdę-ć*, *dasz-li*, *już-że*, agglutinates or abbreviated personal forms of praesens for the verb *być* 'to be': *m/em*, *ś/eś*, *śmy/eśmy*, *ście/eście*, e.g. *ja-m*, *że-ś*, *skąd-eśmy*, *gdzie-ście*, conditional mode operator *by*, e.g. *jakkolwiek-by* or *jakkolwiek-by-m* (with the agglutinate *m*). There is also another form of the pronoun *on* 'he', common for the genitive and accusative case, spelt unspaced: *ń* (*do-ń*, *za-ń*).

In opposition to the above, there is a case of separate (spaced) spelling of inflective forms, for instance *będę czytać/będę czytał*. Again, let us refer to the aforementioned examples of *na\_pewno* and *śmiać się* (*śmiejący się*). While the Polish compound form *na\_pewno* should be treated as a separate lexeme, the forms with *się*, even the ones which do not occur without *się* (such as *bać się*), are mere combinations of two lexemes. This is determined by their semantic properties. Saloni (in: [19, p. 21]) mentions more examples of so-called compound lexemes such as *po polsku* which are regularly derived from adjectives ending with *-ski*, *-cki*, *-dzki*.

It is interesting to consider the Lithuanian equivalents to the aforementioned Polish examples of spaced spelling of two forms which are interpreted as a single lexeme:

Pol. *na\_pewno* — Lit. *tikrai*

Pol. *po polsku* — Lit. *lenkiškai*

**For foundations for identifying parts of speech** in Polish, see: [7] and [18]. Let me just recall the most common criteria applied for identifying parts of speech: ontological/intuitive, morphological, semantic and syntactic. The aforementioned subdivision of Polish lexemes into parts of speech, as proposed by Saloni, is not consistent (nor are the majority of such subdivisions). It seems that Saloni relies most heavily on the criterion of morphology (inflection). When inflection fails to provide an answer, secondary criteria, semantic and syntactic ones, are employed.

## 2 Noun

This paper assumes the definition of noun as proposed by Saloni (for the Polish language). Saloni (in: [19, p. 29]) identifies noun lexemes based on morphological criterion in the inflective category of case (= declension), inflective category of number and selective (i.e. not inflective) category of gender. Saloni includes some forms traditionally considered to be pronouns onto the class of nouns: *ja* 'I', *ty* 'you', *on* 'he / she / it / they', *my* 'we', *wy* 'you', *kto* 'who', *ktoś* 'someone', *ktokolwiek / ktośkolwiek* 'anyone', *co* 'what', *coś* 'something', *cokolwiek / cośkolwiek* 'anything', *cóż* 'whatever', *nic* 'nothing', *się*<sub>1</sub> 'self', *się*<sub>2</sub> 'self', *wszyscy* 'everyone', *toto* 'this thing', *niecoś*, *śmo*, *wasze* 'your/yours' and other, a total of ca. 40 forms. Saloni's proposed interpretation is right. It has simple and obvious semantic and syntactic foundations. However, given that our description of morphosyntactic characteristics for Polish and Lithuanian is expected to conform with the description used in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004, we need to retain the assumptions adopted for MULTEXT-East Morphosyntactic Specifications.

### 2.1 Uninflected words

Saloni formally eliminates the class of uninflected nouns. Based on syntactic criteria and analogy to other, typical nouns (traditionally referred to as inflected nouns) he builds a paradigm for all nouns traditionally described as uninflected, as in the following example for *emu*: Table 4.

Table 4. “Uninflected” Polish nouns.

Case	Singular	Plural
nominative	emu	emu
genitive	emu	emu
dative	emu	emu
accusative	emu	emu
instrumental	emu	emu
locative	emu	emu
vocative	emu	emu

An academic grammar of Lithuanian [1] identifies a small number of forms which are described as uninflected, e.g. *bolero* ‘bolero’, *bruto* ‘gross’, *kredo* ‘creed’, *taksi* ‘taxi’. The way Saloni did for Polish, we suggest assuming that Lithuanian has no uninflected nouns and they should be ascribed the following paradigm:

Table 5. “Uninflected” Lithuanian nouns.

Case	Singular	Plural
nominative	bolero	bolero
genitive	bolero	bolero
dative	bolero	bolero
accusative	bolero	bolero
instrumental	bolero	bolero
locative	bolero	bolero
innesive	bolero	bolero
vocative	bolero	bolero

This will ensure full conformity with the definition of noun.

Moreover, when Polish and Lithuanian borrow new lexemes, they usually ascribe their own paradigms to such lexemes, for instance: Pol. *komputer* (nom.sg.), *komputera* (gen.sg.), Lit. *kompjuteris* (nom.sg.), *kompjuterio* (gen.sg.). The so-called uninflected nouns appear as a result of difficulties in adaptation of phonetic groups found at word end. In Polish the trend to eliminate formal differentiation of cases is manifested in names of occupations, positions, titles referring to women, for instance:

*pani doktor* (nom.sg.) *powiedziła* - ‘the [female] doctor said’  
*pani doktor* (dat.sg.) *powiedziano* - ‘the [female] doctor was told’

A similar phenomenon in Lithuanian is sporadic.

## 2.2 Case

The category of case is identified on the basis of syntactic characteristics imposed on nouns, usually by verbs or prepositions. The following cases exist in the Polish language:

Table 6

Case	Examples
nominative	dom domy
genitive	domu domów
dative	domowi domom
accusative	dom domy
instrumental	domem domami
locative	(preposition +) domu (preposition +) domach

Locative case always occurs with a preposition, for instance *w domu* ‘at home’, *o domu* ‘about home’. This is a not a stand-alone case.

The following cases exist in the Lithuanian language:

Table 7

Case	Examples
nominative	namas namai
genitive	namo namų
dative	namui namams
accusative	namą namus
instrumental	namu namais
locative	name namuose
inessive	name namuose

When we look at data from Tables 5. and 6., we will see that the number of cases is greater for Lithuanian. Lithuanian has a form of inessive. The remaining cases overlap between Polish and Lithuanian. Notably, the locative in Lithuanian takes no preposition, as in Pol. *w domu* – Lit. *namie / namuose*.

Lithuanian has two lative cases in vestigial form: illative (who into? what into? Where to?), allative (who to? towards what?) and the so-called local case, adessive (near whom? near what?). Out of those illative seems most robust, for instance *Lietuvon* (sg.) ‘to Lithuania’, *vežiman* (sg.) ‘onto the cart’, *širdin* (sg.) ‘into the heart’, *turgun* (sg.) ‘to the fair’, *laukuosna* (pl.) ‘to the field.’ While illative has been largely superseded by the structure *į + accusative*, it is still used. One might expect illative to be used largely in literature but data from the largest corpus of Lithuanian Tekstynas (<http://donelaitis.vdu.lt/>) show that even some illative forms are found more frequently in nation-wide and regional press and popular publications than in belles-lettres. The remaining cases mentioned here are more commonly found in dialects. Literary language uses some forms occasionally, e.g. *velniop* ‘to the devil’, *vakarop* ‘near the evening’.

Table 8. presents a summary of cases for the two languages concerned together with the so-called question words which decipher the meaning of cases.

Table 8

Case	Polish		Lithuanian	
nominative	mianownik	kto, co	vardininkas	kas
genitive	dopełniacz	kogo, czego	kilmininkas	ko
dative	celownik	komu, czemu	naudininkas	kam
accusative	biernik	kogo, co	galininkas	ką
instrumental	narzędnik	kim, czym	įnagininkas	kuo
locative	miejscownik	o kim, o czym	vietininkas	kur
inessive	—	—	name	kame

**Vocative** In the linguistic tradition vocative is considered as one of the cases. However, the use of vocative in a text does not confirm the existence of any government imposed on the vocative form by a verb or a preposition. Therefore, vocative should be viewed as a separate category. However, in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) vocative is classified as one of the cases, which is why (according to the fallacious tradition) vocative is included here as another case in Polish and Lithuanian, as in the example below:

Table 9

Case	Polish		Lithuanian	
		Example		Example
	proper name			
[nominative]	[mianownik]	[ <i>Romek</i> ]	[vardininkas]	[ <i>Romukas</i> ]
vocative	wołacz	<i>Romku</i>	šauksmininkas	<i>Romuk</i>
	common name			
[nominative]	[mianownik]	[ <i>dom</i> ]	[vardininkas]	[ <i>namas</i> ]
vocative	wołacz	<i>domu</i>	šauksmininkas	<i>name</i>



### 2.3 Number

The identification of number is based on the semantic difference between a single object: (distributive = non collective) set of objects, e.g. *dom* ‘home’ : *domy* ‘homes’. The grammatical category of number sometimes slightly deviates from the aforementioned semantic opposition. Some nouns do not offer such a distinction, for instance, the so-called plurale tantum:

(a) Polish *nożyczki*, ‘scissors’ *drzwi* ‘door’, *parzystokopytne* ‘even-toed ungulates’, *małżonkowie* (= *małżonka* ‘wife’ + *małżonek* ‘husband’), *narzeczeni* (*narzeczona* ‘fiancée’ + *narzeczony* ‘fiancé’), *Wadowice*, *Kielce*, *Katowice* (proper names).

(b) Lithuanian *žirklutės* ‘scissors’ *durys* ‘door’, *skeltanagiai* ‘even-toed ungulates’, *sutuoktiniai* (= *žmona* ‘wife’ + *vyras* ‘husband’), *sužadėtiniai* (*sužadėtinė* ‘fiancée’ + *sužadėtinis* ‘fiancé’), *Žiemieji Paneriai*, *Prienai*, *Zarasai* (proper names).

The number of plurale tantum in Lithuanian is decidedly higher, cf. [17]) for a discussion of this subject.

In fact, Polish and Lithuanian has no singulare tantum nouns, yet they are sometimes mentioned in literature. One example of singulare tantum is Polish *fizyka* ‘physics’ / Lithuanian *fizika* ‘physics’ or *pierze* ‘plumage’ / Lithuanian *jaunimas* ‘young people / youth’ (the so-called collective nouns), Polish *miłość* ‘love’ / Lithuanian *meilė* ‘love’. However, for any singulare tantum a plural form may be created and a use for it may be found, as noted by Saloni (in: [19]).

There is no need to introduce the dual number in Polish. Contemporary Polish has few dual forms (e.g. in instrumental or locative case) for selected paired bodily parts such as hands, eyes or ears. They should be treated as variants of plural forms: *rękami/rękoma* (instr.), *oczami/oczyma* (instr.), *rękach/ręku* (loc.). The relics of the dual form are visible in proverbs in the two languages, for instance Pol. *Dwie niewieście narobią hałasu w mieście*. ‘Two women will make a noise in the town’ Lit. *Su durnu vis du turgu*. ‘You always bargain twice with a fool.’ Table 10. presents the status quo for the category of number for the two languages concerned.

Table 10

Number	Polish		Lithuanian	
		Example		Example
singular	pojedyncza	dom	vienaskaita	namas
plural	mnoga	domy	daugiskaita	namai

### 2.4 Gender

Gender is identified on the basis of syntactic properties associated with the requirement that a specific form must occur next to a word that combines with a noun. The category of gender in nouns is selective. All nouns in Polish and Lithuanian have a fixed gender. A small group of Polish nouns may have no definite gender. The number of gender-undefined forms in Lithuanian is negligible due to the clear declension forms.

Traditionally, the following genders have been distinguished: masculine, feminine, neuter. Polish has all of these three genders, for instance: *dom* ‘home’ (masculine), *książka* ‘book’ (feminine), *dziecko* ‘child’ (neuter). Lithuanian has no neuter gender.

We do not think it is valid to introduce a separate category for nouns traditionally termed as bi-gendered, such as Polish *ciapa* ‘slowcoach’, *łamaga* ‘butterfingers’, *niedzara* ‘fumbler’, or Lithuanian *dabita* ‘beau’, *mémė* ‘tight-lipped’, *valkata* ‘wałęsa’. The basis for distinguishing bi-genderedness could only be semantic in this case. However, we view these forms as homonymous:

(a) for Polish: *łamaga* described as masculine-human, and *łamaga* as feminine.

(b) for Lithuanian: *dabita* described as masculine, and *dabita* as feminine.

Notably, according to traditional descriptions the group of bi-gendered nouns also includes forms such as *psycholog* ‘psychologist’, *sędzia* ‘judge’ and many other. Such examples can be also described as homonymous forms of masculine-human and feminine. Recently, there has been a strong trend in Polish towards providing a formal distinction between such forms: *psycholog* — masculine-human and *psycholożka* — feminine, *sędzia* — masculine-human and *sędzina* — feminine.

In recent years Lithuanian has shown quite the opposite trend, i.e. equalisation of formal differences between such masculine and feminine forms. We see this phenomenon as an obvious influence of the so-called Western languages. The source of those changes lies in the consistent formal differentiation of surnames (bearing administrative consequences, i.e. names written in passports), for instance *Marcinka* (masculine), *Marcinkienė* (feminine, a married woman, wife of Marcinka), *Marcinkaitė* (feminine, daughter of Mr. and Mrs. Marcinkai).

**Gender in Polish** The distinction into genders is specific to Polish nouns in singular whereas the traditional notion of masculine, feminine and neuter is blurred in plural.

Apart from innovations in plural there are new phenomena in Polish which are characteristic of some classes of singular masculine and neuter nouns. The nature of those new phenomena is syntactic. Therefore, much as gender, the new phenomena are associated with the requirement to adopt a particular form in adjacency to words that combine with nouns.

In the Polish linguistic tradition, initiated by Witold Mańczak [9], three masculine genders are distinguished. They are also adopted by Saloni as three subgenders (in: [19]). In our view, an alternative solution is possible once we have introduced new categories, i.e. «Human» and «Animate». In that case the gender classification adopted within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) will be retained.

In fact, one might talk about two groups of nouns in plural. The first group comprises masculine nouns that are human and pluralia tantum that are human. The second group covers all other plural forms of masculine, feminine and neuter nouns as well as pluralia tantum that are not human.

Table 11

Type	Gender	Number	Case	Human	Animate	Examples		
comon	masculine	singular	nominative genitive dative accusative accusative accusative instrumental locative vocative	+	+	profesor profesora profesorowi profesora  profesorem profesorze profesorze	pies psa psu psa  psem psie psie	dom domu domowi  dom domem domu domu
comon	–	plural	nominative  nominative genitive dative accusative accusative instrumental locative vocative  vocative	+		profesorowie / / profesorzy  profesorów profesorom profesorów  profesorami profesorach profesorowie / / profesorzy	  psy psów psom  psy psami psach  psy	  domy domów domom  domy domami domach  domy

As shown in the tables 11–12, the «Human» category is visible in accusative singular and in nominative, accusative (and vocative) plural whereas «Animate» is visible only in accusative singular.

As far as lexemes of neuter gender are concerned, a split in collocation occurs only when such lexemes combine with numerals. The first group is small and has the following collocation pattern with numerals: *czworo szczeniąt*, *troje dzieci*. The second group is much more numerous

and strongly supersedes the former group. Examples of collocations in this case are: *cztery pola, trzy lata*.

Table 12

Type	Gender	Number	Case	Human	Animate	Examples		
proper	masculine	singular	nominative	+	+	Roman	Burek (dog)	Płock
			genitive			Romana	Burka	Płocka
			dative			Romanowi	Burkowi	Płockowi
			accusative			Romana		
			accusative				Burka	
			accusative					Płock
			instrumental			Burkiem	psem	Płockiem
			locative			Romanie	Burku	Płocku
		vocative	Romanie	Burku	Płocku			
proper	–	plural	nominative	+		Romanowie /		
						/ Romany		
			nominative			Burki	Płocki	
			genitive			Romanów	Burków	Płocków
			dative			Romanom	Burkom	Płockom
			accusative			Romanów		
			accusative				Burki	Płocki
			accusative				Burkami	Płockami
instrumental	Romanami	Burkami	Płockami					
locative	Romanach	Burkach	Płockach					
locative								
vocative	Romanowie /							
	/ Romany							
vocative		Burki	Płocki					

**Gender in Lithuanian** In contrast with Polish, the distinction between genders in Lithuanian is not problematic. Both singular and plural retains the distinction between masculine and feminine. There is no need to distinguish subcategories of human and animate for Lithuanian because such a distinction does not occur at the formal level. Table 13 presents the category of gender in Lithuanian, with examples.

Table 13

Type	Gender	Number	Case	Examples
proper	masculine	singular	nominative	namas
			genitive	namo
			dative	namui
			accusative	namą
			instrumental	namu
			locative	name
			inessive	name
			vocative	name
proper	masculine	plural	nominative	namai
			genitive	namų
			dative	namams
			accusative	namus
			instrumental	namais
			locative	namuose
			inessive	namuose
			vocative	namai

Type	Gender	Number	Case	Examples
proper	feminine	singular	nominative	dantis
			genitive	danties
			dative	dančiui
			accusative	danti
			instrumental	dantimi
			locative	dantyje
			inessive	dantyje
			vocative	dantie
proper	masculine	plural	nominative	dantys
			genitive	dantų
			dative	dantims
			accusative	dantis
			instrumental	dantimis
			locative	dantyse
			inessive	dantyse
			vocative	dantys

One characteristic of Lithuanian is its great freedom in forming feminine equivalents of masculine forms. This is particularly important for names of occupations, positions, titles held or performed by women, for instance:

*profesorius* ‘professor’ — *profesorė* ‘female/woman professor’  
*ministras* ‘minister’ — *ministrė* ‘female/woman minister’  
*statybininkas* ‘builder’ — *statybininkė* ‘female/woman builder’

The list of such Lithuanian masculine and feminine forms is not finite. Notably, masculine and feminine forms differ only in terms of inflection whereas the stem is shared.

Lithuanian seems to have syntactic consequences when some nouns collocate with numerals (cf. p. 153 above, a close phenomenon in Polish). Lithuanian pluralia tantum forms call for collective numerals, e.g.:

*devyni namai* ‘9 houses’

and

*devyneri metai* (pluralia tantum) ‘9 years’

Analysis of the electronic corpus of Lithuanian, Tekstynas (<http://donelaitis.vdu.lt/>) shows that this category is likely to lose importance. Some uses of *devyni metai* instead of the correct *devyneri metai* have already been recorded. In contrast with the analogous phenomenon in Polish, (cf. p. 153 above) the disappearance of this category in Lithuanian is at an early stage, if at all.

## 2.5 Depreciativeness

The category of depreciativeness is identified on the basis of syntactic properties associated with the requirement for a word to occur in a particular form next to a word that combines with a noun. In two cases in plural, nominative and vocative, (the latter being always identical with nominative) some masculine nouns have two forms that are used in parallel, e.g. *chłopacy* ‘boys’ and *chłopaki* ‘[contemptuously about] boys’. If it were not for syntactic differences associated with the use of one or the other form, one could talk about the existence of multivariants and so another subcategory of depreciativeness would not need to be introduced:

*To są silni chłopacy.* i *To są silne chłopaki.* (both: ‘These are strong boys’)

We agree with Saloni that the subcategory of depreciativeness is an inflective one and one that enforces differing syntactic consequences.

As a rule, non-depreciative forms are neutral and considered to be basic. Depreciative forms should be seen as negatively marked, used to show a certain degree of disrespect. As usual in such cases, there are some exceptions such as neutralisation or even a reversal of marking, as described by Saloni.

The table below is an updated version of the relevant elements from Table 11:

Table 14

Type	Gender	Number	Case	Human	Animate	Depreciativeness	Examples
comon	–	plural	nominative	+			profesorowie /
				+			/ profesorzy
			nominative			+	profesory
			vocative	+			profesorowie /
				+			/ profesorzy
			vocative			+	profesory

The category of depreciativeness occurs in the group of masculine nouns which have the attribute of «Human».

## 2.6 Uniformism

The category of uniformism, as distinguished by Saloni, is associated with stylistic differentiation of selected feminine nouns in genitive plural, for instance *kopalni* (neutral form) / *kopalń* ‘coal mines’. As this differentiation does not affect syntactic relations in the sentence, we do not believe it is justified to include this category into the set of morphosyntactic characteristics.

## 2.7 The so-called nominal pronouns

Neither the contemporary grammar of Polish [6] nor Saloni (in: [19]) distinguish pronouns as a separate part of speech. Lexemes which were traditionally regarded as pronouns have been allocated to various classes based on semantic and syntactic criteria, respectively: nouns (e.g. *ja* ‘I’), adjectives (e.g. *ten* ‘this’), numerals (e.g. *wiele* ‘many’) and adverbs (e.g. *tam* ‘there’).

Following the assumptions implemented in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [5] we waive/renounce the elimination of pronouns as a class (albeit we consider such an elimination to be a right step). In order to retain consistency in language descriptions within MULTEXT-East Morphosyntactic Specifications, we recognise the existence of the class of pronouns. For this reason, we do not include the subgroup of nominal pronouns in the description of Polish and Lithuanian nouns.

## 3 Summary

In this paper the authors focus on providing morphosyntactic specifications for Polish and Lithuanian nouns based on the assumptions embedded in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [5]. Due to the innovativeness of Polish nouns a need arose to identify new categories: human, animate and depreciativeness, and to simplify gender in plural. The conservatism of Lithuanian nouns means that the description of the subcategory has become transparent and largely exception-free. The dual number has disappeared in both languages. The only Lithuanian innovation (versus Polish) i.e. the disappearance of neuter gender, is also worth noting.

## Bibliography

- [1] Ambrazas, V., editor (1997). *Lithuanian Grammar*. Baltos lankos, Vilnius.
- [2] Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, Concepts and Relations in the Construction of Polish WordNet. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Global WordNet Conference, Szeged, Hungary January 22–25 2008*, pages 162–177. University of Szeged.

- [3] Dimitrova, L., Koseska, V., Dutsova, R., and Panova, R. (2009). Bulgarian-Polish Online Dictionary — Design and Development. In Koseska-Toszewa, V., Dimitrova, L., and Roszko, R., editors, *Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009*, pages 162–177, Warszawa. Institute of Slavic Studies, Polish Academy of Sciences, Slawistyczny Ośrodek Wydawniczy.
- [4] Dimitrova, L., Koseska, V., Roszko, R., and Roszko, D. (2009). Bulgarian-Polish-Lithuanian Corpus — current development. To appear.
- [5] Erjavec, T., editor (2004). *MULTEXT-East Morphosyntactic Specifications. Version 3.0. May 10th, 2004*. PDF.
- [6] Grzegorzczak, R., Laskowski, R., and Wróbel, H., editors (1984/1998). *Gramatyka współczesnego języka polskiego*. Wiedza Powszechna, Warszawa.
- [7] Koseska, V., Roszko, R. (2008). Remarks on classification of parts of speech and classifiers in an electronic dictionary. In *Lexicographic tools and techniques, Mondilex first open workshop, Moscow, Russia, 3–4 October, 2008, Proceedings*, pages 80–88, Moscow. Russian Academy of Sciences. Institute for Information Transmission Problems (Kharkevich Institute).
- [8] Koskeniemi, K. (1983). *Two-level morphology: a general computational model for word-form recognition and production*. Publication No. 11. University of Helsinki, Department of General Linguistics, Helsinki.
- [9] Mańczak, W. (1956). Ile rodzajów jest w polskim? *Język Polski*, 1956(z.2):116–121.
- [10] Moszyński, L. (1984). *Wstęp do filologii słowiańskiej*. Państwowe Wydawnictwo Naukowe, Warszawa.
- [11] Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 2007(11):151–167.
- [12] Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.
- [13] Przepiórkowski, A. (2004a). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- [14] Przepiórkowski, A. (2004b). *Korpus IPI PAN. Wersja wstępna*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- [15] Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pages 33–40, Budapest.
- [16] Przepiórkowski, A. and Woliński, M. (2003b). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116, Budapest.
- [17] Roszko, R. (2005). Ogólny opis kontrastywny języka litewskiego i polskiego. *Acta Baltico-Slavica*, 29:47–68.
- [18] Roszko, R. (2009). Morphosyntactic specifications for Polish. Theoretical foundations. description of morphosyntactic markers for Polish nouns within multext-east morphosyntactic specifications (version 3.0 may 10th, 2004). In *Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovakia, 15–16 April, 2009. Proceedings*, pages 140–150, Bratislava. L' Štúr Institute of Linguistics. Slovak Academy of Sciences.
- [19] Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R. (2007). *Słownik gramatyczny języka polskiego. Podstawy teoretyczne. Instrukcja użytkownika*. Wiedza Powszechna, Warszawa.
- [20] Tokarski, J. (2002). *Schematyczny indeks a tergo polskich form wyrazowych*. Saloni, Z. (eds.), wyd. 2, Wiedza Powszechna, Warszawa.
- [21] Valeckienė, A. (1998). *Funkcinė lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidybos institutas, Vilnius.
- [22] Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII–XXIII:39–55.
- [23] Zinkevičius, V. Lemuoklis — morfologinei analizei. *Darbai ir dienos*, 24:245–274.
- [24] Korpus IPI PAN: <http://www.korpus.pl/> (corpus of Polish)
- [25] Miłkowski M.: <http://morfologik.blogspot.com/>

- [26] Miłkowski M. [PDF]: <http://nlp.ipipan.waw.pl/NLP-SEMINAR/061016.pdf>
- [27] *Morfologinis anotatorius*: [http://donelaitis.vdu.lt/main.php?id=4&nr=7\\_1](http://donelaitis.vdu.lt/main.php?id=4&nr=7_1) (tagger for Lithuanian)
- [28] National Corpus of Polish: <http://nkjp.pl/>
- [29] Tekstynas: <http://donelaitis.vdu.lt/> (corpus of Lithuanian)

# Description of Morphosyntactic Markers for Polish Verbs within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)\*

Roman Roszko

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw  
roman.roszko@ispan.waw.pl

**Abstract.** This paper refers to [10] and, indirectly, to Danuta Roszko & Roman Roszko (in this volume), which present the conditions for an academically sound subdivision of lexemes into classes (parts of speech, POS) and morphosyntactic specifications for, inter alia, Polish nouns within the standards developed for MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [3]. At present the Author shows morphosyntactic specifications for Polish verbs. It is commonly known that Polish verbs, alongside Polish nouns, represent the most difficult stage of efforts to develop morphosyntactic specifications for the Polish language. Due to a large number of innovations and dynamics of changes the subcategories for Polish verbs are particularly difficult to identify. Those difficulties become even greater if the morphosyntactic specifications are to be consistent with the MULTEXT-East Morphosyntactic Specifications.

**Keywords:** POS : parts of speech, verb, annotation of parallel corpus, Polish.

## 1 Introduction

The problem involving the degree of morphologisation of various meanings in natural language has a significant bearing on the grammatical description of particular languages. A high number of morphological categories, their transparency and absence of exceptions greatly facilitate such a description. However, Polish is not one of the languages where the degree of formalisation of meanings would facilitate grammatical description. It is for a reason that the theoretical foundations for the grammatical dictionary of the Polish language, authored by Z. Saloni, are the object of our interest (Saloni in: [11]).

**Morphosyntactic descriptions for Polish** The system of morphosyntactic markers developed for the Polish language at the Institute of Computer Science, Polish Academy of Sciences (Pol. IPI PAN) (A. Przepiórkowski, M. Woliński: [8], [9], [12], [5], [6]/[7]), is based on a sound methodological foundation comprising linguistic work by authors such as Z. Saloni, M. Świdziński, J. S. Bień. It is thanks to this foundation that the IPI PAN's tagset goes beyond the fossilised traditional framework dating back to Aristotle. On the other hand, the MULTEXT-East tagset, which serves as a point of reference here, is based on the traditional subdivision into parts of speech (this is why, among others, pronouns have been singled out as a part of speech). MULTEXT-East also replicates the deeply-rooted traditional stereotypes regarding the subdivision into parts of speech as well as linguistic categories or morphological subcategories. Moreover, spaced versus unspaced spelling may decide whether or not a category is identified (e.g. articles, interpreted differently for various languages).

The IPI PAN tagset is used not only in the IPI PAN Corpus ([13]; [7]/[6]) but also, in a somewhat modified version, in the National Corpus of Polish [16] and a few other projects, some of which are conducted outside IPI PAN, e.g. in the Polish WordNet project (M. Piasecki: [1], [4]) developed in Wrocław or in Morfologik (M. Miłkowski [14, 15]) and other. These facts clearly indicate that the IPI PAN tagset has become a benchmark for the Polish language.

---

\* The study and preparation of these results have been supported by the EC's Seventh Framework Programme [FP7/2007–2013] under the grant agreement 211938 MONDILEX.



Consequently, the aim of this series of papers is neither to revise the good and highly refined IPI PAN tagset nor to replace it with a new tagset for Polish. The issue in question is what kind of compromise should be sought when developing a joint tagset to be used for simultaneous description of three languages in the BG-PL-LT parallel corpus. For some reasons the MULTEXT-East tagset (developed previously for many languages) has been selected as the leading one for this corpus [2]. Therefore, the aim of this series of papers is to provide a theoretical study of various categories of Polish and Lithuanian, to set priorities (e.g. morphological, semantic, syntactic) in identifying various meanings and to provide a classification of morphosyntactic phenomena which does not contradict the MULTEXT-East standard and does not deviate too strongly from the IPI PAN tagset. In a sense, we seek to establish correspondence/consistency between the two tagsets. If such correspondence proved impracticable, we would confine ourselves to specifying differences (not only significant ones) between the MULTEXT-East tagset and the IPI PAN tagset.

It cannot be excluded that due to the obvious difficulties in achieving consistency of the intertagset the BG-PL-LT corpus will use the IPI PAN tagset for Polish and its modification for Lithuanian. This solution would certainly necessitate a list of more or less close equivalents for the two tagsets: a tagset for Bulgarian on the one hand, and the IPI PAN tagset on the other (for Polish and an extended version for Lithuanian). For some reasons a review of the tagset for Bulgarian is not planned at this stage of work.

It is important to emphasise that only a coherent tagset for a parallel multilingual corpus 1) allows complete linguistic confrontation, 2) enables identification of linguistic facts, 3) enables a search based on pre-defined unambiguous morphosyntactic characteristics.

## 2 Verb — a definition

Verbs are lexemes which are inflected by persons, a feature which is crucial for conjugation. The category of case, inflectional by definition, does not apply to verbal lexemes. Saloni (Saloni : 83) identifies two key subdivisions within verbs:

- (a) proper verbs – non-proper verbs,
- (b) finite verbs – non-finite verbs.

Below given are the respective definitions for the aforementioned groups of verbs.

**Proper verbs** The basis for identification is formal: morphological combined with syntactic. This group of verbal lexemes is characterised by inflection for mood, tense, person, number and gender.

**Non-proper verbs** The basis for identification is formal: morphological combined with syntactic. This group of verbal lexemes is characterised by inflection for mood and tense whereas the inflective categories of number or gender do not apply to it.

**Finite verbs** Identification is based on the syntactic function: this category includes units which lie at the core of the sentence or act as predicates. Other elements of the sentence remain subordinate to the finite verb.

**Non-finite verbs** Identification is based on the syntactic function: this category includes units which are identified as complements (negation) to finite verbs. Examples of non-finite forms: *czytać* (infinitive), *czytając* (present participle), *przeczytawszy* (perfective participle).

## 3 Key inflective categories of verbs

### 3.1 Mood

The following moods should be distinguished in the Polish language:

**Indicative mood** (*czytam*),

**Conditional mood** (*czytałbym*),

**Imperative mood** (2nd person: *czytaj*, *czytajcie*, 1st person: *czytajmy*).

It is important to note the complex structure of verbal forms in the conditional mood. This is an analytical structure which consists of a quasi-participium, conditional mood operator and agglutinative suffix of praesenti personal form for the verb *być* 'to be': *m/em*, *ś/eś*, *śmy/eśmy*, *ście/eście*. In the aforementioned example of *czytałbym* those would be: *czytał* (quasi-participium) -*by* (operator) -*m* (agglutinative suffix).

### 3.2 Tense

Considering the formal features of verbal lexemes, one should identify the following forms:

Past perfect tense (*czytałem był*, *przeczytałem był*),

Past tense (*czytałem*, *przeczytałem*),

Non-past tense (*czytam*, *przeczytam*),

Future tense (*będę czytać* / *będę czytał*).

The past perfect (plusquamperfectum) tense is, in fact, archaic. This is reflected not only in the very low frequency of its use but, first and foremost, in errors in production of those forms for the 1<sup>st</sup> and 2<sup>nd</sup> person, as in a recent bank advertisement which identifies the erroneous form of \**oszczędzałem byłem* as past perfect.

The past tense has an identical structure for the majority of 'typical' Polish verbs. By convention, it is considered as a single orthographic word. However, in this morphosyntactic description this tense is interpreted as being analytical, consisting of quasiparticipium<sup>1</sup> praeteriti activi + agglutinative suffix of the personal form of praesenti for the verb *być* 'to be': *m/em*, *ś/eś*, *śmy/eśmy*, *ście/eście*, for instance: *czytał-em*, *przeczytał-em*.

The non-past tense is a working notion developed to denote simple verbal forms which, depending on their aspect, are either predestined to express the state which is concurrent with the state of the utterance (*czytam*) or one which follows the state of the utterance (*przeczytam*).

Future tense — this refers to analytical forms produced from state-describing verbs (aspect: imperfective) which are intended to express a state that follows the state of the utterance (*będę czytać* / *będę czytał*).

Future tense produces compound expressions consisting either of a future form of the auxiliary verb *być* 'to be' + quasi-participium (*będę czytał*) or of a future form of the auxiliary verb *być* 'to be' + infinitivus (*będę czytać*).

### 3.3 Aspect

With the goal of this description in mind, we assume that aspect is an inflective category. We distinguish between verbal lexemes with an imperfective aspect and those with a perfective aspect. We see the difference between the two aspects (perfective and imperfective) not in the meaning as such but in the set of forms that are typical of each aspect. Therefore, the perfective aspect is characterised with a perfective participle (e.g. *przeczytawszy*) whereas the imperfective aspect is characterised with the forms of future tense (e.g. *będę czytać* / *będę czytał*), active adjectival participle (e.g. *czytający*) and present participle (e.g. *czytając*). The aforementioned forms are unique for a particular value of aspect. Some authors (e.g. [11, p. 86]) identify <present tense> as a feature of the imperfective aspect (e.g. *czytam*), and <synthetic/simple future tense> as a feature of the perfective aspect (e.g. *przeczytam*). In this description, however, we do not use the notions of <present tense> or <synthetic/simple future tense> as they reflect the contamination of two morphological characteristics: non-past tense and aspect (perfective or imperfective).

<sup>1</sup> Identification of quasi-participium forms is given in order to ensure coherence of this description. Therefore, the agglutinative suffix manifested in past tense forms is the decisive factor for identifying quasi-participium.

Table 1. List of inflective forms characteristic of imperfective and perfective aspect

Form	Examples	
	perfective aspect	imperfective aspect
future tense		<i>będę czytać / będę czytał</i>
perfective participle	<i>przeczytawszy</i>	
active adjectival participle		<i>czytający</i>
present participle		<i>czytając</i>

### 3.4 Gender

Gender does not apply to all verbal forms. This category is associated with forms based on quasi-participle and applies to singular only. It assumes the following values:

Masculine (e.g. *przeczytałem, przeczytałeś, przeczytał*),

Feminine (e.g. *przeczytałam, przeczytałaś, przeczytała*),

Neuter (e.g. *przeczytało*).

Polish is said to have the so-called masculine personal gender and non-masculine gender in plural forms (e.g. [11, p.90]). However, we do not use these terms herein and we associate the differentiated forms of plural with the human attribute, taken in separation from gender, as in our description of nouns [10].

### 3.5 Person

The category of person does not apply to all verbal forms. It is associated with the so-called personal verbal forms and takes three values:

1<sup>st</sup> person (e.g. *przeczytałem, przeczytaliśmy, czytam, czytamy*),

2<sup>nd</sup> person (e.g. *przeczytałeś, przeczytaliście, czytasz, czytacie*),

3<sup>rd</sup> person (e.g. *przeczytał, przeczytali, czyta, czytają*).

### 3.6 Number

The inflective category of number assumes the following values:

Singular (*czytałem, czytałeś, czytał, czytam, czytasz, czyta, będę czytać, będziesz czytać, będzie czytać, czytaj, czytałbym*),

Plural (*czytaliśmy, czytaliście, czytały, czytali, czytamy, czytacie, czytają, będziemy czytać, będziecie czytać, będą czytać, czytajcie, czytaliście*).

### 3.7 Voice

Active and passive voice forms are identified in Polish, for example: *Jan czyta książkę*. ‘Jan is reading a book’ (active form) and *Książka jest czytana przez Jana*. ‘A book is being read by Jan’ (passive form) where:

*Jan* (nom. sg.) *czyta* (nonpraeteritum 3 sg.) *książkę* (acc. sg.).

*Książka* (nom. sg.) *jest* (copula) *czytana* (part. praeteriti passivi.) *przez* (preposition) *Jana* (acc. sg.).

As can be seen from this example, the difference between the two sentences is not caused by a difference in inflection of the verb *czytać* ‘to read’. Polish grammar also identifies the so-called reflexive voice. Example: *Jan się myje*. ‘Jan is washing himself’:

*Jan* (nom. sg.) *się* (acc. sg.) *myje* (nonpraeteritum sg.)

As in the case of the active versus passive voice distinction, it is not the verbal inflection that determines reflexivity (i.e. stating something where the subject and the object are the same).

Therefore, voice cannot be considered as a flexive characteristic of verbs. Rather, it is a syntactic characteristic.

## 4 Non-finite forms

### 4.1 Infinitive

Polish verbs have only one form of infinitive, e.g. *czytać*. The infinitive rarely acts as a predicate, e.g. *Po ostudzeniu dodać masła*. ‘Add butter after cooling.’ In most cases, infinitive is syntactically required by other lexemes, e.g. *trzeba czytać* ‘need to read’, *musiał powiedzieć* ‘had to say.’

### 4.2 Present participle

Polish has only one form of present participle. It is formed in a regular way from imperfective verbs, e.g. *czytając*. This participle is intended to express a concurrent state, e.g. *Jan, czytając książkę, nucił melodię*. ‘While reading a book, Jan was humming a tune’ where the state of reading a book and the state of humming a tune are concurrent for the same subject (Jan in this example).

### 4.3 Past participle

Polish has only one form of past participle. It is formed in a regular way from perfective verbs, e.g. *przeczytawszy*. This participle is intended to express a preceding event, e.g. *Jan przeczytawszy książkę nucił melodię*. ‘Having read a book, Jan was humming a tune’ where the event «Jan read a book» precedes the state of Jan’s humming a tune. Both the event and the state relate to the same subject (Jan in this example).

### 4.4 Gerund

Gerunds are treated herein as nouns. Their inflection is typical to that of nouns. We do not consider it valid to include gerunds into the verbal paradigm. This decision is justified by the existence of declension which is typical of nominal forms. However, in order to adhere to the rules adopted within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [3], we are willing to accept an approach where gerunds are viewed as verbal forms. Gerunds have a high frequency of use in Polish. In many cases Polish gerunds take the position which is typically held by infinitives in other languages, e.g. Pol. *Czas zacząć przygotowania* (gerund). ‘Time to start preparations’ – Lit. *Laikas pradėti ruoštis* (infinitive). ‘Time to start preparations’. Moreover, in many cases it requires the same syntax of subordinate elements as verbal forms do, e.g. *Jan gotuje mięso, zupę*. ‘Jan is cooking meat, soup’ / *Trzeba gotować mięso, zupę*. ‘One needs to cook meat, soup’ and *Gotowanie* (gerund) *przez Jana mięsa, zupy*. ‘The cooking of meat, soup by Jan.’

### 4.5 Adjectival participles

Adjectival participles are formed from verbs and have the typical declension of adjectives (inflection for gender, number and case). We do not consider it necessary to include adjectival participles into the verbal paradigm. Also in this case the reason is that such participles have inflection which is typical of nominal forms. However, in order to adhere to the rules adopted within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [3], we are willing to accept an approach where adjectival participles are viewed as verbal forms. Nevertheless, it is important to bear in mind that Polish participles cannot act as stand-alone predicates, e.g. *On czyta*. (nonpraeteritum 3. sg.) ‘He is reading’ and *\*On czytający*. (part.) ‘\*He reading.’ This is possible in other languages, such as Lithuanian: *Jis skaito*. ‘He is reading’, *Jis skaitęs*. (part. praet. act.). ‘He was apparently reading (then),’ *Jis skaitąs*. (part. praes. act.). ‘He is apparently reading (now)’ etc.

**Active adjectival participle** Active adjectival participles are formed from imperfective verbal forms, e.g. *czytający* ‘reading’.

**Past adjectival participle** Past adjectival participles are fossilised forms. Their forms may be generated only from a few (selected) perfective non-transitive verbs, e.g. *poszarzały* ‘grayed’.

**Passive adjectival participle** Passive adjectival participles are formed mostly from transitive verbs, e.g. *czytany* ‘read’, *bity* ‘beaten’.

## 5 Transitive and quasi-transitive verbs

Polish has a class of verbs which govern nouns in either accusative case, e.g. *czytać* ‘to read’: *Jan czyta książkę* (acc.). ‘Jan is reading a book’ or in genitive case (so-called genitive of negation / genetivus negatio), e.g. *Jan nie czyta książki* (gen.). ‘Jan is not reading a book’. Those are the so-called transitive verbs.

When the distinction between an adjective and an adjectival passive participle becomes problematic, e.g. *uśmiechnięty* ‘smiling’, *poszkodowany* ‘wronged’, literature talks about quasi-transitivity of verbs. It is important to bear in mind that the verb *uśmiechnąć się* ‘to smile’ is not transitive whereas the verb *poszkodować* ‘to wrong’ is not used in contemporary Polish.

## 6 Non-transitive verbs

Non-transitive verbs are all the remaining verbs which do not have the attribute of transitivity, e.g. *biec* ‘to run’.

## 7 Untypical verbs

Polish has a small group of untypical verbs. They are considered untypical because of an incomplete inflective pattern and/or lack of nominal derivatives, e.g. *powinien* ‘should’, *winien* ‘ought to’, *rad* ‘glad (to do sth)’, *gotów* ‘ready (to do sth)’.

Another group of untypical verbs consists of the so-called non-proper verbs which do not collocate with a subject in the nominative case. As a result, they have an incomplete set of forms (no distinction into person, number and gender). Those verbs include *braknąć* ‘running short of sth’ (*Zawsze pod wieczór braknie chleba.* ‘Towards the evening one always runs short of bread’), *zabraknąć* ‘become unavailable’ (*Latem zabrakło wody.* ‘Water was unavailable in summer’), *należeć* ‘should’ (*Należało się spotkać, a nie opowiadać kłamstwa.* ‘One should have met instead of telling lies’).

Another group includes the so-called secondary non-proper verbs (= predicatives): *trzeba* ‘should’, *można* ‘may’, *widać* ‘apparently’.

## 8 Modal verbs

Polish has a very limited number of modal verbs, e.g. *musieć* ‘must’. In many cases modal verbs are also non-proper verbs, e.g. *trzeba* ‘should’, *można* ‘may’, *należeć* ‘ought to’, *należeć się* ‘be due’ (e.g. *Za sprzątnięcie należy się 10 zł.* ‘There is a charge of PLN 10 due for the cleaning.’)

## 9 Summary

Given the need to adjust the description of Polish verbs to the requirements posed under MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) [3], it was necessary to abandon the idea of viewing adjectival participles and gerunds as nominal forms. Some complication in the morphosyntactic description of Polish is caused by the formal (inflection-based) distinction between simple forms such as *czytam*, *przeczytam* (nonpraeteritum) which are intended to express the meaning or the present tense or the future tense. The use of nonpraeteritum in a specific meaning is determined by another inflective characteristic, i.e. aspect of verbs. For this reason, the inflective category of tense in Polish cannot be seen as transparent, clear or unambiguous.

## Bibliography

- [1] Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, Concepts and Relations in the Construction of Polish WordNet. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Global WordNet Conference, Szeged, Hungary January 22–25 2008*, pages 162–177. University of Szeged.
- [2] Dimitrova, L., Koseska, V., Roszko, R., and Roszko, D. (2009). Bulgarian-Polish-Lithuanian Corpus — current development. To appear.
- [3] Erjavec, T., editor (2004). *MULTEXT-East Morphosyntactic Specifications. Version 3.0. May 10th, 2004*. PDF.
- [4] Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 2007(11):151–167.
- [5] Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.
- [6] Przepiórkowski, A. (2004a). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- [7] Przepiórkowski, A. (2004b). *Korpus IPI PAN. Wersja wstępna*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- [8] Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pages 33–40, Budapest.
- [9] Przepiórkowski, A. and Woliński, M. (2003b). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116, Budapest.
- [10] Roszko, R. (2009). Morphosyntactic specifications for Polish. Theoretical foundations. description of morphosyntactic markers for Polish nouns within multext-east morphosyntactic specifications (version 3.0 may 10th, 2004). In *Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovakia, 15–16 April, 2009. Proceedings*, pages 140–150, Bratislava. L’Štúr Institute of Linguistics. Slovak Academy of Sciences.
- [11] Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R. (2007). *Słownik gramatyczny języka polskiego. Podstawy teoretyczne. Instrukcja użytkownika*. Wiedza Powszechna, Warszawa.
- [12] Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII–XXIII:39–55.
- 
- [13] Korpus IPI PAN: <http://www.korpus.pl/>
- [14] Miłkowski M.: <http://morfologik.blogspot.com/>
- [15] Miłkowski M. [PDF]: <http://nlp.ipipan.waw.pl/NLP-SEMINAR/061016.pdf>
- [16] National Corpus of Polish: <http://nkjp.pl/>



Part 3  
**Some Semantic Problems of Contrastive Studies  
of Slavic Languages and Related Topics**





# Lexical Functions in Bulgarian and Russian: a Sketch to Digital Comparative Lexicography

Svetlana Timoshenko, Olga Shemanaeva

Institute for Information Transmission Problems RAS, Moscow, Russia

**Abstract.** This paper presents an approach to the creation of Russian-Bulgarian digital dictionary of collocations using the apparatus of lexical functions. The project is aimed not only at the high-quality translation and disambiguation but also at the cross-linguistic analysis and at comparing the semantics and compatibility of the words in Slavic languages (Russian and Bulgarian) by means of digital lexicographical data. Another important application is computer-assisted language learning: Bulgarian data can be incorporated in the educational project being developed for Russian and English at the Institute for Information Transmission Problems of the Russian Academy of Sciences.

**Keywords** lexicon, collocations, lexical functions, digital lexicography, computer-assisted language learning, cross-linguistic study

## 1 Introduction

Lexical functions (LF) were created and developed as a convenient tool to describe lexical co-occurrence and semantic derivation relations between two lexical units. They are applied to disambiguation and idiomatic translation.

The notion of lexical functions is well-known and used in many lexicographic projects for such languages as French, English and Spanish, in addition to Russian. (cf. [12], [13]. [6]).

The theoretical foundations of the software system for Russian language lie within the Meaning  $\leftrightarrow$  Text Theory (MTT) by I. Mel'čuk ([10], [11]) and the Theory of Systemic Lexicography by Ju. Apresjan ([4], [1]).

The goal of this paper is to show how the apparatus of lexical functions may be used for lexicographic purposes cross-linguistically, namely to compare two closely related languages (Russian and Bulgarian) using digital lexicographical data.

It is especially useful because as far as we know there is no such combinatorial dictionary for Bulgarian that has all the data about the syntax and semantics of the word consequently shown.

## 2 Lexical functions: definitions and types

Although lexical functions as a tool of lexicographic description have a long history, we will remind in brief what the lexical functions are and what types of lexical functions generally occur in language.

In principle, there is a strict mathematical definition of lexical functions suggested in [5, p. 207]. For our purposes, it would be sufficient to describe lexical functions as a system of basic senses that are common for all the world languages. These senses can be expressed in many different ways, but since they are basic senses they are often expressed in a regular way. A basic sense can be expressed either morphologically, within the given word (Rus. *дом* 'house'— *домшине* 'big house'), or syntactically, by means of a neighboring word (Engl. *house* – *big house*). Theoretically, Lexical Functions can be applied to describe semantic relations between language units of all kinds. Many senses of lexical functions have parallels in morphology, for example grammatical feature «plural» can be seen as regular, grammaticalized way to express the sense of LF Mult. For more examples see [8, p. 246] We can state the special relations between the senses and classify them.

Lexical functions may be split into three types: substitutes, derivative substitutes and collocates. Such LFs as Anti, Syn, Conv, Gener, some cases of Mult and Sing are substitutes, a value

of such LF is another word that is paradigmatically bound with the given key word. The key word and the value of such lexical function do not necessarily co-occur in text, in fact, such co-occurrence is rather rare.

Bulg. Conv (*мъж*) = *жена* (Conv ('husband') = 'wife')

Rus. Conv (*муж*) = *жена*

Bulg. Anti(*зрящ*) = *сляп* (Anti ('sighted') = 'blind')

Rus. Anti (*зрячий*) = *слепой*

Bulg. Anti(*интересен*) = *неинтересен* (Anti ('interesting') = 'dull / boring')

Rus. Anti (*интересный*) = *неинтересный*

Typical examples of derivative lexical functions are morphologically derived words like Russian *любить* 'to love' (V) – *любовь* 'love' (S), *помогать* 'to help' (V) – *помощник* 'helper, assistant' (S).

Rus. S0 (*любить*) = *любовь*

V0 (*любовь*) = *любить*

Rus. S1 (*помогать*) = *помощник*

V0 (*помощник*) = *помогать*

These examples show one special feature of this type: derivative lexical functions are symmetrical. If one word is the keyword of derivative lexical function, so it is true that we can find another derivative lexical function where the value becomes the keyword and the keyword takes place of the value.

As lexical functions are semantically motivated, they may also describe the so-called semantic derivation [9, p. 462]. These cases, though evident to the native speaker, turn out to be more complicated than it may seem: thus the default name of the agent for the Russian verb *лечить* 'to heal, cure' is not its derivate *лекарь*, but another noun *врач*, morphologically unrelated with *лечить*.

Such is also the case of the Bulgarian noun *любов* 'love' and the verb *обичам* 'to love':

V0 (*любов*) = *обичам*

or the Bulgarian verb *подкова* 'to shoe (a horse etc.)' and *подковач* / *налбантин* 'smith'

S1 (*подкова*) = *подковач*, *налбантин*

Among LF's of this type are: V0, S0, S1, S2, A0, A1, A2, Adv0, Adv1, Adv2 and others.

LF-collocates are probably the most interesting to lexicographers. They establish relations between words that not only exist in the dictionary but also in the text, when the keyword and the value of lexical function may form a stable collocation with the specified meaning and be related by a syntactic relation. Among those functions are nominal LFs Mult and Sing, adjectival LFs Magn, AntiMagn, Bon, AntiBon, Ver, AntiVer, Adv1, Adv2 and verbal LFs Oper 1, Oper 2, Func 1, Labor 1–2, Fin, Incep and so on.

Bulg. Oper1 (*опашка*) = *стоя на (опашка)* (Oper1 ('queue') = 'stand in a queue')

Rus. Oper1 (*очередь*) = *стоять <в> (очереди)*

Bulg. Magn (*опашка*) = *дълга / грамадна (опашка)* (Magn ('queue') = 'long queue')

Rus. Magn (*очередь*) = *длинная / огромная / дикая (очередь)*

There is also a possibility of adding up the meanings of LFs and constructing compound LFs.

Primary LFs are Anti (the opposite meaning) and Magn (high degree), but the combination of the simple meanings can provide AntiMagn (low degree).

Anti (*хубава заплата*) = Anti (Magn (*заплата*)) = Antimagн (*заплата*) = *мизерна (заплата)*

AntiMagn (*зарплата*) = Anti (Magn (*зарплата*)) = Antimagн (*зарплата*) = *мизерная / скромная (зарплата)*

Fin (*вали дъжд*) = Fin (Func0 (*дъжд*)) = Finfunc0 (*дъжд*) = (*дъждът*) *спря*

Fin (*идет дождь*) = Fin (Func0 (*дождь*)) = Finfunc0 (*дождь*) = (*дождь*) *прекратилсѧ / перестал*

FinOper1 (*играта*) = *спирам (играта)*

FinOper1 (*игра*) = *закончить (игру)*

LiquFact0 (*радиото*) = *спирам, изключвам (радиото)*

LiquFact0 (*радио*) = *выключить* (*радио*)

LiquFact0 (*электрический ток*) = *спирям* (*электрический ток*)

LiquFact0 (*ереэлектричество*) = *отключить* (*ереэлектричество*)

CausFunc1 (*впечатление*) = *правя* (*впечатление*) *на*

CausFunc1 (*впечатление*) = *производить* (*впечатление*) *на*

CausOper2 (*опека*) = *поставям под* (*опека*)

CausOper2 (*опека*) = *отдавать под* (*опеку*)

FinOper1 (*блокада*) = *вдигам / махам* *блокадата*

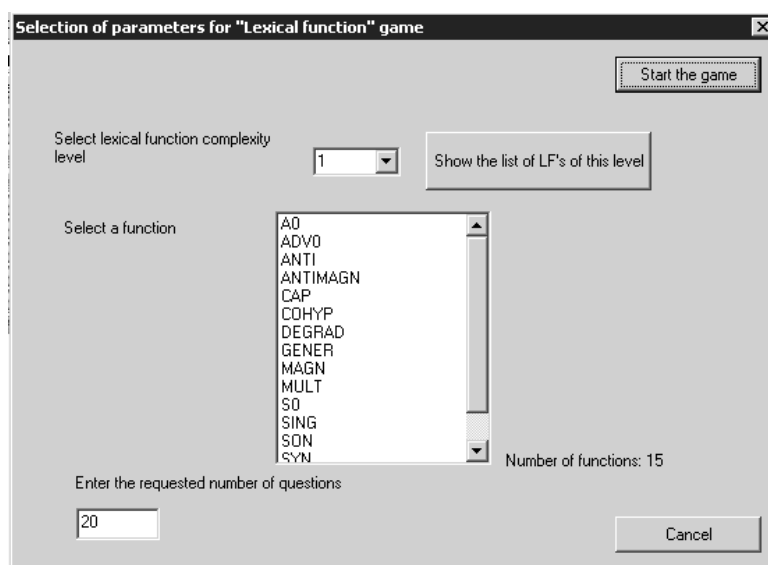
FinOper1 (*блокада*) = *снимать* *блокаду*

Normally, there is a possibility of merging up to three simple meanings in one general meaning of the given situation: for example, *вали дъжд* ‘to rain’ is modified with the meanings Incep ( $\approx$  to start), Anti ( $\approx$  the opposite) and Magn ( $\approx$  very) and the result is *позавалявам* ‘to drizzle’.

**IncepAntiMagn** (*вали дъжд*) = *позавалявам*  
 ‘начинать слегка накрапывать (о дожде)’

### 3 Applications

Lexical functions are used in the lexicographic digital learning device created by Yu. D. Apresjan and L. L. Tsinman which helps the learners study the variety of language and apply their knowledge to modify the meaning of the given situation ([5], [7], [2]).



The functions are ranged from the most intuitively clear and the simplest to the most complicated and elaborated.

Game "Lexical function" - Olga

Function name **MAGN** Finish

Definition

An adjective, an adverb, a preposition + noun collocation or a comparative set expression with a conjunction denoting a large degree or a high intensity of X and fulfilling the function of an attribute or an adverbial modifier of X

Grade

Current  
0  
Normal  
1

*Example 1*

FEELING

deep

profound

*Example 2*

RAIN

heavy

Argument

EXHAUSTED

Your variant of the answer

Show correct answer Answer

A definition and a pair of examples are given to each task and the students then try to apply their language capacity to what is being asked of them.

In fact there are many games in one program. Learners can take definitions and guess words like in the crossword, can take a set of words and fill all their possible lexical functions or choose one or more lexical functions and guess their values for the given words. This game is a good example for creating and usage of multiple access database including lexical-functional information. The database of the same type can be used for a digital dictionary. When the user plays a game, it is the computer who asks and the user who answers. When the user searches information in the dictionary, he asks and the dictionary database must "answer". This similarity means that we can use not only the same database but near the same interface.

We intend that user will input data (collocation in source language), parser will identify the lexical function and will give the translation with respect to lexical function. Translation with the help of lexical functions is more idiomatic and correct.

#### 4 Theoretical extension of the work. Conclusion

The multilingual lexicographic resource based on lexical functions and their cross-lingual correlation may be of high interest both to language learners and researchers. The advantages for the first category of potential users are self-evident.

What advantages do dictionaries with LF marking offer to researchers? In the first place it is useful for such domains of linguistics as theoretical semantics, lexical typology and cognitive linguistics (studies of language concepts and metaphors). For example, we already know that there is a correlation between semantic class and subclass of word and the verb it prefers as a value of lexical function. Russian verbal derivatives with the meaning 'action' have typical Oper2 – *подвергаться* ('undergo'), while derivatives with the meaning 'state' or 'result' choose the verb *находиться* ('be situated') or *быть под* ('be under') as a value of Oper2. If the noun has both

meanings, it can match with both verbs as LF values, but if it has only one meaning, there is no choice, cf. [1, p.105]. Let us compare some semantically close words and see how they collocate with verbs values of LF Oper2.

Oper2	<i>критика</i> (‘criticism’)	<i>ревизия</i> (‘inspection ’)	<i>контроль</i> (‘control’)	<i>власть</i> (‘power’)
‘state’ ( <i>находиться</i> )	–	–	+	+
‘action’ ( <i>подвергаться</i> )	+	+	+	–

We can see that only word *контроль* can match with both verbs. With *критика* and *ревизия* it is possible to use only the action verb, because they signify an action, while *власть* behaves in a different way – as a typical state. It is absolutely impossible to say \**подвергаться власти* or \**быть под критикой*. So we can try to predict lexical compatibility based on semantic class. We can also assume that the opposite is true and we can classify words into different semantic classes or subclasses with respect to values of lexical functions they have. In other words if we weren’t native speakers but had learned that some words match with *подвергаться* and some – with *находиться*, we could hypothesize that there are at least two semantic classes in Russian and describe the slight difference between them. We can apply the same procedure to Bulgarian. Bulgarian and Russian are closely related and the basic hypothesis predicts that in Bulgarian we shall find the same semantic types with similar values of lexical functions. The examples confirm this assumption on a mass scale.

Bulgarian:

Oper2	<i>критика</i> (‘criticism’)	<i>ревизия</i> (‘inspection ’)	<i>контроль</i> (‘control’)	<i>власт</i> (‘power’)
‘state’ ( <i>съм под / намирам се под</i> )	–	–	+	+
‘action’ ( <i>се подлага на</i> )	+	+	+	–

The second step we can do is to try other lexical functions. In Russian the verb that can be the value of LF Oper2 with the nouns corresponding to actions can also be the value of LF Oper1 if it matches with the nouns corresponding to processes and states that have only one actant. We can hypothesize that in Bulgarian *се подлага на* will be compatible with another class – with typical processes as *изменение* ‘change’ (from *изменяться*), *старение* (‘the process of getting old’). But empirical data do not corroborate this hypothesis. We took a short list of typical processes and tested their compatibility with the verb *се подлага на*. We used Bulgarian National Corpus and Bulgarian Internet. Only half of words from our list were bounded with the verb, and the part of examples is from Internet and not from Corpora. In Russian all these nouns have *подвергаться* as the normal value of LF Oper1. We may draw a conclusion that subclass of process nouns in Bulgarian behaves in a different way that the same subclass in Russian.

Bulgarian nouns denoting processes: compatibility with the verb *се подлага на*:

<i>изменение</i> ‘change’	Oper1 ( <i>изменение</i> ) = <i>се подлага на</i>
<i>клатене</i> ‘rocking’	–
<i>мутация</i> ‘mutation’	–
<i>опасност</i> ‘danger’	Oper1 ( <i>опасност</i> ) = <i>се подлага на</i>
<i>разрушаване</i> ‘distruction’	Oper1 ( <i>разрушаване</i> ) = <i>се подлага на</i>
<i>риск</i> ‘risk’	Oper1 ( <i>риск</i> ) = <i>се подлага на</i>
<i>старение</i> ‘the process of getting old’	–
<i>съблазън</i> ‘temptation’	–
<i>облъчване</i> ‘irradiation’	Oper1 ( <i>облъчване</i> ) = <i>се подлага на</i>

We consider these data as reliable because other semantic class, typical states, which have Oper1, show the opposite result. All test keywords, which match with the verb *испытывать* ‘to feel’ in Russian, in Bulgarian have Oper1 *изпитвам*: *болка* ‘pain’, *влечение* ‘attraction’, *вожделение* ‘lust’, *гняв* ‘anger’, *гордость* ‘pride’, *жажда* ‘thirst’, *завист* ‘envy’, *лишение* ‘privation’, *нужда* ‘poverty’, *обич* ‘love’, *радост* ‘joy’, *ревность* ‘jealousy’, *срам* ‘shame’, *страх* ‘fear’, *сърбеж* ‘itch’, *трудность* ‘difficulty’, *ужас* ‘horror’. Bulgarian National Corpus gives representative examples of collocations including these words and the verb *изпитвам*, and that is a reason why we consider the lack of collocations including *се подлага на* and process nouns significant.

There is only one exception from the set of state nouns: the word *глад* ‘hunger’. It matches with the verb *изпитвам*, as it must be according to Russian LF rules, but can also use the verb *се подлага на* as the value of LF OPER1. In other words, it behaves not only as the state but also as the process. In is not typical for Russian. This fact is an indirect proof of the hypothesis for the semantic difference of process nouns in Russian and in Bulgarian. *We are not ready now to offer any final judgment about the nature of this difference. The main point in this rough analysis is to show that lexical functions can be a helpful tool to describe and represent the most subtle differences between the words.*

## 5 Acknowledgments

We would like to thank our Bulgarian colleagues Anna Lipowska and Ivan Derzhansky for their advice and recommendations concerning the lexical data. However, only the authors are responsible for all the unclear cases and data analysis. We would also like to thank our colleagues from the Laboratory of Computational Linguistics IITP RAS, and especially Ju. D. Apresjan, L. L. Iomdin, I. .M. Boguslavsky and P. V. Diachenko.

## 6 Appendix. Examples of Russian-Bulgarian equivalents

In what follows we give a number of LFs in Russian and in Bulgarian with their definitions (based on the dictionary of lexical functions for the Interactive lexical textbook by Ju. Apresjan and L. Tsinman and on the data given and classified in [5]). We give Bulgarian examples first.

### Substitutes:

**Anti = [Antonym, a lexeme of the same part of speech as X with the meaning opposite to that of X]**

Anti (*малък*) = *голям*

Anti (*маленький*) = *большой*

Anti (*момиче*) = *момче*

Anti (*девочка*) = *мальчик*

Anti (*първи*) = *последен*

Anti (*первый*) = *последний*

Anti (*север*) = *юг*

Anti (*север*) = *юг*

Anti (*скъп*) = *евтин*

Anti (*дорогой*) = *дешевый*

Anti (*победа*) = *поражение / загуба*

Anti (*победа*) = *поражение*

Anti (*безопасен*) = *опасен*

Anti (*безопасный*) = *опасный*

**Gener = [A noun, an adjective or a verb denoting the genus under which X is subsumed]**

Gener (*червен / син / черен / бял*) = *цвет*  
 Gener (*красный / синий / черный / белый*) = *цвет*

Gener (*круша / ябълка*) = *плод*  
 Gener (*груша / яблоко*) = *фрукт*

Gener (*ряпа*) = *зеленчук*  
 Gener (*репа*) = *овощ*

Gener (*куче / котка*) = *животно*  
 Gener (*собака / кошка*) = *животное*

**Derivative substitutes:**

**S1 = [A noun semantically derived from X and denoting P1 - the person, thing or situation that does X, has X or is in the state X]**

S1 (*уча*) = *учител*  
 S1 (*учить*) = *учитель*  
 S1 (*лекувам*) = *лекар*  
 S1 (*лечить*) = *врач*

**S2 = [A noun semantically derived from X and denoting P2 - the person, thing or situation that undergoes X or is the object of X]**

S2 (*уча*) = *ученик*  
 S2 (*учить*) = *ученик*  
 S2 (*лекувам*) = *пациент*  
 S2 (*лечить*) = *пациент*

**Adv1 = [An adverb or a preposition + noun collocation semantically derived from X and describing the property, state or action of P1]**

Adv1 (*мижа*) = *мижешком / мижешката*  
 Adv1 (*жмуриться*) = *зажмурясь*

**Loc = [an adverb or a collocation of preposition and noun X denoting the standard spatial or temporal location of something towards X]**

Loc (*лято*) = *през лятото / лете* (ADV)  
 Loc (*лето*) = *летом* (ADV)  
 Loc (*поле*) = *в полето / на полето*  
 Loc (*поле*) = *в поле / на поле*  
 Loc (*ден*) = *през деня*  
 Loc (*день*) = *днем*  
 Loc (*пролет*) = *пролетта / (на) пролет* (ADV)  
 Loc (*весна*) = *весной* (ADV)  
 Loc (*вечер*) = *вечерта*  
 Loc (*вечер*) = *вечером*  
 Loc (*сутрин*) = *сутринта*  
 Loc (*утро*) = *утром*  
 Loc (*зима*) = *през зимата / зиме* (ADV)  
 Loc (*зима*) = *зимой* (ADV)  
 Loc (*есен*) = *есента / (на) есен* (ADV)  
 Loc (*осень*) = *осенью* (ADV)  
 Loc (*нощ*) = *през нощта / ноще / нощем* (ADV)  
 Loc (*нощ*) = *ночью* (ADV)  
 Loc (*залез*) = *по залез слънце*  
 Loc (*закат*) = *на заката*  
 Loc (*сянка*) = *на сянка*



Loc (*тень*) = *в тени*

**Caus** = [To cause X to happen or to exist (a verb including the meaning of X in its own meaning and taking P0 as its grammatical subject and P1 as its primary object)]

Caus (*стоя*) = *поставям*

Caus (*стоять*) = *ставитъ*

**Mult** = [A noun denoting an organized or natural set or group of X's and including the meaning of X in its own meaning]

Mult (*ученик*) = *клас*

Mult (*ученик*) = *класс*

Mult (*книга*) = *библиотека*

Mult (*книга*) = *библиотека*

**Sing** = [A noun denoting one element, instance or sample of X and including the meaning of X in its own meaning]

СИНГ (*кѡдрици*) = *кѡдрица*

СИНГ (*кудри*) = *кудряшка*

**Collocates:**

**Mult** = [A noun denoting an organized or natural set or group of X's and subordinating X syntactically]

Mult (*вълк*) = *гълтница (вълци)*

Mult (*вълк*) = *стая (волков)*

Mult (*птица*) = *ято (птици)*

Mult (*птица*) = *стая (птиц) / косяк (птиц)*

Mult (*пираня*) = *пасаж (пирани)*

Mult (*пиранья*) = *стая (пираний)*

Mult (*риба*) = *пасаж (риби)*

Mult (*рыба*) = *косяк (рыбы)*

Mult (*елен*) = *стадо (елени)*

Mult (*олень*) = *стадо (олений)*

Mult (*овца*) = *стадо (овци)*

Mult (*овца*) = *отара (овец)*

Mult (*кон*) = *табун/ хергеле (коне)*

Mult (*лошадь*) = *табун (лошадей) / косяк*

**Sing** = [A noun denoting one element, instance or sample of X and subordinating X syntactically]

Sing (*кѡдрици*) = *кѡдрица*

Sing (*кудри*) = *кудряшка*

**Magn** = [a large degree or a high intensity of X]

Magn (*пропаст*) = *бездѡнна (пропаст)*

Magn (*пропастъ*) = *бездонная (пропастъ)*

Magn (*радост*) = *безкрайна (радост)*

Magn (*радостъ*) = *безграничная (радостъ)*

Magn (*страдание*) = *безкрайно (страдание)*

Magn (*страдание*) = *невыносимое (страдание)*

Magn (*любов*) = *безмерна (любов)*

Magn (*любовъ*) = *безумная/ горячая (любовъ)*

Magn (*страст*) = *силна (страст)*

Magn (*страстъ*) = *пламенная/ слепая (страстъ)*

Magn (*болка*) = *лота/ адска/ непоносима (болка)*

Magn (*боль*) = *сильная* (*боль*)

Magn (*болен*) = *тежко* (*болен*)

Magn (*больной*) = *тяжело* (*больной*)

Magn (*тишина*) = *абсолютна/ гробна* (*тишина*)

Magn (*тишина*) = *полная/ мертвая/ гробовая/ глубокая* (*тишина*)

Magn (*отличник*) = *пълен* (*отличник*)

Magn (*отличник*) = *круглый* (*отличник*)

**Bon = [An adjective or an adverb expressing a standard positive evaluation of X and fulfilling the function of an attribute or an adverbial modifier of X]**

Part of speech depends on keyword's part of speech. Nouns require adjectives to express the standard positive evaluation and verbs – adverbs.

Bon (*край*) = *благополучен* (*край*)

Bon (*конец*) = *благополучный* (*конец*)

Bon (*обстановка*) = *благоприятна* (*обстановка*)

Bon (*обстановка*) = *благоприятная* (*обстановка*)

Bon (*условия*) = *благоприятни* (*условия*)

Bon (*условия*) = *благоприятные* (*условия*)

Bon (*отзив*) = *благоприятен* (*отзив*)

Bon (*отзыв*) = *благоприятный* (*отзыв*)

Bon (*отговор*) = *благоприятен* (*отговор*)

Bon (*ответ*) = *благоприятный* (*ответ*)

Bon (*памет*) = *силна* (*памет*)

Bon (*память*) = *хорошая* (*память*)

Bon (*репутация*) = *неопетнена/ добра/ безупречна* (*репутация*)

Bon (*репутация*) = *незапятнанная/ блестящая/ безупречная* (*репутация*)

Bon (*слух*) = *тънък/ остър* (*слух*)

Bon (*слух*) = *тонкий/ острый* (*слух*)

Bon (*посрещам*) = *сърдечно/ радушно* (*посрещам*)

Bon (*принимать*) (*гостей*) = *тепло/ радушно* (*принимать*) (*гостей*)

Bon (*влияя*) = *положително/ благотворно* (*влияя*)

Bon (*влиять*) = *положительно/ плодотворно* (*влиять*)

Bon (*благодаря*) = *сърдечно* (*благодаря*)

Bon (*благодарить*) = *сердечно* (*благодарить*)

**Son = [To issue a sound characteristic of X (a collocate verb taking X as its grammatical subject)]**

Son (*котка*) = *мъркам, мяукам*

Son (*кошка*) = *мурлыкать*

Son (*водопад*) = *шумя*

Son (*водопад*) = *шуметь*

Son (*оръдие*) = *гърмя*

Son (*орудие*) = *греметь*

Son (*бухал*) = *бухам*

Son (*филин*) = *ухать*

Son (*муха/ бръмбар*) = *бръмча*

Son (*муха/ жук*) = *жуужжатъ*

**Oper 1 = [to do X, to have X, or to be in the state of X (a collocate verb taking P1 as its grammatical Subject and X as its main Object)]**

Oper1 (*полза*) = *донасямполза*

- Oper1 (*польза*) = *приносить пользу*  
 Oper1 (*влияние*) = *оказываю влияние*  
 Oper1 (*влияние*) = *оказывать влияние*  
 Oper1 (*поддержка*) = *оказываю поддержка*  
 Oper1 (*поддержка*) = *оказывать поддержку*  
 Oper1 (*спротива*) = *оказываю спротива*  
 Oper1 (*сопротивление*) = *оказывать сопротивление*  
 Oper1 (*помощь*) = *оказываю помощь*  
 Oper1 (*помощь*) = *оказывать помощь*  
 Oper1 (*услуга*) = *оказываю услуга*  
 Oper1 (*услуга*) = *оказывать услугу*  
 Oper1 (*участие*) = *вземаю участие*  
 Oper1 (*участие*) = *принимать участие*  
 Oper1 (*апелляция*) = *подаваю апелляция*  
 Oper1 (*апелляция*) = *подавать апелляцию*  
 Oper1 (*виз*) = *надаваю виз*  
 Oper1 (*крик*) = *поднимать крик*  
 Oper1 (*вой*) = *надаваю вой*  
 Oper1 (*вой*) = *поднимать вой*  
 Oper1 (*инъекция*) = *бия / поставям / правя*  
 Oper1 (*укол*) = *делать*

**Oper 2 = [to undergo the action of X or to be in the scope of X (a collocate verb taking P<sub>2</sub> as its grammatical Subject and X as its main Object)]**

- Oper2 (*польза*) = *извлекаю польза*  
 Oper2 (*польза*) = *извлекать пользу*  
 Oper2 (*опека*) = *намираю се под опека*  
 Oper2 (*опека*) = *находиться под опекой*

**Labor 1–2 = [To subject or expose P2 to X (a collocate verb taking P1 as its grammatical subject, P2 as its primary object and X as a secondary object)]**

- Labor1–2 (*критика*) = *подлагам на критика*  
 Labor1–2 (*критика*) = *подвергать критике*

**Real1 = [To use X according to its destination (a collocate verb taking P1 as its grammatical subject and X as its primary object)]**

- Real1 (*гитара*) = *свирия на гитара*  
 Real1 (*гитара*) = *играть на гитаре*  
 Real1 (*шах*) = *игря на шах*  
 Real1 (*шахматы*) = *играть в шахматы*

**Prepar = [To prepare X for use according to its destination (a collocate verb taking P0 or P1 as its grammatical subject and X as its primary object)]**

- Prepar (*легло*) = *постилам легло*  
 Prepar (*постель*) = *стелить постель*  
 Prepar (*маса*) = *слагам маса*  
 Prepar (*стол*) = *накрывать на стол*

## Bibliography

- [1] Апресян, В. Ю., Апресян, Ю. Д., Бабаева, Е. ЕРЕВ., Богуславская, О. Ю., Иомдин, Б. Л., Крылова, Т. В., Левонтина, И. Б., Санников, А. В., Урысон, Е. В. (2006). *Языковая картина мира и системная лексикография. Языки славянских культур*, Москва.

- [2] Apresjan, Ju. D. (1996). Enseignement du lexique assisté par ordinateur. In *Lexicomatique et dictionnaires. Ives Journées scientifiques du réseau thématique "Lexicologie, Terminologie, Traduction"*, pages 1–10, Lyon & Montréal.
- [3] Apresjan, Ju. D., Boguslavsky, I. M., Iomdin, L. L., Tsinman, L. L. (2007). Lexical functions in actual nlp-applications. In *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*, Wanner, L. (ed.), pages 203–234, Amsterdam. John Benjamins.
- [4] Апресян, Ю. Д. (1995). *Избранные труды. Том 1. Лексическая семантика. Синонимические средства языка*. Языки русской культуры, Москва.
- [5] Апресян, Ю. Д., Дяченко, П. В., Лазурский, А. В., Цинман, Л. Л. (2007). О компьютерном учебнике лексики русского языка. *Русский язык в научном освещении*, 2((14)):48–112.
- [6] Boguslavsky, I., Baggio, Rodriguez, M., Diachenko, P. (2006). Callex-esp: A software system for learning spanish lexicon and collocations. *Current developments in Technology-Assisted Education*, 1:22–26.
- [7] Дяченко, П. В. (2008). *Разработка компьютерных методов обучения владению языком с помощью аппарата лексических функций. Дисс. ... к.ф.н.* Москва.
- [8] Крылов, С. А. (2005). Влияние МСТ на общую лингвистику (в свете подведения предварительных итогов «семантической революции»). In *Восток – Запад: Вторая международная конференция по модели "Смысл «-» Текст": Языки славянской культуры*, pages 237–247, Москва.
- [9] Мельчук, И. А. (1995). *Русский язык в модели «Смысл ↔ Текст»*. Wiener Slawistischer Almanach, Москва – Вена: Школа.
- [10] Mel'čuk, I. A. (1996). Lexical functions: A tool for the description of lexical relations in a lexicon. In *Lexical Functions in Lexicography and Natural Language Processing*, Wanner, L. (ed.), pages 37–102, Amsterdam. John Benjamins.
- [11] Mel'čuk, I. A. (1998). Collocations and lexical functions. In *Phraseology, Theory, Analysis and Applications*, Cowie, A. P. (ed.), pages 23–53, Oxford. Clarendon Press.
- [12] Polguère, A. (2005). *DiCouèbe, Dictionnaire en ligne de combinatoire du français*. Université de Montréal.
- [13] Polguère, A. (2007). Lexical function standardness. In *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*, Wanner, L. (ed.), pages 43–96, Amsterdam. John Benjamins.
- [14] <http://search.dcl.bas.bg/>
- [15] <http://www.ruscorpora.ru/>

# Translation of Polish Uninflected Present Participle in Bulgarian Literature – on the Basis of *Pan Tadeusz* (*Mr Thaddaeus*) by Adam Mickiewicz\*

Joanna Satoła-Staškowiak

Institute of Slavic Studies, Polish Academy of Sciences, Poland

**Abstract.** The article describes the realization of sentences with present participles found in the Bulgarian translation of *Pan Tadeusz* by Adam Mickiewicz done by Blaga Dimitrova [11]. The possibilities of translating participle constructions observed in the translation have been classified by me into separate groups, divided with regard to linguistic character of translation.

## 1 Constructions with uninflected participles

Constructions with uninflected participles in Polish literature on the subject are called “*participle gerund clauses*” or “*participle declarative clauses*”. For a number of years a view was held that participle constructions with reference to a clause joined to them are always subordinate<sup>1</sup>, that is *semantically dependent on the main clause* [5, p. 184].

Other grammar authors, such as Z. Klemensiewicz [6], [7], [8], [9], [10], J. Kuryłowicz, D. Butlerowa [1], Z. Saloni [14], [15], A. Łęgowska-Grybosiowa [3] rather agreed with this view, or sometimes, like W. Doroszewski and B. Wiczorkiewicz [2], H. Wróbel [17], proposed a different approach not excluding subordination and coordination of participle constructions regarding a combined clause.

According to R. Grzegorzczkova, constructions of the type: *he walks limping “constitute the borderline between singular and compound sentences: they hide in themselves two predications”* [4, p. 83].

### 1.1 The occurrence of participle constructions in the Polish language and Bulgarian language

**In the Polish language** participle constructions with uninflected participles are still relatively often used. It is true that they are not as frequent as sentences with inflected participles but they are present in our grammar.

**In the Bulgarian language** participle constructions with uninflected participles occur less frequently. It is caused by the disappearance of a Bulgarian equivalent of perfect participle as well as the unnaturalness of style, which results from the use of present participle constructions.

### 1.2 Henryk Wróbel methods of research and classification

H. Wróbel, unlike many other linguists, distinguishes coordinate clauses (paratactical system) and subordinate ones (hypotactical) among sentences with participle constructions. He engages in polemics with other researchers.

The two systems isolated by Wróbel: paratactical and hypotactical, belong to participle constructions as equivalents of a clause.

---

\* The study and preparation of these results have received funding from the FP7 under grant agreement Mondilex.

<sup>1</sup> I carry out a review of scientific positions concerning the status of clauses with uninflected participles in a work “Clauses with participles in *Pan Tadeusz* and their equivalents in the Bulgarian translation by Blaga Dimitrova”. The book will soon be published by a WSHE printing house in Łódź.

## 2 The choice of my own method of description and classification

For the purposes of this article, I worked out my own method of description and classification of the material in question. In this undertaking I used a way of analysis carried out by Henryk Wróbel since he discusses most accurately the problems of participle constructions that I examine. It allows explicit realization of constructions, which enables paraphrase.

In my research, like Wróbel, I see a need for distinguishing two types of clauses with constructions consisting of uninflected present participles (paratactical and hypotactical clauses) and naming them respectively: concurrent and anterior.

For the purposes of this article I analyse only the concurrent group. The rest of the research will be presented in other author's works. Additionally, in the concurrent group I single out the types of translation of equivalents thus isolating groups comprising present participle in places where it occurs in the original as well as other equivalents where for some reason participle could not be applied and was expressed with the use of other linguistic means.

### 2.1 Participle constructions in the paratactic system

Constructions with present participles can be regarded as complementary sentences. At the same time they can be described as semantically coordinate to combined clauses but formally dependent. *“As a result of this formal dependence of participle, the text with participle construction is more coherent than a compound sentence ( in particular parentetical ), from which it was derived. It involves a greater freedom of word order of participle construction in comparison with a combined clause.”* [17]

Wróbel quotes examples where both forms of the verbs can be in turns transformed into participle construction, e.g.:

Tolo milczał, **patrzac** na niego z namysłem. (Tolo was silent, **looking** at him thoughtfully.)

Jakub stał z opuszczonymi rękami i patrzył na niego **milczac**. (Jacob was standing with lowered hands and looking at her, **keeping** silent.)

It is the sender of the message who decides what constitutes complementary information at a particular moment in these two statements. Naturally, sometimes the freedom of such operation- as the author observes- *“is restricted by the reasonableness of putting one action against a background of the other”*.

Bulgarian sentences are constructed differently. Below are examples where participles were replaced by a clause and a predicate remained unchanged, e.g.

**Давайки** висока оценка на всеотдайния и хуманен труд на здравните работници и **изразявайки** признателността на народа, партията и правителство определиха 7 април за ден на здравния работник. – **Като дават** висока оценка на всеотдайния и хуманен труд на здравните работници и **изразявайки** признателността на народа, партията и правителството определиха 7 април за ден на здравния работник. – **Давайки** висока оценка на всеотдайния и хуманен труд на здравните работници и **като изразяват** признателността на народа, партията и правителството определиха 7 април за ден на здравния работник [13].

Unfortunately, I have not found a single example in Bulgarian grammars, which would support what Wróbel claims. It is another proof of the differences between participle systems.

Wróbel noticed a certain group of verbs which are characterised by particular susceptibility to serving the role of a background for a different action. These are the following verbs: *stand, sit, lie, kneel, walk, go, drive, etc.*

E.g. Gadali ze sobą **stojac** grupkami w bramach albo przed zamkniętymi sklepikami **siedzac** na malutkich wyplatanych krzeselkach.

(They chatted **standing** in groups in gateways or **sitting** on tiny wicker chairs in front of closed small shops.)

Jakub nie odpowiadał, lecz słuchał, **lezac** oparty na łokciach.

(Jacob didn't answer but listened **lying** and leaning on his elbows.) (examples after Wróbel)

Such participle constructions resemble adverbial expressions of time, manner, etc. and are getting closer to them.

Among the paratactical systems, sentences with generally inclusive relation predominate. Two types of sentences are important here: - “where the second sentence is a more accurate version of the first sentence” (that is so-called inclusive sentences) - and parentetic sentences, introduced in some contexts with expressions such as: *and in addition to which, and before that*.

### 3 Description and classification of the material

#### 3.1 Concurrence in the original and in the translation

**Translation equivalent has the form of present participle** A group of equivalents having the form of present participle is not too numerous. It includes sentences in which it was possible to capture the correspondence between Polish present participle and Bulgarian participle on *-айки, -ейки* conveying the same content.

##### **Księga I [12]**

1.

Tak **mówiąc** na Sędziego *mrugał*, widać z miny,  
Że miał i tań inne, ważniejsze przyczyny.

**Говорейки** така, той *смигна* дяволито,  
но ясно бе, че тук е нещо друго скрито.

Among the sentences belonging to this group one can find examples- disturbances which convey modified content despite built-in correspondence participle- participle. It is because of the fact that the modification takes place on the level of Bulgarian predicate different from the original.

##### **Księga II**

1.

Hrabia jeszcze chwilę w miejscu *bawił*  
Śmiejąc się i **klnąc** razem tej nagłej przeszkodzie;

Пан Графът *се сконфузи*,  
**проклинайки** го скрито, дето му попречи.  
В градината погледна – няма никой вече.

2.

Już goście zgromadzeni w wielkiej zamku sali,  
**Czekając** uczty wkoło stołu *rozmawiali*,

Тълпа от много гости празничната зала,  
**очаквайки** гощавка, шумно *се е сбрала*.

Another irregularity occurring in the presented group is the condensation of content. As an example one can quote a sentence from Book XI in which the content conveyed by two present participles was narrowed and expressed by the construction with one participle.

##### **Księga XI**

1.

Na to Zosia *rzecze*  
**Wznosząc** głowę i **patrzac** w oczy mu nieśmiało:

Свян девойката изпитва.  
**Издигайки** очи, му *отговаря* скромно:

#### **Translation equivalent is not a present participle**

**Compound sentence in translation** When Polish present participle has its Bulgarian equivalent, which is a personal form of the verb, we can find two forms of a predicate in one sentence, which results in a compound sentence in translation. The formed compound sentences were linked by means of conjunctions: *u, a* or without conjunctions, separated only by commas.

**Conjunction *u* as an index of concurrence** This combination index characterizes inclusive coordinate clauses.

Content **S1** combines with content **S2**, semantic dependence between them is not very strong. According to *Граматика на съвременния български книжовен език* [16]: “conjunction *u* indicates the simplest links without emphasizing any special relations between them.”

#### Księga I

1.

Z wieku mu i z urzędu ten zaszczyt należy,  
**Idąc kłaniał się** damom, starcom i młodzieży;  
 заслужил тая чест по ранг и по години;  
 той **тръгва** и *се кланя*, край когото мине –  
 на дами и на старци, на младежи чинни.

One of the more important issues that should be looked into is the problem of aspect. It should be emphasized that the translator did not always have an opportunity to convey the content identical with the original. We can notice modifications of the translated material in many sentences. We can find aspect differences which arose between the original text and the translation.

#### Księga I

1.

Teraz ręce przy boku miał, w tył wygiął łokcie,  
 Spod ramion *wytknął* palce i długie paznokcie,  
**Przedstawiając** dwa smycze chartów tym obrazem;

Сега той пред гърди ръце на лакти скръсти,  
 под мишници *показа* с дълги нокти пръсти  
 и тъй **изобрази** два ремъка за хрътка

The participle in the original (derived from an imperfective verb) was expressed by an equivalent of participle- a perfective verb.

In other examples we can notice analogous situations:

#### Księga II

1.

Gdy skrzypiące stodoły drzwi otwarto z trzaskiem  
 I bernardyn ksiądz Robak *wszedł* z węzlastym paskiem,  
 “Surge, puer!” **wołając** i ponad barkami  
 Rubasznie **wywijając** pasek z ogórkami.

Вратата изтрещя. На прага сам Монаха  
**застана**: “Surge puer!” – **викна** и **размаха**  
 над момъка сънлив колана бял със възел  
 и свойски го зашиба, гъсти вежди свъсил.

**Conjunction *a* as an index of concurrence** This conjunction is typical of a certain kind of coordinate clauses which are called opposing clauses in the Polish linguistics. In these sentences content **S2** opposes content **S1**. This construction is perfectly shown by an example from Book I.

#### Księga I

1.

Spór był wielki, już potraw ostatnich nie jedli.  
 Stojąc i **pijąc** obie *klóciły się* strony,  
 A najstraszniej pan Rejent był zaciętrzewiony.  
 Забравят, че пред тях е пиршеводство голямо,  
*не вкусват* от блюдата, а на крак все **пият**.  
 Нотариусът вече е от гняв най – злият,  
 говори, без да спре, изпънал вратна жила  
 и жарката си реч със жестове подсиля.



### Comma as an index of concurrence

#### Księga I

1.  
 konie, porzucone same,  
**Szczurіаc** trawę *siągnęty* powoli pod bramę.  
 Сами оставени конете,  
**заскубаха** трева, така *с юзди несети*.

2.  
**Nucаc** *chwyciła* suknie, biegła do zwierciadła;  
**Zatanаника**, *грабна* роклята простряна,  
 към огледалото затича,

**Conditional clauses** *Грамматика на съвременния български книжовен език* describes conditional clauses as a kind of clauses “that expresses a relation in which one activity is dependent on the circumstances of doing the other activity”. Conditional accomplishing of one activity becomes the basis for accomplishing the other.

#### Księga II

1.  
 Kto z nas tych lat nie pomni, gdy, młode pachole,  
 Ze strzelbą na ramieniu świszcząc szedł na pole,  
 Gdzie żaden wał, płot żaden nogi nie utrudza,  
 Gdzie **przestępując** miedzę *nie poznasz*, że cudza!

Кой може да забрави времето, в което  
 щастлив, нарамил пушка, свиркал е в полето?  
 Ни синор, ни окоп младежите прокужда;  
**прекрачиш** ли межда – *не гледаш*, че е чужда.

- Ако прекрачиш - не гледаш, че е чужда

The content included in the hypothesis (crossing the baulk) conditions the accomplishment of the thesis (lack of awareness that the baulk is someone else's.)

**Singular clause in the translation** The group includes such cases where a structure derived from coordinate clauses is expressed by means of a singular clause. The content of the original was not modified in the majority of cases.

#### Księga II

1.  
 “[...]”  
 Kto łaska, proszę za mną“ - *rzekła*, koło głowy  
**Obwijając** czerwony szal kaszemirowy;  
 А който би дошел със мене, да побърза! –  
 С червен кашмирен шал косата си **завърза**.  
 На Подкоможи тя поведе дъщерята,  
 до глезен вдигайки полата си развята.  
 Lack of the equivalent of a predicate in the original.

**Clauses containing syntagmatic relations in the function of adverbials of manner as equivalents of participles**

**Adverbial function of manner expressed by adverb**

#### Księga VI

1.  
 Sędzia słuchając, z wolna okulary składał  
 I **wpatrując się** mocno w Księdza, nic *nie gadał*,  
 Westchnął głęboko, w oczach łzy się zakreśliły...
- А Съдията вслушан, сгъна очилата  
 и втрещено **се вгледа** в Робак *мълчешката*.  
 В очите му сълзи се бяха появили.

**Adverbial function of manner expressed by prepositional phrase** The sub-group presented here can be internally divided into two even smaller sub-groups. The first one includes prepositional phrases with abstractum, in which predicative content has been maintained whereas the second includes prepositional expressions, in which a noun relating to the subject is the modified content of the original.

#### Prepositional phrase with abstractum

- Księga I**
1.  
 Tak **mówiąc** *spójrzal* zyzem, gdzie wśród biesiadników  
 Siedział gość Moskal; był to pan kapitan Ryków;
- При тия хладни думи** Робак изпод вежда  
 към капитана руски бързешком *поглежда* –
2.  
 to **mówiąc**, ręce *ciągnął* wzdłuż po stole
- При тия думи** той ръцете *си протегна*

#### Prepositional phrase containing reflection of the content carried by participle

- Księga VII**
1.  
 W końcu wszyscy przez długą zaściankę ulicę  
*Puścili się* w cwał **krzycząc**: “Hajże na Sopleć!”
- Подковен звек *отеква*, дълга върволица  
 в галоп полита **с възглас**: Хей, срещу Соплица!

**Translation equivalent has the form of the passive participle** The presence of the passive participle in the translation, as part of a noun phrase, gives it the function of an incongruent attribute. Additionally, one should draw attention to the fact that with the passive participles a shade of resultativeness is possible, which places clauses with these constructions on a borderline between concurrence and anteriority. It is most noticeable in the examples of sentences of the type:

- Wysmukłą postać tylko aż do piersi *kryje*,  
**Odsłaniając** ramiona i łabędzią szyję.
- Стои девойка млада  
 във бяло облекло, като цветец в саксия,  
**с открити рамене** и лебедова шия.

An element of resultativeness is visible in the translation- the young girl's shoulders have been bared earlier. Now we can see the result of that activity- the shoulders are bare.

**Structure derived from coordinate clauses expressed by separate clauses** In this kind of sentences we are faced with a situation in which the coordinate construction of the predicate of

the main clause and participle has been separated. The content carried by participle is represented by a separate clause, not connected with a sentence containing the predicate of the original.

### Księga II

1.

Postrzegłem wtenczas kulę, wpadła w piersi same,  
Pan **słaniając się** palcem *ukazał* na bramę.

Но трясва в тоя миг гърмеж откъм вратата,  
пан Столникът **се люшва** и глава отмята,  
избликва от гръдта му кърваво поточе.  
Той бледен към вратата с пръста си *посочи*.

### 3.2 Concurrence in the original and its disturbance in translation

In the sentences classified into this group one can notice a shift of the concurrence in the original to the anteriority in the translation. The class is further internally divided into three sub-classes:

**Translation equivalent has the form of a participle on -л** In the Bulgarian language this participle has two forms: perfective- минало свършено причастие and imperfective — минало несвършено причастие. In the sentences I present the translator uses only one of them- the perfective form of the participle, replacing Polish present participle. As a result of this operation, concurrent content represented by the Polish sentence was moved to anteriority in the Bulgarian sentence. The Bulgarian participle on -л has built-in features of anteriority regarding a different activity which is not expressed by this participle.

### Księga II

1.

**Wstrzymując** oddech, usty *chwycił* jej westchnienie  
I okiem łowił wszystkie jej wzroku promienie.

Дихание горещо, дъх **стайл** *погълна*  
и поглед впи в зеница, с парещ блясък пълна.

**Translation equivalent has the form of a sentence in perfectum** The past indefinite tense-perfectum consists of the participle of the past perfective tense and personal forms of an auxiliary word “*съм*” in the present tense. It only shows the fact that a certain activity took place in the past without specifying the moment, time correlation and accompanying circumstances, as it happens in other tenses (e.g. aorist). Perfectum most often indicates a resultative state.

### Księga II

1.

Dalej maków białawe górują badyle;  
Na nich, myślisz, iż rojem *usiadły* motyle  
**Trzepiocząc** skrzydełkami, na których się mieni  
Z różnaitością tęczy blask drogich kamieni;

И макови стебла навред с ръка досягаш –  
*накацали* по тях са пеперуди сякаш  
и **пърхат** със крилца, обагрени във шарки  
на писана дъга и на рубини ярки -

**Translation equivalent has the form of anterior clause, thanks to the index -цом** According to *Граматика на съвременния български книжовен език* “index of the sequentiality of two activities can be a conjunction *цом* (...) Sentences with this conjunction express quick, direct sequence of the second activity.”

**Księga I**

1.

**Widząc** gościa, na folwark *dążył* po kryjomu**щом чу** за госта нов, от смайване забързан,  
из стаите странични *се укри***Translation equivalent corresponds to the form of the conditional Księga II**

1.

**Mogąc** żyć u Hrabiego na łaskawym chlebie,  
*Nie chciał*, bo wszędzie tęsknił i czuł się niezdrowym,  
Jeżeli nie oddychał powietrzem zamkowym.Постлал си е в една от стаите пустинни,  
а **би могъл** при Графа даром да помине,  
но старият прислужник чак се поболява  
без въздуха на тази сграда величава.**4 Summary**

The above comparison of the Polish and Bulgarian linguistic material has proved how many differences divide both languages despite the fact that they belong to one group of Slavonic languages.

A point of departure for this work became Polish sentences with present participles which I later confronted with their Bulgarian translation.

Despite the fact that the Bulgarian language possesses present participle- *деепричастие*, I have found few examples of accurate translation: participle- participle. I do not know exactly what the reasons for such state of affairs are. The situation may be dictated by the specificity of the text studied- *Pan Tadeusz* and its rhythmicity. It may also result from limited distribution of this participle in spoken Bulgarian. Most often Polish constructions with the present participles after translating into Bulgarian did not change semantic or functional relation existing between the elements of initial structures (however, we can also find even such cases, cf. successive articles). Most often Polish sentences with these participles were also transformed into Bulgarian compound coordinate clauses, with indicators of junction: *i*, *a* or without conjunction, which correctly highlighted the character of the present participles. A group of equivalents of the present participles in the Bulgarian translation by Blaga Dimitrova was very diverse. Apart from the above mentioned options, there were also equivalents in the form of prepositional phrase, functioning as adverbials of manner. I have also found a small group of equivalents- passive participles placing the content on the borderline between concurrence and anteriority.

**Bibliography**

- [1] Buttler, D. (1971). Zasady poprawnego użycia imiesłowowych równoważników zdań, Siatkiewicz, H., Kurkowska, H., Buttlera, D. (eds.). In *Kultura języka polskiego*, pages 412–421, Warszawa.
- [2] Doroszewski, W., Wieczorkiewicz, B. (1959). *Gramatyka opisowa języka polskiego z ćwiczeniami, t. II*. Państwowe Zakłady Wydawnictw Szkolnych, Warszawa.
- [3] Grybosiova, A. (1978). *Rozwój funkcji imiesłowów nieodmiennych w języku polskim. Związki z nomen*. Wrocław.
- [4] Grzegorzyczkowa, R. (1999). *Wykłady z polskiej składni*. Wydawnictwo Naukowe PWN, Warszawa.
- [5] Jadacka, H. (1994). Próba określenia normy składniowej dotyczącej użycia równoważników imiesłowowych na *-ąc*. In *Polszczyzna i/a Polacy u schyłku XX wieku*, Handke, K., Dalewska-Greń, H. (eds.), pages 97–113, Warszawa.

- [6] Klemensiewicz, Z. (1937). *Składnia opisowa współczesnej polszczyzny*. Kraków.
- [7] Klemensiewicz, Z. (1963). *Zarys składni polskiej, wyd. IV*. Wydawnictwo Naukowe PWN, Warszawa.
- [8] Klemensiewicz, Z. (1980). *Historia języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- [9] Klemensiewicz, Z. (1982). *Składnia, stylistyka, pedagogika językowa*. Wydawnictwo Naukowe PWN, Warszawa.
- [10] Klemensiewicz, Z. (1983). *Podstawowe wiadomości z gramatyki języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- [11] Мицкевич, А. (1979). *Пан Тадеуш. Шляхтишка история от годините 1811–1812 в дванадесет стихотворни книги*. Народна Култура, София.
- [12] Mickiewicz, A. (1986). *Pan Tadeusz czyli Ostatni Zajazd na Litwie. Historia szlachecka z roku 1811 i 1812 we dwunastu księgach wierszem*. Książka i Wiedza, Warszawa.
- [13] Москов, М. (1974). *Български език и стил*. Наука и изкуство, София.
- [14] Saloni, Z. (1971). *Błędy językowe w pracach pisemnych uczniów liceum ogólnokształcącego. Próba analizy językoznawczej*. Warszawa.
- [15] Saloni, Z., Świdziński, M. (1998). *Składnia współczesna języka polskiego, wyd. IV*. Wydawnictwo Naukowe PWN, Warszawa.
- [16] Стоянов, С., (ed.) (1983). *Граматика на съвременния български книжовен език в три тома. Том II Морфология. Том III Синтаксис*. София.
- [17] Wróbel, H.: (1975). *Składnia imiesłowów czynnych we współczesnej Polszczyźnie*. Prace Naukowe Uniwersytetu Śląskiego w Katowicach, Uniwersytet Śląski, Katowice.

# Exponents of Adnumeral Approximation in Polish and Russian\*

Maksim Dushkin

Institute of Slavic Studies, Polish Academy of Sciences, Poland

**Abstract.** This paper will attempt at presenting the basic theoretical problems that can be encountered while trying to procure a list of Polish and Russian exponents of adnumeral approximation.

These exponents are, primarily, some lexemes, indicating approximate numbers in conjunction with numeric expressions or names of units of measurement. These are, for example, Polish exponents *około* (*Jaś naliczył około 30 osób, siedzących przy biurkach wzdłuż sali*), *przeszło* (*W ciągu roku statek "Gwiazda morską" przewiózł przeszło 10 tysięcy pasażerów*), Russian lexemes *более* (*Обычно Вася едет на работу более часа*), *примерно* (*Вася подождал автобуса примерно полчаса и решил идти пешком*) etc. Exponents of approximation can also be used in constructions, such as conjoining two numerals, like *5-10* (*5-10 человек, 5-10 osób*), numeric expressions like *sto kilkadziesiąt* and so on.

The exponents of approximation were analysed "from content to form". Approximation can be understood as a type of information, linguistically transmitted by various means. First, types of content that can be labeled as "approximation" were defined, and then an attempt was made at establishing, what linguistic devices are used to express that content in Polish and Russian.

1. Approximation is widely connected with numbers, but it can also be understood in a broader context ([6], [12], [10]). An approximate description requires a predicate and a special exponent of approximation. This exponent denotes that the described state differs or could differ from the state communicated by the predicate (Cf. *Jan ma 30 lat* and *Jan ma około 30 lat* <Jan has> 'little less than 30 or 30 or little more than 30 years'). If a numeric predicate is used, the approximation is called numeric (adnumeral), e.g. Pol. *Ekipa pokonała mniej więcej 200 kilometrów piechotą*; Rus. *Команда преодолела примерно 200 километров пешком*. If the predicate is non-adnumeral, the approximation has a broader meaning (e.g. Pol. *Do pokoju weszły dziewczyny wyglądające mniej więcej jednakowo*; Rus. *В комнату вошли девушки, выглядевшие примерно одинаково*). In the broader context, approximation can denote not only designates of numbers, but also of "any states" [6, p. 29].

This paper, however, will only be concerned with adnumeral (numeric) approximation. Such approximation should be juxtaposed with numeric accuracy. Both accurate and approximate descriptions of numbers refer to the **arithmetic sequence** (Cf. *5* and *około 5*), but in a different manner. Precise descriptions of numbers refer to a specific point in the sequence (e.g. *50*). Approximate descriptions of numbers, on the other hand, refer to a segment of the sequence. A segment is a set of points within the sequence. One of these points is a correlate of the specified number. However, the point itself is unknown (*bez mała 50* '... 47 or 48 or 49').

2. This definition of numeric approximation (as a way of referring to the arithmetic sequence) allows to separate some types of descriptions, which are treated (but perhaps should not be) by some researchers as approximates.

First, the inaccurate use of definite "round" numerals, such as Pol. *W Polsce mieszka 40 milionów osób*, Rus. *В Польше живет 40 миллионов человек* should be separated from the field of approximation. Such usage has no formal indicators of inaccuracy, while the "round" numeral

---

\* The study and preparation of these results have received funding from the FP7 under grant agreement Mondilex.

does not indicate an accurate number (the phrase Pol. *40 milionów*, Rus. *40 миллионов* does not indicate precisely 40.000.000 with the exception of all other numbers, such as 40.000.001). Such descriptions are not normally separated from accurate descriptions of numbers by researchers. Sannikov 1999 treats them as approximate expressions. This article treats them as “rounded” usage and separates from the field of approximation. The “roundedness” is based on reference to the sequence, points of which are numbers with orders of magnitude higher than “1” (the points are tens, hundreds, thousands, etc.). What is counted is whole orders, e.g. millions, but smaller numbers are not “calculated”. Approximate descriptions of numbers (consisting of a numeral and an exponent of approximation) could also be rounded, if counted with units of multiples of tens. Cf. the ambiguity of the description *około 40 milionów osób*:

1. ordinary approximate description: ‘little less than 40.000.000 or 40.000.000 or little more than 40.000.000’ (people)
2. “rounded” approximate description: ‘little less than 40 or 40 or little more than 40’ (million people)<sup>1</sup>.

Secondly, expressions such as *dużo/m mało*, recognized by Grochowski [6, p.29–31] as exponents of approximation, were left out from the field of exponents of approximation. Such expressions do not refer to the arithmetic sequence (the infinite strain of numbers, starting from 1 up). They express subjective quantitative assessment, based on quantitative comparison of the type: such number/quantity is greater/lower than the number/quantity, that would not attract attention [2]<sup>2</sup>.

**3.** An analysis of works by other researchers concludes that Polish and Russian exponents of approximation set the segment of the arithmetical sequence (or the parameter scale) in different ways:

- they set the lower boundary of the segment exclusively (e.g. Pol. *ponad*, Rus. *свыше*) or inclusively (e.g. Pol. *co najmniej*, Rus. *не меньше*)
- they set the upper boundary of the segment exclusively (e.g. Pol. *niespełna, mniej niż*, Rus. *без малого, менее*) or inclusively (e.g. Pol. *co najwyżej*, Rus. *не больше*)
- they set both boundaries of the segment (*20-30*)
- they set the centre of the segment (Pol. *około*, Rus. *около*).

So, 6 methods of defining the segment are available. All of these have their exponents in Polish and Russian.

**4.** Individual exponents of numeric approximation differ not only in the way they define the segment they are representing, but also in several other features.

**4.1.** Simply speaking, two types of expressions are in most cases considered to be exponents of approximation - both (a) expressions that can only be connected with numerals and numeric expressions, e.g. *około 5*, and (b) expressions that can be connected not only with numerals and numeric expressions, e.g. *prawie: prawie 5* and *prawie pusty*. By conjoining with numerals and other numeric expressions, these expressions designate a numeric segment.

Some authors have made only expressions of the former type, i.e. the ones that can only be connected with numerals and numeric expressions, an object of their study [9], [5]. This allows to standardize the field of analysed expressions, not only with regard to their syntactic connectivity, but also semantically: this leads to separation of expressions, whose meaning is connected with the concept of number and the definition of a segment of numbers, and no other concepts. Such expressions are exponents of only numeric approximation.

<sup>1</sup> More on the phenomenon of roundedness, see: [4].

<sup>2</sup> Cf.: “[...] *A is tall*= *A’s size* (possibly: *which is conspicuous*) *is greater than his possible size which would not attract attention*” [2].

**4.2.** It is noticeable, that some researchers only consider expressions which, in connection with a numeral, set **small segments of numbers** (e.g. [12]) to be exponents of approximation, while others treat expressions which, while setting segments, do not **denote their size** (e.g. [3], [9], [6]). For example, *około*, *bez mała* are expressions, setting only a small segment; while *co najmniej* sets a segment, but does not determine if the segment is small.

**4.3.** An attempt was made to come up with the list of Polish and Russian expressions, designating a numeric segment, and to divide them according to the following criteria:

1. connectivity or non-connectivity with expressions other than numeric expressions
2. the lack or presence of the component 'little' (information about the small size of the designated numeric segment or the smallness of the difference between the state communicated via the predicate and the factual state).

Accepting these criteria allowed to separate four groups of exponents of adnumeral approximation:

- expressions, appearing only in conjunction with numeric expressions, and carrying the component 'little' within their meaning (e.g. *około*);
- expressions, appearing only in conjunction with numeric expressions, but not carrying the component 'little' within their meaning (e.g. *ponad*);
- expressions, appearing not only in conjunction with numeric expressions, but also many others, at the same time carrying the component 'little' within their meaning (e.g. *prawie*)<sup>3</sup>;
- expressions, appearing not only in conjunction with numeric expressions and not carrying the component 'little' within their meaning (e.g. *więcej niż*).

For the division of known exponents of adnumeral approximation into the abovementioned groups, see table 1 on page 192<sup>4</sup>.

What is most interesting in this classification, is the observation that some exponents can only be connected with numeric expressions (they are strictly adnumeral), while at the same time they do not contain the component informing about a small difference between the factual state and the state communicated by the predicate (e.g. *ponad 5*). Moreover, some other exponents contain the component of this small difference between the factual state and the state communicated by the predicate, but they are not strictly adnumeral (e.g. *prawie 5*, *prawie zielony*), and as such they are not in the field of strictly numeral approximation.

It should be noted that taking the difference between the components, concerning the semantic component 'little', into account, is especially important while translating. Translating an exponent, not containing the component 'little', using an equivalent, containing this component (and vice versa), changes the meaning of a sentence, cf.: *В каждой из групп без малого 20 человек* (it could be 18, 19, but not 10) and *В каждой з групп jest mniej niż 20 osób* (it could be 14, 18, 19 and even 10).

**5.** Finally, the specifics of descriptions of numbers, designating a unit of measurement, should also be mentioned. It turns out that exponents of approximation appear mostly with names of round numbers (cf. e.g. *około 15 osób* vs. *?około 13 osób*), but this phenomenon is neutralised when a numeral designates the number of some units of measurement, cf. e.g. *\*około 2 osób* vs. *około 2 litrów*, *około 2 godzin*, *około 2 kilometrów*. In other words, exponents of approximation could appear alongside not only round numerals, but also unround ones, when the numeral designates a number of units of measurement. Authors of some works (e.g. [12]) mention the limited connectivity of exponents of approximation with unrounded numerals, but none of them mention the specificity of the context of units of measurement.

<sup>3</sup> Wierzbicka [12] describes the meaning of not only adnumeral expressions, using the broader concept of 'difference' and not the narrower concept of quantitative difference of the 'more' – 'less' type.

<sup>4</sup> It should be noted that the table does not contain exponents of the sense 'X-Y', because some specific problems that merit a larger description are connected with this group.



A special case of the peculiarity of connections, expressing the approximate number of units of measurement, can be seen in conjunctions, containing the ellipsis of the numeral 1 (*około metra, przeszło godzinę, ponad kilometr* etc.) or the numeral 1 (*około jednego litra*).

The specificity of conjunctions of exponents of approximation with names of units of measurement stems from the characteristics of these units. A unit, whose name is in the description, could be – at least theoretically – considered a sum (a number) of units ten/hundred etc. times smaller (a *meter* for example, as sum of 10 ten-times smaller units – *10 decimeters*). Units of measurement are objects that can be counted and, on the other hand, divided into smaller objects.

The abovementioned examples of the peculiarities of approximate descriptions in relation to units of measurement are characteristic of both Polish and Russian.

6. To summarize, it should be stated that approximation is a mechanism of numeric designation, based on defining a segment instead of one point of the arithmetic sequence. Thus defined, the concept has numerous exponents in both Polish and Russian.

The most significant differences between the exponents have been discussed in this paper and a classification, accounting for these differences, has also been presented.

Table 1. Division of exponents of approximation

Sense	Exponents with the component 'little'		Exponents without the component 'little'	
	Group A (Adnumeral)	Group B (not only adnumeral)	Group C (Adnumeral)	Group D (not only adnumeral)
'<little> less than X or X or <little> more than X'	Pol. <i>circa, gdzieś, jakieś, około, około, rzędu, z;</i> Rus. <i>где-то, около, порядка, с</i>	Pol. <i>mniej więcej, plus minus, w przybliżeniu</i> Rus. <i>приблизительно, примерно</i>	–	–
'<little> less than X'	Pol. <i>blisko, niepełna; pod</i> Rus. <i>под</i>	Pol. <i>bez mała, niecały, niemal, prawie</i> Rus. <i>без малого, едва (ли) не неполный, чуть (ли) не</i>	–	Pol. <i>mniej niż; poniżej;</i> Rus. <i>менее (чем), меньше (чем)</i>
'<little> more than X'	Pol. <i>przeszło</i> Rus. –	–	Pol. <i>ponad, po;</i> Rus. <i>свыше; за</i>	Pol. <i>więcej niż; powyżej;</i> Rus. <i>более (чем), больше (чем)</i>
'X or <little> less than X'	–	–	Pol. <i>do</i> Rus. <i>до от силы</i>	Pol. <i>co najwyżej, maksimum, najwyżej, nie więcej niż;</i> Rus. <i>максимум, не более (чем), не больше (чем), самое большее</i>
'X or <little> more than X'	–	–	–	Pol. <i>co najmniej, minimum, najmniej, nie mniej niż, przynajmniej;</i> Rus. <i>как минимум, минимум, не менее (чем), не меньше (чем), по меньшей мере, самое меньшее</i>

## Bibliography

- [1] Bogusławski, A. (1966). *Semantyczne pojęcie liczebnika i jego morfologia w języku rosyjskim*. Zakład Narodowy im. Ossolińskich, Wrocław.
- [2] Bogusławski, A. (1994). Measures are measures in defence of the diversity of comparatives and positives. In *Sprawy słowa (Word matters)*, pages 323–329, Warszawa. Veda.
- [3] Bogusławski, A., Karolak, S. (1973). *Gramatyka rosyjska w ujęciu funkcjonalnym*. Wiedza Powszechna, Warszawa.

- [4] Duszkin, M. (2003). Dokładniej o dokładności: “zaokrągloność” dokładnych określeń ilościowych. *Prace Filologiczne*, XLVIII:133–142.
- [5] Grochowski, M. (1996). O wykładnikach aproksymacji: liczebniki niewłaściwe a operatory przyliczebnikowe. In *Studia z leksykologii i gramatyki języków słowiańskich, IV Polsko-Szwedzka Konferencja Slawistyczna, Mogilany 1–3 października 1995*, Wróbel, H. (ed.), pages 31–37, Kraków. Polska Akademia Nauk, Instytut Języka Polskiego.
- [6] Grochowski, M. (1997). *Wyrażenia funkcyjne. Studium leksykograficzne*. Polska Akademia Nauk, Instytut Języka Polskiego, Kraków.
- [7] Koseska-Toszewa, V. (1991). O języku-pośredniku i badaniach konfrontatywnych. In *Problemy teoretyczno-metodologiczne badań konfrontatywnych języków słowiańskich*, Běličova, H., Nieszczimienko, G., Rudnik-Karwatowa, Z. (eds), pages 7–19, Warszawa. Instytut Słowianoznawstwa PAN.
- [8] Koseska-Toszewa, V. (1993). *Gramatyka konfrontatywna rosyjsko-polska. Składnia*. SOW, Omnitech Press, Warszawa.
- [9] Мельчук, И. А. (1985). Поверхностный синтаксис русских числовых выражений. *Wiener Slawistischer Almanach*, 16.
- [10] Сахно, С., Л. (1983). Приблизительное именование в естественном языке. *Вопросы языкознания*, 6:29–36.
- [11] Санников, В., З. (1999). Русский язык в серкале языковой игры. *Языки русской культуры*.
- [12] Wierzbicka, A. (1991). *Cross-Cultural Pragmatics. The Semantics of Human Interaction*. Mouton de Gruyter, Berlin, New York.

# Definitions of Prepositions, Conjunctions and Particles in the Explanatory Dictionaries

Oleg Bugakov

Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine

**Abstract.** Definition formulas of prepositions, conjunctions and particles in the Ukrainian Language Dictionary are defined. The peculiarities of presenting lexical semantics for these parts of speech in the explanatory part of the dictionary entry are described.

**Keywords:** explanatory dictionary, Ukrainian Language Dictionary, preposition, conjunction, particle, lexical semantics, meaning, definition formula.

A 20-volume explanatory Ukrainian Language Dictionary (ULD) as well as its electronic version is now created in the Ukrainian Lingua-Information Fund (ULIF) of the National Academy of Sciences of Ukraine. So far as a large group of lexicographers is working on this dictionary, the problems have appeared with developing definitions and selecting the illustrations for each meaning, especially for the polysemantic words. So the necessity of creating the unified definition formulas for each part of speech and for various lexical-semantic word groups has appeared.

The investigation is concerned to the prepositions, conjunctions and particles. Each of these parts of speech is not large and seems to be simple word class. But there are a number of peculiarities in the definitions for each of them.

For each representative of these classes we specify its belonging to the part of speech in the shortened form just after the register word in the left part of the dictionary entry: “прийм.” – for prepositions, “спол.” – for conjunctions, “част.” – for particles. Let’s see each class separately.

## *Prepositions*

The explanatory (right) part of the register prepositions in the ULD is divided to the meanings, shades of the meaning and illustrations according to the principles for the lexicographic description of the register words lexical semantics:

**БІЛЯ**, *прийм.* з *род. в.* 1. Уживається на позначення невеликої відстані між кимсь, чимсь; коло, близько. ...

The peculiarity of the right part of the prepositions dictionary entries is that this part is divided into 2 fragments for the description of the simple prepositions polysemy. The first fragment begins with the phrase “Сполучення з + (реєстровий прийменник) + виражають:” (“Combinations with + (register preposition) + express”):

**ДО<sup>1</sup>**, *прийм.*, з *род. в.* Сполучення з **до** виражають:

The colon finishes the first fragment of the explanatory part for the register preposition.

The second fragment of the explanatory part describes quasilexical semantics of the register preposition and has through numbering of the meanings – semantic relations that a preposition expresses. All meanings are divided into parts – semantic relation types. Each of them has a name that is given with expanded spacing between letters. The names of semantic relation types are given in the nominative case, plural:

Об’єктні відношення (Objective relations)

Просторові відношення (Spatial relations)

Часові відношення (Temporal relations)

...

In the limits of a meaning there may be shades of the meaning that are marked with “//” (mark of shade of the meaning) and/or letters of the Ukrainian alphabet – а), б) etc., when the meaning generalize quasilexical semantics of the preposition and require its concretization, specification or detailed description. Thus, for example, the explanatory part for preposition

**В<sup>2</sup> (У)**, *рідко УВ* – *із знах., місц. і род. відмінками; ц-с., уроч.* **ВО**, *рідко ВВІ (УВІ)*, **ВІ** – *із знах. і місц. відмінками; прийм.*

has the following structure: initially the first fragment is given –

“Сполучення з **в(у)** виражають:”,

then the semantic relation types that unify the meanings are given:

Просторові відношення

1. *із знах. і місц. в.* Уживається на позначення предмета, місця, простору: а) всередину якого, куди спрямована дія (*знах. в.*) .. б) в якому, де відбувається дія чи хто-, що-небудь міститься, перебуває (*місц. в.*) ...

Об'єктні відношення

6. *із знах. і місц. в.* Уживається на позначення дії: а) в яку хто-небудь включається, втручається (*знах. в.*) ... б) в якій хто-небудь бере участь (*місц. в.*) ...

7. *із знах. і місц. в.* Уживається на позначення сфери діяльності, організації, установи і т. ін.: а) до якої вступає, приступає, куди переходить або приймається хто-небудь (*знах. в.*) ... б) в якій бере участь, де працює, вчиться хто-небудь (*місц. в.*) ...

In the 11-volume explanatory Ukrainian Language Dictionary that is a basis for creating a new dictionary there are 15 semantic relation types that prepositions can express: temporal, quantitative, quantitative-determinative, objective, objective-adverbial, adverbial, determinative-adverbial, determinative, relations of *modus operandi*, relations of goal, relations of measure, causative, spatial, conditional and concessive.

When a preposition has one or some meanings and mainly expresses semantic relations that belong to one type, the division into parts with pointing the semantic relations is not given. That is each meaning is given as in the entries of other parts of speech:

**З-ПО#ЗА**, *прийм.*, *з род. в.* Уживається при вказуванні на предмет, з протилежного або зворотного боку якого спрямовано рух, дію.

When defining concrete lexical meaning of the preposition, that is semantic state of the preposition, the following parameters are important: grammatical meanings of the main and dependent words, lexical meanings of the main and dependent words, semantic relation that a preposition expresses, and a case by means of which preposition manages a noun [1].

The universal definition formula for preposition looks as follows. The definition begins with the words: “Уживається на позначення / при вказуванні...” (Used for determination / for specifying). Then information on the lexical and grammatical meanings of the dependent word is given. After that information on the lexical and grammatical meanings of the main word is given. But grammatical information on the main and dependent words may be specified implicitly [1]. Thus, the number of preposition meanings depends on the number of possible sets of the meanings pointed above. For example, one of the meanings of the preposition *mi1zh* (between) looks as follows:

“6. *із знах. в.* Уживається на позначення місця, сукупності предметів чи осіб, куди спрямована дія”: *застромити між пальці, дертися між каміння.*

The main word is a verb or a noun with a meaning of movement, and the dependent word is a noun with a meaning of place, space or a subject. The main word semantics is more important for this meaning. It's also important that only accusative case is chosen from the three possible cases (genitive, accusative and instrumental), by means of which the preposition manages the dependent word.

### **Conjunctions**

The description of lexical (quasilexical) semantics of conjunctions also has several peculiarities. Particularly, the conjunction category should be specified:

1. If a conjunction has one meaning, its category is specified in the left part of the dictionary entry next to the remark “спол.” – conjunction:

**АЙ<sup>2</sup>**, *спол. протиставний*. Уживається для поєднання сурядних речень або членів речення із значенням протиставлення; та, але.

2. If a conjunction has several meanings that concern to the same category, its category is also specified in the left part of the dictionary entry:

**АДЖЕ#<sup>2</sup>**, *спол. причиновий* 1. Поєднує підрядне речення причини з головним; тому що, бо...

2. Приєднує речення, яке служить обгупрунтуванням думки попереднього речення.

3. If a conjunction has several meanings that concern to the different categories, its category is specified in the right part of the dictionary entry at the beginning of each meaning in italics:

**АБИ**, *спол.* 1. *умовний*. Починає підрядні речення умови; коли б лише, тільки б. ...

2. *мети*. Починає підрядні речення мети; щоб. ...

4. If conjunction meanings concern to the different categories, and there are several meaning in the limits of one category, then additional numbering with the Roman numerals is added. But in the limits of one category the numbering with Arabic numerals begins from 1:

**А<sup>2</sup>**, *спол.* I. *протиставний*. 1. Поєднує речення, протиставні за змістом одне одному; значенням близький до але, проте, навпаки, та...

2. Поєднує речення (або члени речення), не відповідні одне одному змістом, причому зміст другого суперечить сподіваному змістові, що впливає з першого; але, проте, однак, та...

II. *зіставний*. Поєднує члени речення або й цілі речення, в яких зіставляються одночасні дії; значенням наближається до тим часом, у той же час...

III. *приєднувальний*. 1. Приєднує нові речення або члени речення при послідовному викладі думок, описі ряду предметів чи явищ. ...

The universal definition formula for conjunction looks as follows. The definition begins with the verb in the third person in the present tense: “Поєднує...”, “З’єднує...”, “Приєднує...”, “Починає...” (“Combines...”, “Adds...”, “Begins...”) etc. Then there is an indication to the type of the sentence or the word that are combined. After that the peculiarities of the meaning are specified:

**АДЖЕ#<sup>2</sup>**, *спол. причиновий* 1. Поєднує підрядне речення причини з головним; тому що, бо...

2. Приєднує речення, яке служить обгупрунтуванням думки попереднього речення.

### **Particles**

The definition of the particle contains the peculiarities that define its functional-semantic meaning in the modern Ukrainian language. As for the conjunctions, the category should be specified:

**АЯ#КЖЕ**, *част., розм.* 1. *ствердж.* Уживається для ствердження якоїсь думки; авжеж, звичайно.

2. *запереч.* Уживається для вираження незгоди з чим-небудь, заперечення, відмови.

The definition begins with the words: “Уживається для творення...” (Used for forming...), “Уживається для означення...” (Used for determination...), “Уживається при вказуванні...” (Used for specifying), “Уживається на початку / в кінці речень...” (Used at the beginning / at the end of the sentence) etc.:

**БИ**, *після голосного Б*, *част.* 1. Уживається для творення дієслівних форм умовного способу. ...

2. Уживається для означення бажаності або можливості здійснення дії, вираженої дієсловом. ...

Besides the definition formulas for prepositions, conjunctions and particles described above there is one more way for representation of definitions for these parts of speech. If there are synonymous relations between the pairs of prepositions, conjunctions and particles, the following definition formulas are used.

For prepositions there is a definition formula “Уживається у знач. прийм.” (Used in the meaning of prep.) + specifying the word (words) with expanded spacing between letters:

**БЕЗ<sup>1</sup>**, *прийм.*, *з род. в.*

2. Уживається у знач. прийм. к р і м . . . .

For conjunctions there is a definition formula “Уживається у знач. спол.” (Used in the meaning of conj.) + specifying the word (words) with expanded spacing between letters:

**А<sup>2</sup>**, спол. . . .

В. *еднальний, діал.* Уживається у знач. спол. і . . . .

For particles there is a definition formula “Уживається у знач.” (Used in the meaning of) or “Уживається у знач., близькому до” (Used in the meaning close to) + specifying the word (words) with expanded spacing between letters:

**БИ**, після голосного **Б**, част. . . .

4. *розм.* Уживається у знач. м о в , н е м о в , н і б и , б у ц і м т о .

The definition formulas described will help lexicographers to make definitions more precise. Of course such definition formulas are made for other parts of speech. It's also important to make such formulas for various semantic word groups, for example, plants, animals, transport, minerals etc.

## Bibliography

- [1] Bugakov, O. (2008). Semantic states of ukrainian prepositions. *Études Cognitives, Warszawa*, 8.
- [2] Бугаков, О. В. (2008). Формулы толкований предлогів для словаря українського мови. In *MeqLing'2007. Прикладна лінгвістика та лінгвістичні технології. Сбірник наукових праць*, pages 94–101, Київ.
- [3] Широков, В. А. (2005). *Елементи лексикографії*. Київ.
- [4] Широков, В. А., Бугаков, О. В., Грязнухина, Т. О., та ін (2005). *Корпусна лінгвістика*. Київ.
- [5] СУМ (1970–1980). *Словник української мови. В. 11 т.* Київ.

# Quelle description pour les préverbes polonais ?

Ewa Gwiazdecka

Université de Varsovie

**Abstract.** In this paper, we are trying to provide the semantic description for Polish verbal prefix. We are particularly interested in the prepositional “history” of this operator and the way the diachronic relation contributes to aspectuality. We use topological operators for description of spatiality, temporality and activities encoded by preposition and verbal prefix. The topological intervals serve to represent basic aspectual situations: *state*, *process* and *event*. These aspects are regarded as a property of the whole predicative relation. We claim that in the binary predicative relation,  $P_2T^2T^1$ , the compositionality between verbal prefix and the second argument  $T^2$  is of aspectual relevance. It follows that aspect encoded by prefixation does not focus on the verb only, but it can determine a spatial place or an object. In fact, in the majority of the cases, verbal prefix encodes an achievement understood as a closure of the right boundary of the *process* associate with the predicate, which coincides with the focus on *some* topological zone. Achievement is understood qualitatively and does not always implies the “completion” of action, it can indicate “achievement of the beginning”, “achievement of the end phase”, etc.

## 1 Le dictionnaire aspectuel des verbes

La description de l’aspect perfectif engendré par le préverbe pose plusieurs problèmes. Le lien diachronique et sémantique avec la préposition fait que le préverbe, outre les changements aspectuels introduit très souvent des modifications de signification dans le verbe de base. Au surcroît, plusieurs préverbes peuvent participer à la formation d’une paire aspectuelle et le choix du préverbe ne présente aucune régularité morphologique. Ainsi, certains verbes possèdent plusieurs correspondants aspectuels construits avec des préverbes divers (*zółknąć - pożółknąć - zżółknąć* ‘jaunir’).

Tenant compte de ces difficultés, W. Cockiewicz et A. Matlak ont conçu le dictionnaire structural et aspectuel du polonais [2]. Le but de ce travail est de présenter le système dérivationnel du verbe en montrant à la fois ses relations aspectuelles. Pour ce faire, Cockiewicz adopte une théorie de l’aspect [1] dans laquelle la perfectivité comprend deux sous-catégories: l’aspect au sens stricte (les verbes qui indiquent la fin de l’action) et l’aspect au sens large (les déterminatifs engendrés régulièrement avec *po-*). Formellement, cette distinction se base sur le critère suivant: on considère qu’un imperfectif et un perfectif forment une paire aspectuelle lorsque parmi de nombreuses formations préfixées du verbe de base, on trouve celle qui ne peut plus former un correspondant imperfectif, comme dans: *czytać -> przeczytać -> \*przeczytywać*.

Dans ce cadre théorique, le réseau dérivationnel du verbe *pisać* ‘écrire’ (le fragment ci-contre) se présente comme suit (la flèche montre la direction de dérivation et les deux-points indiquent les relations aspectuelles):

<b>pisać</b> ->: napisać			
->: popisać			
-> dopisać	->: dopisywać	->: podopisywać	
-> odpisać	->: odpisywać	->: poodpisywać	
-> opisać	->: opisywać	->: poopisywać	
....			
etc.			

La première position du réseau occupe (dans 75 % de cas) un verbe imperfectif simple. Dans la deuxième colonne, nous trouvons les formations préfixées. Remarquons que, selon l’acception

de l'aspect adoptée par les auteurs, seuls *napisać* et *popisać* sont considérés comme aspectuels. La troisième position est réservée aux verbes imperfectifs engendrés par les suffixes. Les formations perfectives déterminatives créées par le préverbe *po-* occupent la dernière place. Les auteurs proposent la traduction anglaise pour chaque verbe préfixé.

Il est clair qu'en plus de l'effort de systématisation du système verbal, ce dictionnaire est un formidable outil pour l'apprentissage du polonais. Cependant, on peut se poser la question du choix des critères pour la formation d'une paire aspectuelle dans le cas de la préverbation. Ces critères, comme on le sait, ont été sujets de maints débats qui portaient sur le sémantisme du préverbe. Sans rentrer dans les détails de ces discussions, il nous semble que la description du préverbe devrait se faire dans un cadre théorique plus large que celui d'une paire aspectuelle. Ainsi, l'approche que nous proposons ne consiste pas à chercher les critères d'opposition entre les formes imperfectives et perfectives, mais vise, bien au contraire, à nous interroger sur le changement sémantique introduit par le préverbe, sur les origines prépositionnelles de ce changement et sur sa contribution à l'aspectualité.

Les études que nous développons utilisent les opérateurs topologiques (voir plus loin) et font référence à la Grammaire Applicative et Cognitive [6] qui est une extension de la Grammaire Universelle de Shaumjan [11]. Cette grammaire s'articule sur trois niveaux de représentation interliés par des formalismes applicatives.

## 2 Les opérateurs topologiques, les espaces abstraits et l'aspect

La topologie est un formalisme largement employé en linguistique. On se sert des opérateurs topologiques de l'intériorité (INT), de l'extériorité (EXT), de la frontière (FRO) et de la fermeture (FER) pour décrire les marqueurs d'espace, de temps et d'aspect. Cependant, la topologie classique qui divise l'espace en région de l'extérieur et de l'intérieur les séparant par une frontière ne répond pas toujours aux besoins d'un linguiste. On voit donc naître des calculs qui attribuent des épaisseurs à la frontière. Nous retrouvons cette idée dans la *locologie* de M. De Glas [3] et dans la *théorie des lieux abstraits* développée par J.-P. Desclés et son équipe [4]. Ce dernier formalisme, qui nous servira pour les travaux présentés ici, introduit les notions de frontière extérieure (FRO\_EXT) et de frontière intérieure (FRO\_INT).

Nous utilisons les opérateurs topologiques pour décrire la signification de la préposition spatiale et du préverbe correspondant. Ainsi, nous considérons la préposition spatiale comme un opérateur topologique qui détermine un lieu. Cependant, puisque la préposition peut marquer un lieu spatial, mais aussi la temporalité, l'activité et ce que Bernard Pottier [10] appelle la notion, plutôt que de privilégier la catégorie spatiale, nous allons parler d'un *lieu abstrait*.

Historiquement, le préverbe s'est développé à partir de l'adverbe et de la préposition. Ce procès diachronique implique certaines modifications syntaxiques comme la transitivisation et le changement d'arité du verbe. Au niveau sémantique, le lien entre la préposition et le préverbe se remarque surtout dans le verbe de mouvement, mais on l'observe également dans des constructions temporelles et les opérations indiquant les changements sur l'objet:

- (a) *Anna biegła do domu* -> *Anna dobiegła do domu*  
'Anna courait à la maison' -> 'Anna est arrivée à la maison (en courant)'
- (b) *Jan czekał do rana* -> *Jan doczekał rana*  
'Jan attendait jusqu'au matin' -> 'Jan a attendu jusqu'au matin'
- (c) *Maria czytała książkę (do końca)* -> *Maria doczytała książkę*  
'Maria lisait un livre (jusqu'à la fin)' -> 'Maria a terminé le livre'

Le préverbe possède donc une signification venant de la préposition. Cet opérateur souvent modifie la signification du verbe, mais il marque aussi l'aspect perfectif. La question qui se pose concerne le choix du métalangage capable de décrire ces deux opérations.

Nous pouvons représenter l'aspect en interprétant les points de l'espace topologique comme les intervalles d'instant avec des bornes (frontières) ouvertes ou fermées. Après Desclés [5], nous allons distinguer trois aspects : *état*, *événement* et *processus*. Nous utilisons les mêmes concepts pour définir l'aspect lié aux marqueurs grammaticaux et l'aspectualité liée aux prédicats.



L'aspect *état* représente une situation stable, où ni le début ni la fin ne sont pris en compte. En termes topologiques, cet aspect se réalise sur un intervalle ouvert (les bornes n'appartiennent pas à cet intervalle):



Figure 1: Aspect ETAT

L'aspect *événement* représente une situation prise dans sa globalité. Il se réalise sur un intervalle fermé.



Figure 2: Aspect EVENEMENT

L'aspect *processus* représente un changement initial. Le processus s'oriente vers la fin sans pourtant l'atteindre. Il se réalise sur un intervalle fermé à gauche et ouvert à droite.



Figure 3: Aspect PROCESSUS

Le processus qui a atteint la borne droite engendre un événement. Ce processus peut être *accompli* ou *achevé*. Un processus est *achevé* lorsqu'il a atteint la borne au-delà de laquelle il ne peut plus se poursuivre. Un processus est *accompli* lorsqu'il a atteint la borne qui n'est pas nécessairement finale.

L'approche que nous proposons ici suppose la continuité de situations aspectuelles. Il s'ensuit que les aspects de base, à savoir *état*, *événement* et *processus* sont interdépendants. Cette propriété distingue la théorie proposée de celle de Vendler [12] ou Mourelatos [9] où les concepts aspectuels sont organisés hiérarchiquement.

Les trois aspects de base (**ASP**) s'appliquent à toute la relation prédicative et donnent comme résultat le schéma prédicatif suivant: **ASP** (Prédicat, Terme1, Terme2...).

Le choix aspectuel s'effectue dans une situation de l'énonciation qui implique, entre autres, un énonciateur et un co-énonciateur. L'énonciateur insère le schéma prédicatif dans son système référentiel et l'organise par rapport à son acte d'énonciation. En termes topologiques, nous qualifions cet acte comme de *processus* ce qui nous permet de saisir l'évolution de la production de la parole: "JE, énonciateur, je suis en train de parler". La relation entre le *processus énonciatif* et le schéma prédicatif (coïncidence, antériorité, postériorité) indique les temps grammaticaux.

Nous pouvons représenter le schéma prédicatif, le processus d'énonciation et les relations temporelles dans une expression applicative suivante:

$$\text{PROC}_I (\text{DIT} (\text{ASP}_J (\text{Prédicat, Terme1, Terme2...}) I) \& (\text{I REL J}))$$

Dans ce modèle aspecto-temporel, il serait possible de considérer le préverbe polonais comme un opérateur qui s'applique à un *processus* pour en créer un *événement achevé*. Cependant, une telle description « écraserait » sa dimension sémantique supprimant la distinction entre *napisać list*, *dopisać list*, *popisać list*, *przepisać list*, etc. En effet, pour expliquer certains phénomènes, nous aurons besoin de considérer les zones qualitatives liées au lexique. Nous faisons la distinction entre les prédicats *processuels*, *événementiels* et *statiques*. Nous associons aux prédicats dynamiques (*processuels* et *événementiels*) les sept zones lexico-aspectuelles engendrées en projetant les lieux de la théorie des lieux abstraits (avec des frontières épaisses) sur l'axe temporel :

extériorité: *avant*  
 frontière extérieure: *zone de préparation*  
 frontière intérieure: *zone de commencement*  
 intériorité: *pendant l'action, la continuité*  
 frontière intérieure: *la résultativité*  
 frontière extérieure: *la fin*  
 extériorité: *après*

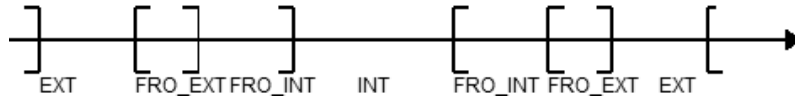


Figure 4: Zones aspecto-temporelles

### 3 L'achèvement spatial

Comment, avec les outils présentés, modéliser les opérations du préverbe ? Prenons quelques exemples de verbes de mouvement.

- (1) *Agata biegła<sup>IMPF</sup> przez park*  
courir à travers parc.ACC  
'Agata a couru à travers le parc'
- (2) *Agata prze-biegła<sup>PERF</sup> przez park*  
à travers-courir à travers parc.ACC  
'Agata a traversé le parc'
- (3) *Agata biegła<sup>IMPF</sup> do parku*  
courir jusqu'à parc.ACC  
'Agata a couru jusqu'au parc'
- (4) *Agata dobiegła<sup>PERF</sup> do parku*  
jusqu'à-courir jusqu'à parc.ACC  
'Agata est arrivée dans le parc'

Dans (1) la préposition *przez* 'à travers' détermine le lieu spatial 'parc' dans sa frontière, son intérieur et la deuxième occurrence de la frontière. Du point de vue aspectuel, l'agent vise l'occurrence de la deuxième frontière, mais il ne l'atteint pas.

Dans l'exemple (3), la préposition *do* 'à', 'jusqu'à' indique la frontière extérieure du lieu 'parc', mais tout comme dans l'exemple précédent, cette frontière n'est pas atteinte. Observons maintenant l'opération engendrée par le préverbe. Dans (2) et (4), l'agent termine son mouvement et il se trouve dans un lieu spatial indiqué par le préverbe. En effet, *prze-* et *do-* créent un événement, mais ils renvoient à des lieux différents selon leurs significations respectives. Examinons de près l'exemple (2): l'agent a terminé le mouvement au moment où le lieu 'parc' a été parcouru. En termes topologiques, nous dirons que ce mouvement a été achevé lorsqu'on a atteint la deuxième occurrence de la frontière du lieu 'parc'. Nous illustrons ces relations dans le diagramme à 2-dimensions, où l'abscisse indique l'aspectualité du prédicat et l'ordonnée, les lieux. La flèche marque l'achèvement spatial qui dépend de la valeur des deux axes.

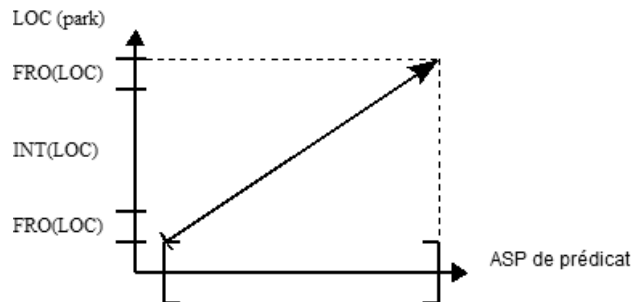


Figure 5: *Przebiegła przez park*

En comparaison, l'exemple (4) est plus complexe, car le préverbe *do* indique ici la frontière d'un lieu mais il marque aussi la phase finale du mouvement. Ainsi, l'achèvement dans ce cas n'est pas global, mais porte seulement sur la phase finale.

En général, l'achèvement spatial consiste en la fermeture de la borne droite d'un processus ce qui coïncide avec l'atteinte d'une zone topologique d'un lieu spatial. Cet achèvement n'indique pas nécessairement la fin de l'action, mais peut renvoyer à l'achèvement du début, l'achèvement de la fin, etc.

## 4 L'affectation de l'objet

Les travaux récents soulignent le lien entre l'aspect et la détermination de l'objet. Dans certaines langues comme le finnois ou l'estonien, on observe une relation entre l'aspect et le cas partitif [8]. En français, le choix du déterminant en co-occurrence avec le passé composé peut influencer l'aspect de toute la relation prédicative en indiquant soit l'achèvement soit l'accomplissement<sup>1</sup>.

Il semble qu'en polonais le préverbe marque une modification de l'objet et que le résultat de ce changement soit aspectuellement pertinent. Examinons quelques cas :

- (5) *Agata zbudowała<sup>PERF</sup> dom*  
 prev-construire maison.ACC  
 'Agata a construit la maison'
- (6) *Jan napisał list*  
 prev-écrire lettre.ACC  
 'Jan a écrit la lettre'
- (7) *Anna wypila<sup>PERF</sup> herbatę*  
 prev-boire thé.ACC  
 'Anna a terminé son thé'

Dans les deux premiers exemples, l'action progresse simultanément avec la construction de l'objet pour s'achever au moment où cet objet ('maison', 'lettre') acquiert une existence. Dans l'exemple (7), au contraire, l'objet subit une "déconstruction" progressive. Ainsi, ces trois cas présentent l'achèvement lié à une opération sur l'objet. Le préverbe est ici un opérateur qui engendre la fermeture d'un *processus* sous-jacent et, en même temps, il affecte l'objet.

Les exemples traditionnellement considérés comme des modalités d'action, s'expliquent dans le cadre du même modèle:

- (8) *Anna dopiła<sup>PERF</sup> herbatę*  
 jusqu'à-boire thé.ACC  
 'Anna a terminé son thé'
- (9) *Anna odbudowała<sup>PERF</sup> dom*  
 re-construire maison.ACC  
 'Agata a reconstruit la maison'

Dans (8), il faut faire appel à la signification du préverbe *do* et examiner ensuite sa compositionnalité avec le verbe *pić*. En effet, *do* marque la modification de l'objet dans une phase finale de sa "disparition". Dans l'exemple suivant, nous analysons *od-*. La notion de *odbudować* suppose bien évidemment une construction (*budować*), mais *od-* fait implicitement appel à une déconstruction qui peut s'exprimer par des verbes polonais *zburzyć*, *zniszczyć* 'détruire'. Nous avons donc deux situations saillantes, l'une où la maison est détruite et l'autre où elle est reconstruite. C'est au moment de la reconstruction que le processus est achevé. Avec les mêmes outils, nous pouvons décrire les formations délimitatives créée avec le préverbe *po-*, telles que *poczytać książkę*, *popisać list*. Nous allons les interpréter comme des achèvements du début.

Pour conclure cette partie, nous dirons que le préverbe dans les exemples (5) - (9) engendre la fermeture de la borne droite d'un processus ce qui coïncide avec une modification de l'objet (la construction, la déconstruction, les changements d'état). En fonction de la signification du préverbe (dans les cas où le préverbe n'est pas totalement grammaticalisé) cet achèvement peut aussi signifier l'achèvement du début ou de la fin, etc.

<sup>1</sup> Comparons: (a) *Il a bu un verre de vin* et (b) *Il a bu du vin*. Dans (a) le processus est achevé, alors que dans (b) l'article partitif nous indique que l'action de boire pourrait se poursuivre. Le processus est donc accompli.

## 5 Transition d'un état à un autre

Nous avons jusqu'à maintenant analysé les verbes qui marquent le changement sur le lieu spatial ou sur l'objet. Prenons des exemples où les modifications concernent un élément  $T^1$  (sujet) de la relation prédicative.

- (10) *Jan wytyślał*<sup>PERF</sup>  
 pref-devenir chauve  
 'Jan est devenu chauve'

Traditionnellement, (10) est analysé comme une constructin résultative. En effet, dire *Jan wytyślał* implique l'état résultant '*jest łysy*', mais nous pouvons inférer également un état antérieur '*nie jest łysy*'. La transition d'un état à un autre est exprimée par un processus souvent lexicalisé (*łysieć, chudnąć...*). L'application du préverbe ferme la borne de ce processus et marque l'état résultatif qui est concomitant. Il semble que le polonais se focalise non pas sur le résultat (celui est seulement impliqué), mais sur le moment-même de l'achèvement. L'exemple en bas, tiré du corpus IPI PAN montre une succession d'événements:

*Malowała powoli, portret robiła wprost latami. Model się zestarzał, wytyślał, ożenił, schudł, musiał pozować, chciał czy nie chciał, chyba że umarł.*

Nous pouvons, de la même manière, analyser les formations inchoatives:

- (6) *Darek pokochał Marię*  
 prev-aimer Maria.ACC  
 'Darek est tombé amoureux de Maria'

En effet, *pokochać* 'tomber amoureux', *zachorować* 'tomber malade', etc. impliquent un état initial contraire *nie kochać* 'ne pas aimer', *nie być chorym* 'ne pas être malade'... et un état final. Le processus (non lexicalisé) qui mène d'un état à un autre est achevé dans son début. Remarquons que pour expliquer la résultativité et l'inchoativité nous procédons à une décomposition d'un processus sous-jacent en zones qualitatives. Alors que l'inchoativité se focalise sur la frontière extérieure associée à la phase initiale du processus, la résultativité fait appel à la frontière en relation avec sa fin.

## 6 L'opération de l'achèvement

Dans les exemples analysés, le préverbe engendre un achèvement. Il est clair que cette opération met en jeu plusieurs opérateurs aspectuels qui se situent sur des niveaux différents. Pour l'illustrer, donnons quelques éléments de l'analyse formelle du préverbe dans le cadre d'une relation prédicative binaire  $P_2T^2T^1$ . Introduisons un opérateur aspectuel **ASP3** qui représente les propriétés aspectuelles inherant au prédicat (*processuel, événementiel, statique*):

1.  $((\mathbf{ASP3}P_2)T^2)T^1$

Notre analyse a montré qu'une relation entre le prédicat aspectualisé et le terme  $T^2$  introduit de nouvelles caractéristiques aspectuelles. Nous désignons cette relation par **ASP2**:

2.  $((\mathbf{ASP2}(\mathbf{ASP3}P_2)T^2)T^1)$

Introduisons enfin l'opérateur de l'aspect grammatical **ASP1** (état, événement, processus) qui porte sur toute l'expression applicative:

3.  $[\mathbf{ASP1} ((\mathbf{ASP2}(\mathbf{ASP3}P_2)T^2)T^1)]$

L'expression (3) est l'opérande de processus d'énonciation qui comprend l'énonciateur  $S^\circ$ :

4.  $\text{PROC}\{\text{DIT}[\mathbf{ASP1} ((\mathbf{ASP2}(\mathbf{ASP3}P_2)T^2)T^1)]S^\circ\}$

Cette expression signifie que l'énonciateur est en train de dire que la relation prédicative est aspectualisée (**ASP1**), que le prédicat possède les propriétés aspectuelles relatives à sa signification (**ASP3**) et que la relation entre ce prédicat et le terme  $T^2$  est aspectuellement pertinente (**ASP2**).

Il semble que le préverbe polonais, dans la relation binaire, est la trace de la composition formelle entre les opérateurs **ASP2** et **ASP3**. Il s'ensuit que l'aspectualité dans ce cadre porte toujours sur un opérande. Quant à l'aspect **ASP1**, cet opérateur représente le choix de l'énonciateur entre l'événement achevé (l'antériorité par rapport au processus d'énonciation) et l'événement qui

se réalisera après le processus d'énonciation. Pour le détail de ce calcul effectué dans le cadre applicatif (logique combinatoire) voir [7].

## 7 Conclusion

Nous avons présenté une analyse du préverbe polonais autant qu'un marqueur d'un achèvement.

Cette opération complexe se présente comme une fermeture du *processus* inhérent au prédicat lexical (une complétude temporelle de l'action) qui coïncide, dépendamment de l'argument, avec (i) l'atteinte d'une zone topologique d'un lieu spatial ; (ii) la modification de l'objet; (iii) le changement de l'état d'une entité.

Grâce aux outils topologiques, nous avons pu établir les zones qualitatives de l'achèvement, dans les cas, où la signification du préverbe introduit les changements aspectuels. Ainsi, l'achèvement peut être défini plus largement que "la fin de l'action".

Il est clair qu'au stade actuel cette étude ne pas complète. Par exemple, les préverbes qui indiquent de la quantité et de la mesure (*na-*, *po-*, *wy-*) nécessitent une analyse à part. D'un autre côté, il semble la compositionnalité entre le préverbe et *się* 'se' (*przespacerować się* 'se faire une promenade', *nabiegać się* 'courir beaucoup'...), ne renvoient pas à l'achèvement, marquant plutôt un accomplissement qui se focalise sur un état (satisfaction, satiété, plaisir). L'autre problème constitue le classement aspectuel des verbes.

## Références

- [1] Cockiewicz, W. (1992). *Aspekt na tle systemu słowotwórczego polskiego czasownika i jego funkcyjne odpowiedniki w języku niemieckim*. Uniwersytet Jagielloński, Kraków.
- [2] Cockiewicz, W., Matlak, A. (1995). *Strukturalny słownik aspektowy czasowników polskich*. Uniwersytet Jagielloński, Kraków.
- [3] De Glas, W. (1991). Locological spaces: knowledge representation in an intensional setting. In *Proceedings of the third COGNITIVA symposium*, pages 229–337, Amsterdam. North-Holland Publishing Co.
- [4] Desclés, J.-P. (septembre 2006). Opérations métalinguistiques et traces linguistiques. *Colloque en Hommage ? Antoine Culioli, Centre International de Cerisy*.
- [5] Desclés, J.-P. (1980). Construction formelle de la catégorie grammaticale de l'aspect. In *La notion d'aspect*, David, J., Martin, R. (éds), pages 198–237, Paris. Klincksieck.
- [6] Desclés, J.-P. (1990). *Langages applicatifs, langues naturelles et cognition*. Hermès, Paris.
- [7] Gwiazdecka, E. (2005). *Aspects, prépositions et préverbes dans une perspective logique et cognitive. Application au polonais: przez/prze-, do/do-, od/od-*. Thèse de doctorat. Université de Paris IV, Sorbonne.
- [8] Kiparsky, P. (1998). Partitive case and aspect. In *The projection of Arguments: Lexical and Compositional Factors*, Butt, M, Geuder, W. (eds), pages 275–269, Stanford. CSLI Publications.
- [9] Mourelatos, A. (1978). Events, states and processes. *Linguistics in Philosophy*, 2:415–34.
- [10] Pottier, B. (1995). *Sémantique générale*. Presses Universitaires de France, Paris.
- [11] Shaumyan, S. K. (1977). *Applicative Grammar as Semantic theory of Natural Language*. Chicago University Press, Chicago.
- [12] Vendler, Z. (1967). Verbs and times. In *Linguistics in philosophy*, pages 97–121, Ithaca. Cornell University Press.

# On the Lexicographic Representation of Relational Nouns

Petya Osenova

Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria  
petya@bultreebank.org

**Abstract.** This paper discusses the problem of representing the relational nouns in a lexicon. In order to comment on this representation, various approaches towards their classification are presented. The author relies on Pustejovsky's idea about the generative lexicon for building a hierarchy of relational nouns. Qualia structure and argument structure are seen in their interdependence. The relational nouns are viewed as continuum rather than a dichotomy. Also, an experiment with nouns from an explanatory dictionary of Bulgarian is presented and analysed.

**Keywords:** relational nouns, argument structure, qualia structure, continuum, relations

## 1 Introduction

The relational nouns can be defined as nouns, whose referents always presuppose some semantic relation to another referent. However, it is difficult to delimit the boundaries of these nouns. We support the idea that the nouns do not exist in dichotomy (*relational vs. non-relational*), but they rather exist in continuum, depending on their sense.

The classic cases of relational nouns are kinship (*mother, father, sister*) and part-whole (*leg, hand, head*) nouns. The problem remains, which other groups are to be included here. For example, Hölzner [3] thinks that the argument-taking nouns in German fall into the following groups: nouns for events (*siege, visit*), nouns for thematic roles (*discovery, examiner*) and relational nouns (*king, father*). On the other hand, Pitha [4, p.220] says that it is very difficult to determine all the relational nouns. He raises the question whether words like *doctor* and *patient* are as relational as *brother* and *friend*. Gentner and Kurtz [2] propose the idea that the relational nouns participate in the so-called *relational scheme*. For example, the scheme *theft* comprises also *thief, victim, and stolen property*.

From the brief review above, it became clear that there is no common opinion on the scope of the relational nouns. Thus, the following questions need their answers: a) what are the criteria for defining a noun as being relational?, b) what types of relations are denoted by these nouns?, c) are the relational nouns contrasted to non-relational, or is there a gradual approach?

In our opinion, these nouns have to be classified not only according to some semantic criterion, but also with respect to the following relations: symmetricity, transitivity, reflexivity, etc.

## 2 Two hypotheses

To solve the above-mentioned problems we rely on two hypotheses: Barker and Dowty [1] and Pustejovsky [6].

Barker and Dowty [1] rely on entirely semantic approach towards delimiting the relational nouns. They consider the relational nouns as nouns having one additional argument. For example, the notation within the set theory would be as follows:

- a. [human] = { x | x is human }
- b. [friend] = { x |  $\exists$  y such that x is friend of y }

Barker and Dowty [1] choose the part-whole relation as a central one to the relational nouns in contrast to the possessivity relation. The authors include the following groups: kinship relations, body parts, other partitive nouns, nouns for abstract characteristics, deadjectival nouns. At the same time we accept Pustejovsky's idea that each noun is, in general, relational. I.e. it has at

least a zero argument, which is in practice its ontological restriction - superconcept (*knife* is an *instrument*; *leg* is a *limb*, *man* is a *human*, etc.). Pustejovsky introduced the so-called qualia structure, which includes the following roles: AGENTIVE (origin), CONSTITUTIVE (content), TELIC (purpose), FORMAL (differentia specifica). However, even the qualia structure is not enough for differentiating among the nouns. Thus, we use the connection between the qualia structure and the argument structure. Here are examples, which show this connection:

*brother*  
 ARGST = [ARG0 = x:human]  
           [D-ARG = y:human]  
 Qualia = [CONST = man (x)]  
           [**FORMAL = brother\_of(x,y)**]

The zero argument (ARG0) indicates the superconcept ‘human’. The default argument (D-ARG) indicates the connection to another human being. Within the qualia structure FORMAL characteristics encodes the relation *brother\_of*.

*arm*  
 ARGST = [ARG0 = x:limb]  
           [D-ARG = y:human]  
 Qualia = [FORMAL = x]  
           [**CONST = part\_of(x,y:body)**]

The zero argument ( ARG0) indicates the superconcept ‘limb’. The default argument ( D-ARG) indicates the *part-whole* connection to the human being. Within the qualia structure CONST characteristics encodes the relation *part\_of* .

*honda*  
 ARGST = [ARG0 = x:car]  
 Qualia = [FORMAL = x]  
           [**TELIC = drive(e,y,x)**]  
           [**AGENTIVE = create(e,Honda,x)**]

In contrast to the previous two examples, this one lacks a default argument. It indicates only the superconcept ‘car’. Within the qualia structure TELIC and AGENTIVE relations have values, which relate to events.

### 3 Towards a solution

Following Pustejovsky’s ideas, where all nouns have at least one zero argument ( **ARG 0**) and often express at least one relation within the qualia structure, we need a more restrictive criterion. Thus, we will additionally rely on the information within the argument structure. More precisely, when there is an argument, which is different from the zero one, a test can be made for detecting the degree of ‘relational-nounness’. Let us sum up what the various types are.

1. The kinship relational nouns are expressed via the FORMAL qualia.

In the argument structure, the zero argument shows that the referent is human, but there is also a default argument (D-ARG), i.e. the other entity to which the referent relates. This other entity is also human.

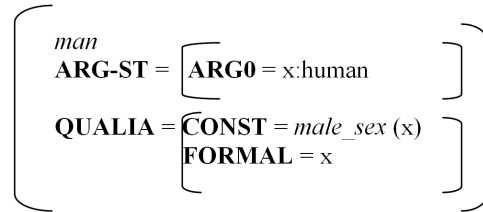
2. The relational partitive nouns are expressed by the qualia CONSTITUTIVE.

Within the argument structure the zero argument says that the *body\_part* is limb or another respective superconcept, but there is again a default argument, which shows that the *body\_part* is part of human. In the qualia structure the CONST(itutive) qualia encodes the relation *part\_of*.

3. The relational nouns for purpose and origin are expressed by the qualia TELIC and AGENTIVE. There is not a default argument, because these relations reflect situations rather than entities.

Therefore, we can conclude that the qualia FORMAL and CONSTITUTIVE are more central than TELIC and AGENTIVE as tests for detecting relational nouns.

From the above considerations, the following conclusion can be made: the relational noun requires for each own referent the existence of another referent or situation, which it depends on. Thus, when all the specified relations of a noun have value X instead of relation to another object, it is considered as non-relational. For example, the word *man* is non-relational in the meaning of ‘human with a certain sex’:



Our idea is to make a hierarchy of qualia, based on the interdependence among referents. For prototypical ones we can consider the *kinship relations* and the *social roles*, which comprise two referents. They are expressed by the characteristics FORMAL. Next comes the relation *part\_of*, which has also two referents. They are expressed by the qualia CONSTITUTIVE. Then, provisionally, we order the relations TELIC and AGENTIVE. They usually include situations. In the farthest end, the abstract and deadjectival nouns are positioned. They have one referent and relation to one of his characteristics. We call this newly added qualia CHARACTERISTICS.

Thus, in our opinion the qualia structure characteristics exist in a ranked order with respect to relational nouns as follows:

FORMAL > CONSTITUTIVE > TELIC > AGENTIVE > CHARACTERISTICS

From left to right the relational nature of the relational nouns decreases, but still is present within their lexical structure.

#### 4 Dictionary-based experiment

In order to gather some realistic language data, we decided to use the information from a lexicographic source. All nouns were excerpted from the Bulgarian explanatory dictionary [5], which had in their definition the expression ‘related to’. The number of these nouns was 64. These nouns were also classified with respect to relations, such as *inversed*, *symmetrical*, etc. As expected, it turned out that the expression ‘related to’ was not a stable criterion due to the variety of ways in the definition presentation. For that reason, the list of relata is not exhaustive. For example, the word *patient* was excerpted, but not the word *doctor*. Two thirds of the nouns refer to persons. The same was the situation with the partitive nouns. We used the expression ‘part\_of’. However, it was present in the definition for the word *head*, for example, but not for the word *arm*. Our task in this paper is not to discuss the non-homogeneity of definitions in a dictionary, but this problem is worth to be mentioned. Below comes our classification of the nouns.

1. Inversed relations If there is a relation R (x,y), where x is R to y, then there exists also a relation R1 (y,x), where y is R1 to x. Examples: grandmother related to granddaughter; granddaughter related to grandmother.  
 Semantic types:
  - kinship (father, son-in-law, wife, sister, etc.)
  - social roles (alumni, guest, patient, student, etc.)
  - partitives (head to body, province to country, etc.)
2. Symmetrical relations If there is a relation R (x,y), where x is R to y, then there is also the same relation R (y,x), where y is R to x. Examples: I am a companion (fellow-traveller) to you, and you are companion (fellow-traveller) to me.  
 Semantic types:
  - kinship relations (brother to his own brother; sister to her own sister)



- other relations (co-religionist, fellow-student, fellow countryman, etc.)
- 3. Other relations These relations are called ‘other’ just because they bear a qualia with lower degree of ‘relation nounness’. These are AGENTIVE, TELIC and CHARACTERISTICS. Please note that the first and the second, presented below, are of type inverted.
  - If there is a relation R (x,y), where x is a product of y, then there is also relation R1, where y is a manufacturer of x and produces x. For example, for the word *opium* the following definition is provided: it is a narcotic substance, which *is derived* from a solidified poppy juice.
  - If there is a relation R (x,y), where x is used for y, then there is also relation R1, where in y x is used. In the dictionary, one of the possible detecting expressions was ‘serve for’ (fan, test-tube, etc.)
  - *is* \_ X, where X is a characteristics (redness, height, symmetry, reflex).

## 5 Conclusion

The semantic criterion does not facilitate the detection of relational nouns, since all the nouns are relational from a certain point of view. Thus, we should rely on the combination of the ontological and lexico-compositional approaches. These nouns form a continuum, in which there is a different type of domination qualia. In our opinion, it is misleading to look for which relation is part of the lexical meaning of the word. Rather, we should consider the dominating one in the specific context.

The relational nouns should be explicated in lexicons with respect to the connection between argument structure and qualia structure in Pustejovsky’s sense. The presence of an additional argument is pointed out also by Barker and Dowty [1]. However, we defend the relativity of the set of relational nouns. There exist stronger as well as weaker relational nouns. The noun meaning also plays a crucial role for the lexical entry specification.

## Références

- [1] Barker, Ch., Dowty, D. (1993). Nominal thematic proto-roles. In <ftp://ftp.ling.ohio-state.edu/pub/dowty/nvthr-pt1.ps.gz>.
- [2] Gentner, D., Kurtz, K. (2005). Learning and using relational categories. In *Ahn, W. K., Goldstone, R. L., Love, B. C., Markman, A. B., Wolff, P. W. (Eds.), Categorization inside and outside the laboratory.*, Washington, DC. APA.
- [3] Holzner, M. The syntax-semantic interface of german nouns. In <http://linguistics.buffalo.edu/research/rrg/Matthias%20Holzner%20-%20The%20syntax-semantic%20interface%20of%20German%20nouns.pdf> .
- [4] Pitha, P. (1981). On case frames of nouns. *Prague studies in Mathematical Linguistics*, 7:215-224.
- [5] Попов., Д. (1999). *Български тълковен речник*. Наука и изкуство, София.
- [6] Pustejovsky, J. (1998). *The Generative Lexicon*. The MIT Press, Cambridge, Massachusetts, London, England.



Part 4  
**Abstracts**

# The Lexicographic Description of Modals in Polish

Björn Hansen

Institut für Slavistik, Universität Regensburg

The aim of the contribution is to present a practical application of a lexicographic model which combines features of the known Russian dictionary ‘Tolkovo-kombinatornyj slovar’ with elements of the Polish ‘Składnia gramatyki współczesnego języka polskiego’. This lexicographic model could be the basis for a new valence dictionary of Polish. The idea is to combine semantic with syntactic information in a maximally compact and precise way. The model is presented on the basis of the category of modals, i.e. a class of elements which due to their abstract semantics are located between lexicon and grammar. Modals form a cross-linguistic category found in all European languages and can be grasped in the following way: A fully-fledged modal is a polyfunctional, syntactically autonomous expression of modality which shows a certain degree of grammaticalisation. ‘Polyfunctional’ is understood as covering a domain within the semantic space of modality. A fully-fledged modal functions as an operator on the predicational and/or the propositional level of the clause (Hansen /de Haan in press). According to this definition, modals are mainly characterised by their polyfunctionality in the semantic field of modality (dynamic, deontic and epistemic).

A specific feature of the model is that it combines the advantages of an explaining with a valence dictionary. Every dictionary article consists of the following sections:

1	<b>LEMMA</b>	
2	meaning 1 ‘meaning label’	
	meaning 2 ‘meaning label’	
	...	
3	Formal syntactic features	
	1. government (meaning 1)	semantic selectional restrictions
	2. government (meaning 2)	
4a	meaning 1 ‘EXPLICATION’	
4b	illustrating example	
5a	meaning 2 ‘EXPLICATION’	
5b	pr illustrating example	
6	Pragmatic features	

The explication is written in a normalized version of Polish (Apresjan: ‘pod’-język języka ob’ekta) and contains the semantic elements relevant for the individual modals: the modal primitives CAN (possibility) or MUST (necessity) or IT IS BETTER IF (weakened necessity as in *mieć*), the type of modality (dynamic, deontic and epistemic) and variables which are filled by the syntactic context. We distinguish propositional variables P, Q filled with verbal phrases or subordinate clauses from individual variables. A major advantage of the model is that the lemma contains a direct link between the meaning explications and the formal syntactic features: the variables in the explication show up in the government model which indicates the number of valence slots and the selectional restrictions. We suggest a notation which allows to distinguish between cases where the valence slots are filled by a semantic actant from cases where the verb opens a syntactic valence slots which, however, is filled by a semantic actant inherited from the infinitival verb (raising construction).

## Références

Apresjan, Ju.D. (1994) O języku tolkowanij i semantičeskich primitivach. *Izv. RAN, Serija Lit. i Jaz.* 4, 27-40

- Hansen, B. (2001)** *Das Modalauxiliar im Slavischen. Semantik und Grammatikalisierung im Russischen, Polnischen, Serbischen/Kroatischen und Altkirchenslavischen. (Slavolinguistica 2)*. München.
- Hansen, B. (2006)** Mapy semantyczne w konfrontacji językowej. In: Koseska-Toszewa, V. & Roszko, R. (red.): *Semantyka i konfrontacja językowa* t. 3. Warszawa, 141-151
- Hansen, B., de Haan, F. (in Print)** Concluding chapter: Modal constructions in the languages of Europe. In: Hansen, B. & de Haan, F. (eds.) *Modals in the Languages of Europe*. Berlin.
- Kątny, A. (1980)** Die Modalverben und Modaladverben im Deutschen und Polnischen. Rzeszów
- Ligara, B. (1997)** *Polskie czasowniki modalne i ich francuskie ekwiwalenty tłumaczeniowe*. Kraków.
- Mel'čuk, I. A. & Žolkovskij, A.K. (1984)** *Tolkovo-kombinatornyj slovar' russkogo jazyka*. Wien.
- Polański, K. (red.) (1980-92)** *Słownik syntaktyczno-generatywny czasowników polskich I-V*. Wrocław.
- Rytel, D. (1982)** *Leksykalne środki wyrażania modalności w języku czeskim i polskim*. Wrocław
- Topolińska, Z. (red.) (1984)** *Gramatyka współczesnego języka polskiego. Składnia*. Warszawa.
- Wierzbicka, A. (1987)** The semantics of modality. *Folia Linguistica* 21/1, 25-43.
- Zabrocki T. (1978)** Status syntaktyczny czasowników modalnych w języku polskim i angielskim. *Biuletyn Polskiego Towarzystwa Językoznawczego* 36, 43-57

# Situational and Information Structures of Discourse

Ewa Miczka

University of Silesia

The article regards relations between situational and information structures of discourse. Analyzing the possible configurations between these two types of structures, the author aims to present their role in discourse comprehension – the process which implicates creation of discourse representation.

The situational structures are defined as a sequence of frames (E. Goffman: 1991). Each frame permits to conceptualize one event forming a part of information introduced in discourse. The author proposes to apply the notion of cognitive event and their typology introduced by R. Langacker (1995) to describe the variations of situational structures of discourse.

The information structures are defined as hierarchically organized thematic-rhematic structures. The author distinguishes three levels in their thematic part represented by: global theme, theme of group of sentences and theme of sentence. The rhematic part is divided in two levels: the first one contains rhematic groups, the second rhemes of sentences.

The author focuses her attention on the highest level of information structure and describes the relations between the units of situational structures – frames and cognitive events – and choices regarding the global theme of discourse.

# Part of Speech Assignment as a Type of Semantic Information about a Word

Jadwiga Wajszczuk

University of Warsaw

The title of the paper has a slightly provocative character because when both the traditional label *part of speech* and the idea of semantic information are invoked in their mutual combination, what becomes readily associated with that kind of nomenclature is a threefold division of vocabulary according to its relevant aspects, a division well known since antiquity, where semantic characterization of a word has always involved a component corresponding to the concept of a “part of speech”; whereas it is nowadays clear that this conceptual bias should be abandoned for more reasons than one. It goes without saying that attempts to solve the primordial question of how to partition a lexicon from a functional-syntactic point of view (its division into “parts of speech” being the foremost and to a very great extent efficient materialization of the partition) by relying on such “ontological categories” as ‘substantiality’, ‘attribute’, ‘number’, ‘action’, ‘state’, ‘process’, etc., are damned to fail.

However, it has hitherto not been made clear what categories that would be apt at revealing the principles of such a division are in fact needed. No doubt morphological categories must be relegated to a subsidiary position; the reason is that they are plainly non-universal. Unfortunately, bold syntactic projects of partitioning a vocabulary (in studies of Polish these are, basically, Laskowski 1984, 1998, Wróbel 2001 with his draft correction) are flawed by circularity. This is because they are based on properties of words as sentence components, with ‘sentence’ adopted as an initial category, while we at the same time lack a syntactic definition of sentence that would not invoke its constituents (such as verb, nominal phrase, verbal phrase).

It is my contention that the direction of analysis should be reversed. The starting point in my own project is the word and its combinatory properties rooted in meaning: properties such as the word’s opening of certain positions with semantic validity vs. its opening of certain positions of another kind, its filling of certain positions with semantic validity vs. its failing to fill such positions, its unilaterally opening of certain positions vs. its bilaterally opening of certain positions (in a further perspective also right-hand vs. left-hand opening should be reckoned with – a distinction respecting the direction of the relation *vector*). These properties are taken to be **formal indices of certain meanings, to be described in syntactically relevant terms**, such as referentiality, predicativity, metapredicativity, on the one hand, and, to put it in general terms, metatextuality, on the other.

It can subsequently be shown that the successive nodes of the classification achieved in this way admit of a general semantic interpretation that would allow us to account for the function a given class is called to exercise while co-constituting a unit of a higher order (what is meant here are partly units of the next-higher level, and partly – not in a direct way! – of the sentence level).

# Facilitating Access to Digitalized Dictionaries

Janusz S. Bień

Formal Linguistics Department, University of Warsaw

One of the best formats for scanned documents is DjVu. An essential feature of the format is the hidden text layer, usually containing the results of Optical Character Recognition. Another important feature is the ability to store (and serve over Internet) the documents as a collection of individual pages.

From the very beginning it has been used also for dictionaries, as exemplified by „The Century Dictionary” (<http://global-language.com/CENTURY/>). There are also several Polish dictionaries available in this format. So the question is how to search efficiently the text layer in such large multi-volume works. For this purpose we intend to adapt Poliqarp (Polyinterpretation Indexing Query and Retrieval Procesor), a GPLed corpus query tool developed in the Institute of Computer Science of Polish Academy of Sciences. Some preliminary experiments are described in the talk.

In our „quick and dirty” approach we treat every page as a single document with the metadata consisting of the name of the document index and the name of the file with the page content. For every word, instead of grammatical tags, we provide its localization on the page in the form of the line number and its position in the line. All the data taken together allow to link the search results to the appropriate fragments of the original scans.



# The Confluence of the Dative and Middle Voice in Croatian and Polish

Mateusz-Milan Stanojević, Barbara Kryżan Stanojević

Faculty of Humanities and Social Sciences, University of Zagreb

In Croatian and Polish various constructions with the reflexive marker *se/się* may or may not involve a noun in the dative case. In Croatian one may say *govori se o ovome problemu* ‘this problem is discussed’ as well as *stalno im-DAT se govori o tom problemu* ‘they are being told about this problem all the time’. Other examples include, for instance, *Kto wie, co się zdarzy za dziewięć miesięcy* (Polish) ‘Who knows what will happen in nine months’ as opposed to *A jeżeli zdarzy im-DAT się coś złego?* ‘And what if something bad happens to them?’. In this paper we will discuss the way in which the *se/się* construction interacts with the dative case in the construction of meaning. A corpus study was conducted on the IPI PAN corpus of Polish (<http://korpus.pl/>) and the Croatian National Corpus (<http://www.hnk.ffzg.hr>) to find examples where the *se/się* construction coincided with the dative construction. The results show that there are two basic semantic groups: the allative/competitor group and the transfer group, which partially corresponds to semantic groups found for various dative senses (Stanojević and Tuđman in press). In the allative/competitor group the dative serves as an abstract goal, and the *se/się* construction marks the self-movement of the agent (i.e. the fact that it has internal energy). As opposed to that, in various transfer subsenses the *se/się* construction is grammaticalized to defocus the agent, and the dative gradually changes its role from a potentially affected recipient (as in *stalno im-DAT se govori o tom problemu* ‘they are being told about this problem all the time’) to a completely affected experiencer (*Meni-DAT kad se plače plačem* ‘When I feel like crying I cry’; *Wszystko można, tylko człowiekowi-DAT się nie chce* ‘Anything can be done, but a person simply doesn’t feel like it’). In these senses both the dative and the *se/się* construction are grammaticalized in respect to their other senses, and are hence semantically bleached. Therefore, in those senses new constructional meaning occurs, which is not present in any senses of the two components taken alone: dative as the experiencer of its internal change of state. Constructional meaning is possible only in the bleached senses, which are less detailed in respect to the “basic”, diachronically older senses.

## (Mini) Portraits of the Words *mistrz* and *uczeń*. Semantic Relations

Katarzyna Drożdż-Łuszczak, Zofia Zaron

University of Warsaw

The paper presents the results of research carried out within the project *Synchronic and Diachronic Research into Contemporary Proper Names* (financed by the Ministry of Science and Higher Education). The subject matter of the given paper is the relation between the lexemes *mistrz*<sup>2</sup> (master) and *uczeń*<sup>3</sup> (apprentice)<sup>1</sup>.

The phrases *mój mistrz* (my master), *mistrz Nowaka* (Nowak's master) and *uczeń Kowalskiego* (Kowalski's apprentice) suggest a kind of convertive relation. Yet it is not the case. The *mistrz-uczeń* relation is a unilateral one – it is the apprentice who takes a master and it is the apprentice who can say *jestem jego uczniem* (I am his apprentice). The master will confirm this only if *uczeń*<sup>3</sup> (apprentice) is taught by the master and if he is regarded by the master as a skilled, knowledgeable person.

However, it is quite unlikely that someone will seriously state about him- or herself: *Jestem mistrzem Iksińskiego* (I am Iksiński's master). It is only Iksiński, who can state *Jestem uczniem Kowalskiego* (I am Kowalski's apprentice). The occurrence of a two-way relation (*Iksiński jest uczniem Kowalskiego, a Kowalski jest mistrzem Iksińskiego* [Iksiński is Kowalski's apprentice and Kowalski is Iksiński's apprentice] ) can be indicated only by the speaking subject.

Summary:

- the relation *mistrz1-uczeń3* is a three-actant one: *ktoś, czyjś, w jakiejś dziedzinie wiedzy (/sztuki)* (someone, someone's, in a field of science(/art)).
- the occurrence of the relation *mistrz1-uczeń3* depends on the apprentice as he chooses his master (*Jeśli jest uczeń, jest i mistrz; żeby był mistrz, musi być uczeń* [If there is an apprentice, there must be a master; in order to be a master, one must have an apprentice]).
- *mistrz1* does not have to know that he or she is a master for someone.
- *mistrz1*, as perceived by the apprentice is a good (the best) guide / teacher in the field of science or art which is important for both of them.
- *uczeń3* can acquire the knowledge directly from the master, but he or she may also continue his / her master's work or develop his / her idea without being taught by the master him- or herself.

---

<sup>1</sup> Numbering of *Uniwersalny Słownik Języka Polskiego* (the Universal Dictionary of the Polish Language)

# Idiom Variability in Croatian: the Case of the CONTAINER Schema

Jelena Parizoska

Faculty of Humanities and Social Sciences, University of Zagreb

In cognitive linguistics most idioms (multi-word units which have figurative meanings and relatively stable forms) are considered to be motivated by various cognitive mechanisms which link the meaning of idioms with the meanings of their constituents (cf. Lakoff 1987, Gibbs 1994). One of these mechanisms is the CONTAINER image schema, which is reflected in Croatian expressions with the preposition *u* ('in'). This schema serves to structure abstract conceptual domains like SITUATIONS, EVENTS and STATES. For example, acting in a difficult situation is conceptualized as being in a container (e.g. *biti u sosu* (lit. be in a sauce.LOC)) or entering a container (e.g. *upasti u dugove* (lit. fall into debts.ACC)). In addition to motivating the idioms with the constituent *u*, the CONTAINER image schema also constrains their variability.

The aim of this paper is to demonstrate the systematic motivation of the Croatian idioms describing difficult situations by the CONTAINER schema. We performed a study of the Croatian National Corpus, and ended up with 681 instances of variant realizations of the given idioms. We analyzed the type of constructions and the verbs used.

Our results show that there are 51% of dynamic (accusative) construals and 49% of static (locative) construals. Dynamic construals predominantly relate to self-propelled motion (e.g. *ući u žrvanj* (lit. go into a millstone.ACC)) and the remaining examples are realized as energy transfer from one entity to another (e.g. *uvući koga u klopku* (lit. drag someone into a trap.ACC)). The results show that the construction *u* + NP constitutes the conceptual core of the given idioms, which serves as the basis for variant realizations that reflect the different ways in which the relation between the trajector and the landmark is conceptualized.

# RAMKI or How Verbs Were Framed

Magdalena Derwojedowa, Jadwiga Linde-Usiekniewicz, Magdalena Zawisławska

University of Warsaw

Project RAMKI (Rigorous Application of the Cognitive-Interpretational Methodology (Interpretative Frames) for Polish Language Description) is an attempt to apply a methodology of the frame semantics for the Polish language. A template and a benchmark for our project is the Berkeley's FrameNet. The aim of the project is to provide a description of about 200 Polish verbs within the frame semantics. The verbs were chosen according to two criteria: frequency in a large (ca. 100 million words) Corpus of Polish, and lexical equivalence to lexical unit already described in other (i.e. Berkeley, German and Spanish) FrameNets. As in Berkeley FN each lexical unit entry will be tagged with the surface syntactic properties, the interpretative frame activated by the lexeme and also examples from the corpus, tagged both syntactically and semantically, with the appropriate frame elements. A lexicographer is provided with a dedicated application, a computer program that helps to select examples from the corpus, annotate the data syntactically and semantically; it also keeps the data in a database to facilitate searching according to various criteria, such as sentence patterns, frames and frame elements, and outputs the data for www presentation.

# Dictionary Sense Division and Relation to Frames

Dorota Kopcińska, Jadwiga Linde-Usiekiewicz

University of Warsaw

The paper examines the degree in which sense division in Polish language dictionaries matches frame relations. The word in question is *jechać* in the sense in which it takes a human subject, e.g. *jechać pociągami, samochodem* 'go by train, by car'. Polish dictionaries tend to focus on the use of a vehicle for this sense, in contrast to *iść*, which means going on foot. Such a vague sense of *jechać* matches a non-lexical Motion frame, which is too general to account for the meaning of the Polish verb. In most cases actual examples can be easily matched with more detailed frames, i.e. Operate\_vehicle, Ride\_vehicle and Travel. It seems, however that semantic and syntactic behavior of the verb in question warrants a more detailed sense division that matches the three frames. The sense Operate\_vehicle is distinguished from the sense Ride\_vehicle by the fact that in the former the Frame element SPEED can be expressed by the adjunct *z szybkością . . .* 'at a speed of . . .', connected directly to the verb *jechać*. In the latter, this element has to be introduced in the subordinate clause. The Travel sense differs from the other two on semantic grounds. While in the Operate\_vehicle and, Ride\_vehicle senses the verb *jechać* can refer only to motion on the ground (by car, train, coach, bicycle, on horseback) and is contrasted with *lecieć* 'fly, go by plane' and *płynąć* 'sail', in the Travel sense it can refer to a plane trip or a sailing trip as well.

# Description of Verbs in Polish FrameNet Project Based on the Example of IŚĆ ('to go')

Witold Kieraś

University of Warsaw

The aim of my talk is to present how verbs in Polish FrameNet project (RAMKI) are described, based on the example of the verb IŚĆ 'to go'. Twenty homonymous lexical units IŚĆ were extracted from a dictionary *Inny słownik języka polskiego* (ISJP) and initially assigned to four different FrameNet frames (MOTION, BECOMING, ATTACK,

CHANGE\_POSITION\_ON\_THE\_SCALE). Lexicographer's task was to assign these lexical units properly either to one of the existing FrameNet frame, or to some newly created frame, find corpus examples for these units, assign semantic roles and choose a proper syntactic sentence scheme.

My talk is based on the examples of PATH\_SHAPE, BEING\_OPERATIONAL, MOTION and SELF\_MOTION frames. At least in some of these frames a lexicographer may identify certain problems. It needs to be decided whether some lexical units should be merged or rather split into parts and included into other units (and as a consequence into other frames). These decisions are not straightforward because of the differences between FrameNet (and RAMKI project) and traditional dictionaries. Some problematic examples from ISJP will be presented.

As a result selected lexical units were assigned to 15 different frames, two of them are new (non-existing in FrameNet). Also three new lexical units were identified and added on the basis of corpus research. Some others were merged or found idiomatic and deleted.

# Some Questionable Issues of the FrameNet: the Case of the Death and Killing Frames

Magdalena Zawisławska

University of Warsaw

The theory of Ch. Fillmore places emphasis on the connection between the conceptual level and its linguistics description. The main premises of Fillmore' idea is that frames are tools which can be used to explain lexical and grammatical meaning. The first application of the Fillmore's idea is the project FrameNet. Although the Fillmore's ideas seem very promising, not all realisations are satisfying. The analysis of two examples: frames Death and Killing shows, that it is not clear enough what are the relations between frames, on what bases particular lexical units are assigned to the particular frame and what situation exactly is by the frame described (the frame definitions are more like glosses than scenarios). The lexical units assigned to the same frame can also differ very much semantically, but it is not considered in the frame description.

## Authors

- Janusz S. Bień**, Formal Linguistics Department, University of Warsaw, Warsaw, Poland. → 215
- Oleg Bugakov**, Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine, Kyiv, Ukraine. → 194
- Magdalena Derwojedowa**, University of Warsaw, Warsaw, Poland. → 219
- Ludmila Dimitrova**, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria. → 76, 123
- Katarzyna Drożdż-Łuszczak**, University of Warsaw, Warsaw, Poland. → 217
- Matej Ďurčo**, Austrian Academy Corpus, Austrian Academy of Sciences, Vienna, Austria. → 128
- Peter Ďurčo**, St. Cyril and Methodius University, Trnava, Slovakia. → 128
- Maksim Dushkin**, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland. → 189
- Ralitsa Dutsova**, Veliko Tŕrnovo University, Veliko Tŕrnovo, Bulgaria. → 76
- Tomař Erjavec**, Department of Knowledge Technologies, Jořef Stefan Institute, Ljubljana, Slovenia. → 106
- Darja Fiřer**, Department of Translation, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia. → 106
- Radovan Garabík**, Slovak National Corpus department, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia. → 123, 128
- Ewa Gwiazdecka**, University of Warsaw, Warsaw, Poland. → 198
- Björn Hansen**, Institut für Slavistik, Universität Regensburg, Germany. → 211
- Leonid Iomdin**, Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. → 69
- Witold Kieraś**, University of Warsaw, Warsaw, Poland. → 221
- Dorota Kopcińska**, University of Warsaw, Warsaw, Poland. → 220
- Małgorzata Korytkowska**, University of Łódź, Lodz, Poland. → 18
- Violetta Koseska-Toszeva**, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland. → 9, 24, 76
- Barbara Kryřan Stanojević**, Faculty of Humanities and Social Sciences, University of Zagreb, Croatian. → 216
- Jadwiga Linde-Usiekiewicz**, University of Warsaw, Warsaw, Poland. → 219, 220
- Daniela Majchráková**, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia. → 128
- Antoni Mazurkiewicz**, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland. → 24
- Ewa Miczka**, University of Silesia, Poland. → 213
- Petya Osenova**, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia Bulgaria. → 115, 205



- Rumyana Panova**, Veliko Tŕrnovo University, Veliko Tŕrnovo, Bulgaria. → 76
- Jelena Parizoska**, Faculty of Humanities and Social Sciences, University of Zagreb, Croatian. → 218
- Maciej Piasecki**, Instytut of Informatics, Politechnika Wroclaw University of Technology, Wroclaw, Poland. → 32
- Adam Przepiŕrkowski**, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland. → 138
- Danuta Roszko**, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland. → 145
- Roman Roszko**, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland. → 9, 145, 159
- Joanna Satoła-Staŕkowiak**, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland. → 180
- Olga Shemanaeva**, Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. → 169
- Volodymyr Shyrov**, Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine, Kyiv, Ukraine. → 89
- Kiril Simov**, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria. → 115
- Mateusz-Milan Stanojević**, Faculty of Humanities and Social Sciences, University of Zagreb, Croatian. → 216
- Mária Ŗimková**, L'. Ŗtŕr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia. → 123
- Svetlana Timoshenko**, Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. → 169
- Jadwiga Wajszczuk**, University of Warsaw, Warsaw, Poland. → 219
- Andrŕ Wlodarczyk**, CELTA — Centre de Linguistique Thŕorique et Appliquŕe, Paris, France. → 44
- Hŕlŕne Wlodarczyk**, CELTA — Centre de Linguistique Thŕorique et Appliquŕe, Paris, France. → 56
- Zofia Zaron**, University of Warsaw, Warsaw, Poland. → 217
- Magdalena Zawisławska**, University of Warsaw, Warsaw, Poland. → 219, 222