

COGNITIVE STUDIES | ÉTUDES COGNITIVES

COGNITIVE STUDIES
ÉTUDES COGNITIVES

9

Editorial Board

VIOLETTA KOSESKA-TOSZEWA – the editor-in-chief,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

Members of Editorial Board

WIESŁAW BANYŚ,
Uniwersytet Śląski, Katowice

BJÖRN HANSEN,
Universität Regensburg
Institut für Slavistik

STANISŁAW KAROLAK,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

ANDRÉ WŁODARCZYK
Centre de linguistique théorique et appliquée (CELTA)
Université de Paris-Sorbonne

HÉLÈNE WŁODARCZYK
Centre de linguistique théorique et appliquée (CELTA)
Université de Paris-Sorbonne

ROMAN ROSZKO – secretary of Editorial Board,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

INSTITUTE OF SLAVIC STUDIES
POLISH ACADEMY OF SCIENCES

COGNITIVE STUDIES
ÉTUDES COGNITIVES

Vol. 9



SLAWISTYCZNY
OŚRODEK
WYDAWNICZY

WARSAW 2009

COGNITIVE STUDIES
ÉTUDES COGNITIVES

9

Comité de rédaction

VIOLETTA KOSESKA-TOSZEWA – rédacteur en chef,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

Membres du comité de rédaction

WIESŁAW BANYŚ,
Uniwersytet Śląski, Katowice

BJÖRN HANSEN,
Universität Regensburg
Institut für Slavistik

STANISŁAW KAROLAK,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

ANDRÉ WŁODARCZYK
Centre de linguistique théorique et appliquée (CELTA)
Université de Paris-Sorbonne

HÉLÈNE WŁODARCZYK
Centre de linguistique théorique et appliquée (CELTA)
Université de Paris-Sorbonne

ROMAN ROSZKO – secrétaire de la rédaction,
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa

ACADÉMIE POLONAISE DES SCIENCES
INSTITUT D'ÉTUDES SLAVES

COGNITIVE STUDIES
ÉTUDES COGNITIVES

Vol. 9



SLAWISTYCZNY
OŚRODEK
WYDAWNICZY

VARSOVIE 2009

Reviewed by
ZBIGNIEW GREŃ

The preparation of this edition of Cognitive Studies | Études Cognitives has been supported by the EC's Seventh Framework Programme [FP7/2007–2013] under the grant agreement 211938 MONDILEX

Cover design
MONIKA HANDKE

Editor of the volume and computer design
ROMAN ROSZKO \TeX

© Copyright by Sławistyczny Ośrodek Wydawniczy
Printed in Poland

ISSN 2080-7147

Sławistyczny Ośrodek Wydawniczy
Instytut Sławistyki PAN
Pałac Staszica, ul. Nowy Świat 72, 00-330 Warszawa
tel./fax [+48] 22 827 17 41 tel. [+48] 22 657 28 49
sow-ispan@wp.pl www.ispan.waw.pl

CONTENTS

Foreword.....	9
Stanisław Karolak	11
Hélie Włodarczyk	15
From Ontological Attributes to Semantic Feature Structures — Experimental Research on Aspect in Polish	
Violetta Koseska-Toszewa, Małgorzata Korytkowska, Roman Roszko	33
Contrastive Studies and Semantic Interlanguage	
Antoni Mazurkiewicz	53
Formal Description of Temporality (Petri Net Approach)	
Violetta Koseska, Antoni Mazurkiewicz	65
Net-Based Description of Modality in Natural Language (on the Example of Conditional Modality)	
Katarzyna Drożdż-Łuszczczyk, Zofia Zaron	79
Semantic Interrelations Between the Words <i>mistrz</i> and <i>uczeń</i>	
Darja Fišer, Tomaž Erjavec	89
Semantic Concordances for Slovene	
Peter Ďurčo, Radovan Garabík, Daniela Majchráková, Matej Ďurčo	101
Contrastive Dictionary of German and Slovak Collocations	
Ludmila Dimitrova, Violetta Koseska	117
Classifiers and Digital Dictionaries	
Ludmila Dimitrova, Violetta Koseska	133
Bulgarian-Polish Corpus	
Diana Blagoeva	143
Electronic Corpora and Bulgarian New-Word Lexicography	
Dorota Kopcińska, Jadwiga Linde-Usiekiewicz	151
Matching FrameNet Frames with Polish Sense Divisions: the Case of <i>jechać</i>	
Janusz S. Bień	161
Facilitating Access to Digitalized Dictionaries in DjVu Format	
Jelena Parizoska	171
Idiom variability in Croatian: the case of the CONTAINER schema	
Mateusz-Milan Stanojević, Barbara Kryżan-Stanojević	181
Levels of Constructional Meaning: the Confluence of the Dative and Middle Voice in Polish and Croatian	

Maksim Duškin	199
Concerning Exponents of Adnumeral Approximation in Polish and Russian	
Joanna Satoła-Staškowiak	211
Translating into Something that does not Exist... Literary Ways of Translating Polish Sentences with Uninflected Perfect Participles into the Bulgarian Language	
Julia Mazurkiewicz-Sułkowska, Agata Mokrzycka	223
From the Works on the Bulgarian-Polish Dictionary of Verbo-Nominal Analytical Constructions	
Ewa Miczka	233
Les structures situationnelles et informationnelles de discours	
Ewa Gwiazdecka	243
Quelle description pour le préverbe polonais ?	

FOREWORD

The **Editorial Committee** of our yearbook consists of two members from France and three members from Poland. At present, we propose to extend the Editorial Committee to incorporate Prof. Björn Hansen (Universität Regensburg, Institut für Slavistik), specialist in Slavic and Polish studies, and co-author of Polish-German grammar. The scientific editorial board of the journal reflects not only international cooperation between the respective Academies and Universities, but also the interdisciplinary character of the Team. At first, France was represented in the Editorial Committee of the yearbook by Prof. Jean-Pierre Desclés (computer scientist and mathematician) and Prof. Zlatka Guentchéva (linguist, language theoretician). At present, France is represented by Prof. Héléne Włodarczyk (specialist in Slavic studies, language theoretician), Director of the CELTA (Centre de Linguistique Théorique et Appliquée <http://celta.paris-sorbonne.fr/>) research centre at the Paris-Sorbonne University (Paris 4), and Prof. André Włodarczyk (specialist in Japanese studies, and linguist-computer scientist in CELTA at the Paris 4 University).

The scholars participating in the Editorial Committee from the **Polish side** include: Prof. Wiesław Banyś (linguist and specialist in Romance studies, language theoretician — Rector of the Silesian University), and a representative of the Institute of Slavic Studies PAS — Prof. V. Koseska-Toszewa (specialist in Slavic and Bulgarian studies, language theoretician, Editor-in-Chief, currently the representative of ISSPAS in the European Clarin Project and head of the Polish Side in the European Mondilex Project: FP7 Research Infrastructure, MONDILEX Project “Representing Semantics in Digital Lexicography” <http://www.mondilex.org>). Another IS PAS scientist on the Polish Side of the Committee used to be Prof. Stanisław Karolak (specialist in Slavic and Romance studies, language theoretician), who, to our great regret, died on June 5, 2009. The secretary of the yearbook is Associate Prof. R. Roszko (specialist in Slavic and Lithuanian studies, language theoretician — ISSPAS).

Members of the Editorial Committee have been dealing with cognitive problems of natural languages for years, using for that purpose the methods and theories of logic, as well as representations and tools originating from the field of computer science.

The articles are published in English (which has been the basic language of the yearbook since Issue 8), and — more rarely — in French. Each article is accompanied by an abstract, keywords and references.

The research presented in the **yearbook** is conducted in the spirit of the newest linguistic tendencies: semantics and confrontation of languages, semantic interlanguage (tertium comparationis), elaboration of language facts for the needs of com-

puter processing. This allows for analysis of not only formal grammatical language means, but also their lexical counterparts. Thank to this, the yearbook realizes the postulate of linguistic cognitive research, which consists in not separating the grammar from the lexis.

The yearbook is available both in a book (printed) version and in an electronic version, among others, at the Internet Journal Bookshop CEEOL, where one can buy the individual volumes or selected single articles. Some of its distribution places include:

- Internet bookshops, e.g.: the Internet Journal Bookshop CEEOL (<http://www.ceeol.com/asp/publicationlist.aspx>), Internet Bookshop of the Slavistics Publishing Centre (SPC / Slawistyczny Ośrodek Wydawniczy — SOW) at the Institute of Slavic Studies PAS (<http://www.ispan.waw.pl/sklep/>), and others www.kapitalka.pl, www.ksiegarnia-polska.com.
- traditional bookshops, e.g: the SPC bookshop (in the Staszic Palace in Warsaw),
- other distribution channels of the Slavistic Publishing Centre (cooperating warehouses and bookshops, mainly university ones, e.g. bookshops of: the Silesian University, the Pedagogical University of Krakow, Université Paris-Sorbonne (Paris 4), Université Denis Diderot (Paris 7), Université Charles de Gaulle (Lille 3), Institute of Mathematics and Computer Science of the Bulgarian Academy of Sciences in Sofia, Università degli Studi “G. d’Annunzio” di Chieti, and others.

In addition, the journal is ordered by libraries and scientific centres: Natsionalna Biblioteka Byelarussi, Minsk; Rossijskaya Gosudarstvennaya Biblioteka, Moskva; Narodna Biblioteka na Sofijskijat Universitet, Sofia; Slovanska Knihovna, Praha; Slovansky Ustav, Praha; Biblioteka Rossijskoj Akademii Nauk, St. Petersburg; Ros. Natsion. Bibl., St. Petersburg; INION RAN (Scientific Institute of Information and Social Sciences), Moskva; Institut Slavianovedeniya i Balkanistiki, Moskva; Univerzitetna Kniznica, Bratislava; Ustredna Kniznica SAV, Bratislava, Slovakia; Slaviska Institutionen vid Lunds Univer., Lund; Uppsala University, Dept. of Slavic Languages, Uppsala, Sweden; Congress Library, Washington; Indiana University Library, Bloomington.

Violetta Koseska

**Prof. Stanislaw Karolak, Dr h.c.,
(Institute of Slavic Studies, Polish Academy of Sciences, Poland)
died on 5 June this year**



Ceremony of decoration prof. Stanisław Karolak with the Officer Cross of the Polonia Restituta Order (Photo. Roman Roszko © 2004)

Each death is too fast and unexpected, even if sometimes we must slowly start coping with that thought, and accept the idea that it may happen any day now. However, when it finally comes, it is always premature and unexpected..... for there are still so many things to do and discuss together!

One of the most eminent linguists of the late 20th century has left us. An unusual man, with unusual will to live and scientific passion, unusual fortitude, unusual energy and unusual charisma, infecting others with his commitment and determination in discovering general language laws.

Prof. Karolak, a graduate of Warsaw University, in 1973 established Romance studies at the Silesian University, and chaired them for many years. He also reinforced and developed Romance studies at the Higher Paedagogical School (later Paedagogical Academy) in Krakow, and was, among others, a member of the Linguistics Committee of the Polish Academy of Sciences, of the Slavic Studies Committee PAS, and a member of the Neophilological Committee PAS. He cooperated with numerous linguistic centres in Poland and abroad, and was, among others, a Honorary Professor of Université Libre in Brussels, where he lectured for many years, and a Honoris Causa Doctor of the University Paris-XIII, which he cooperated with for over 30 years. He was the Knight of the French Order of Academic Palms, a member of the Polish Linguistic Society, a member of PAU, the Chair of the International Grammatical Commission at the International Committee of Slavists, and the Chair of the Aspectological Sub-commission affiliated at the International Grammatical Commission.

He educated a multitude of his faithful disciples and alumni, and the author of these words is honoured to be one of them.

His insight threw new light on the whole language system, since Prof. Karolak – “Staszek” to his friends — did not try to describe details — which are plentiful everywhere, including the language — for the sake of details themselves, but he studied them as part of a general conception and vision of the language, of a semantics-based grammar which he developed using the notions of a predicate and arguments borrowed from logic and adjusted to linguistics (see e.g. *Zagadnienia składni ogólnej* (1972), *Składnia wyrażen predykatywnych* In: *Gramatyka współczesnego języka polskiego* (1984)). That vision stemmed from the view of the world he had accepted, which followed in general terms from the principles of analytic philosophy, especially that developed by scholars like G. Moore, B. Russell, G. Ryle, P. Strawson or M. Dummett. The core of such a philosophy were relations and correspondence among logic, language and the external world. This resulted, among others, in seeking means for analysing problems, also philosophical ones, taking into consideration the language perspective, including the issues of meaning and reference and the way of using words. This was one of the reasons why Prof. S. Karolak gave its text entitled *Gramatyka a kształtowanie świadomości poznawczej* (Prace Filologiczne, 1992, Vol. XXXVII) a motto taken from *Myśli nieuczesane* by Polish humorist S. J. Lec, which renders very well the need for studying one of the aspects of those relations: *Obserwujemy ciekawe zjawisko: bełkot jako środek porozumiewawczy między ludźmi* [We can observe an interesting phenomenon: gibberish as means of communication among people].

Prof. Karolak’s interests were not limited to a general description of natural languages — where his work concerned not only French and Polish, but also Russian and Bulgarian (by education, he was a specialist in Slavic and Romance studies, and his Ph. D. thesis discussed verbal rection in Russian (*Zagadnienia rekcji przyimkowej czasownika w języku rosyjskim*, 1966, see also e.g. *Gramatyka rosyjska w*

układzie systematycznym, 1990), as well as English and Italian. Staszek was also involved in studying great areas of language operation, such as the communication structure of the language (thematically-rhematic structure), meaning, reference, quantification, article (see e.g. *Kwantyfikacja a determinacja w językach naturalnych* (1990), *Histoire et grammaire comparée des langues slaves* (1993), *Études sur l'article et la détermination* (1995), *Od semantyki do gramatyki. Wybór rozpraw* (2001)), and recently values of grammatical tenses and aspect.

His works are of fundamental importance for every researcher of the mentioned problem area. Prof. S. Karolak was able to discover even in the most complex nature of things the general rules and mechanisms that regulate the operation of the whole language system. One of the reasons why his works are so incomparable is because they show the invisible mechanisms that govern the operation of the visible world — of the surface of the language. His works on the French article (and the comparison of its functioning with English) take linguistics even deeper into the field of experimental sciences, since the intensional and meaning-related mechanisms presented there, as well as article usage, are described with sets of very precise rules — a feat which nobody managed to achieve earlier (see e.g. *Gramatyka kontrastywna rodzajnika francuskiego i angielskiego* (2002), *Rodzajnik francuski w ujęciu funkcjonalnym* (2004), *Jak stosować rodzajnik francuski* (1999)). Prof. S. Karolak always presented his arguments and defended them with great fervour. I remember one of his discussions which he conducted under one of the numerous international projects he participated in, with Prof. J.-P. Desclés from Paris-Sorbonne. After some time, both the discussion partners realized that the opponent's arguments were very much justified, so they rolled up their sleeves and started yet another round of discussion, which was both extremely inspiring intellectually and very emotional.

We will always retain a fond memory of Prof. S. Karolak. We are grateful to him for existing, for being her with us, for doing so much for us and for the world linguistics - for he has not wholly departed! He has left us, his disciples, with the research methods and linguistic aims he implanted in our minds, and the extremely inspiring achievements that have set the directions of linguistic research in Poland and abroad for many years.

Wiesław Banyś

Hélène Włodarczyk

Université Paris-Sorbonne

CELTA — Centre de Linguistique Théorique et Appliquée

FROM ONTOLOGICAL ATTRIBUTES
TO SEMANTIC FEATURE STRUCTURES
— EXPERIMENTAL RESEARCH ON ASPECT IN POLISH

Abstract. Our aim is the formalisation of semantic concepts using experimental Computer-aided Acquisition of Semantic Knowledge (CASK) method. This method is being used at CELTA with the SEMANA (SE-Mantic ANALyser) software which was especially designed for this purpose. The case study we propose is the experimental research on Aspect in Polish: we show how although rich nevertheless intuitive notions resulting from years of traditional research can be formalised with computational methods of data analysis.

We use the notation close to that of Semantic Feature Structures as a representation language for describing the category of aspect in Polish taking into account not only the grammatical core of this category but also the whole semantic field (called aspectuality¹ by A.V. Bondarko) regardless of its pertaining to various levels of expression (morphological, syntactical or lexical). To model the Category of Aspect we treat ontological 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.

The database of Polish Aspect was analysed using KDD statistical tools (Sauvet, G. [20]): thus, we obtained the first preliminary experience-based 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 such ontologies that were especially designed in order to account for

¹ In Russian *apektual'nost'*, [3].

linguistic objects (expressions). We claim that this kind of research tasks should be interactive (computer-aided), if we want to avoid the lack of precision and the variability of semantic parameters in traditional linguistic research, notably risky in the semantic domain.

We define the semantic content of a linguistic expression as a function mapping this expression, although through intermediary representations, onto 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 precisely 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 only the specific set of semantic feature structures for the aspect description in a given 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 (meta-theoretical) knowledge of the category they are studying and specific knowledge of the language they are describing.

2 Towards Experimental Semantics

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 been conducting this research in the field of aspect semantics in Slavic languages in contrast with non Slavic Indo-European languages, mostly French and English. In our view, the study of verbal aspect has been developed for such a long time and by so many researchers that the only hope to make any progress in this field is to look to new formalised methods making it possible to treat rich but not always precise information.

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. Needless to speak of the possibility to store large language data, to access them easily and share them with other researchers.

2.1 The CASK method

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 (to build semantic databases), so that linguistic theories consist in describing and analysing these data using symbolic and statistical methods.

This approach differs drastically from hypothetico-deductive linguistic theories (which, as a rule, use merely few examples as illustrations) but it is confronted with the serious problem of meta-data input. In fact, we consider that raw facts in the domain of language do not exist as such, and any description of linguistic reality relies on a felicitous set of meta-theoretical assumptions.

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

The difficulty of data input is partly due to that the linguistic expressions in their contexts exhibit not only explicit meanings but also entail as well presupposed as inferred knowledge. Researchers should be therefore conscious of the fact that it is often difficult to establish which part of the presupposed or inferred knowledge is pertinent in a given context.

2.2 SEMANA: a software designed for interactive (man-machine) semantic research

The problem of communication between linguists and computer scientists comes from the times when the latter used to analyse the formers' needs for their specific domain at a given stage of their research in order to build adapted tools. Most often such tools reflected the image (knowledge) of the domain 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 software², 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 and (2) tools for automatic KDD analyses.

Due to its interactivity and interoperability the db-Builder module (dynamic database builder) in Semana, it is always possible, during the research process which heavily relies on the linguist's expertise in a given domain, to modify features and their values (and the structure they belong to) as soon as the progress in research proves it necessary. Each card in 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 module is complemented by that of Tree Builder Assistant which allows the linguist to organise the attributes and values chosen for the description of a semantic field in a tree structure. Any change in the feature tree of the Tree Assistant is transferred to each record of the database after the

² <http://www.celta.paris-sorbonne.fr/anase/m/papers/>

linguist is asked whether he accepts changes in the tree builder to be echoed in the database. Since 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 restructuration 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 input of each card and makes it possible to display the global and partial feature trees (in Tree Builder Assistant). Many automated checks are available and help the linguist to verify the consistency of his description.

The second part of Semana is main and contains algorithms for various KDD analyses integrating tools on the three following theories: 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 [25].

3 The categorial and contextual meaning of Aspect

It is often a very delicate issue to distinguish the explicit categorial aspect meaning of an expression used in an utterance from its inferences and presuppositions. The latter are sometimes referred to globally as the “pragmatics” of the category 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.

First, properly aspecto-temporal senses of verbs in context must be taken into account. Such senses do not depend only on the paradigmatic categorial meaning of the Aspect category but also on the utterance in which the verb is used. Much has been written about this problem by e.g. Jakobson, R. [7], [8] (*Gesamtbedeutung / Sonderbedeutung*), Kuryłowicz, J. [12] (*primary* and *secondary* meanings). The Russian linguist Bondarko, A. V. also a specialist of verbal aspect, devoted much attention to the formalisation of this problem for which he used feature; he opposed the *semantic potential* of a grammatical category to its *particular meanings*, moreover among all features belonging to the *potential meaning* of a category, the one which is present in all *particular meanings* is considered *dominant* [2, p. 103].

For instance, in some languages, usages 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.

Second, 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 (cf. [31]). We have devoted and continue to devote much attention to this part of the aspectual problem, which reveals very important when contrasting languages: both the grammaticised noun determination systems (articles) and grammaticised aspect systems are involved in the marking of the meta-informative Old or New status of utterances (cf. [17]). In this paper we do not develop this problem because, for the time being, we do not tackle this problem in the Semana database in which we limited the description of aspect uses to the first sort of uses, i.e. what we consider as properly aspectual uses.

4 Aspect as a hypercategory

It is well-known that verbal aspect in Slavic languages consists in the opposition of two values (perfective and imperfective) only; however, we claim that this binary aspectual opposition is present 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 perfective³. Among them, two classes can be distinguished though the border between these classes is not sharp : (1) lexical derivatives are verbs with a new lexical meaning and (2) aspectual derivatives are verbs which keep the same lexical meaning but differ from the root verb by some specific aspectual feature(s). Among the aspectual derivatives, slavacists distinguish traditionally (since the beginning of the 20th century, cf. [1]) between a series of derived verbs representing the aspect-related meaning named *Aktionsart* (or *lexical* aspect) and *one* derived perfective verb considered as the unique grammatical perfective partner of the simple verb. Recently, this distinction has been reappraised by many researchers (including, Sémon [22], Padučeva [17], Karolak [10], Xrakovskij [35]). 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. The first pioneer work in that direction was that of Cezar Piernikarski [18] who showed that there exist several different semantic types of “aspectual” oppositions and that not only derived perfective verbs with resultative meaning should be considered as real aspectual partners of simple imperfective verbs.

³ Exceptions are extremely rare and due to diachronic reasons, e.g. in Polish, imperfective verb *należć* (to belong) is derived from imperfective simple verb *leżeć* (to lie).

Our cognitive approach makes it possible to bring closer both sub-categories of Aspect and *Aktionsart* in order to substitute the notion of “aspectual pair” by that of aspectual nest when defining the perfective aspect of Slavic languages [26]. 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 belonging to the ontological class of dynamic situations, most often expressed by transitive verbs, e.g. *budować* imp. / *zbudować* perf. (to build). Verbs referring to static situations like *leżeć* (to lie) were considered as *imperfectiva tantum* and such derivative like *poleżeć* (to lie for a while) were not treated “pure” as a perfective partner of the simple verb but as an *Aktionsart* derivative.

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* verbs. 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 database⁴ of the Polish Frequency Dictionary (SFPW) we studied the relative frequency of simple imperfective verbs and the different perfective verbs derived from them [26]. It appears that the frequency of the verb traditionally considered as the unique or “real” perfective partner is much higher than the frequency of verbs considered as *Aktionsart*. That is probably the reason why the most frequent perfective partner was generally considered as *the pure grammatical perfective partner*. Below we give a small sample in the form of a comparative table (Table 1) of the frequency of simple imperfective verbs and their derived perfectives⁵, traditionally distinguished as *Aktionsart* verbs and perfective partners.

In view of this quantitative data we put forward the hypothesis that the perfective aspect should be regarded as a hypercategory. According to this hypothesis, each verbal prefix inherits a specific bundle of aspectual features. No verbal prefix can be viewed as semantically void, on the contrary different configurations of features are possible for each prefix. As a matter of fact, since the perfective aspect may be expressed by more than one verb derived from a single verbal root, none of them can be considered as a synonym of the root verb. Prefixed verbs that are traditionally described as *Aktionsart* verbs should be described as differ-

⁴ The electronic file of SFPW was made available to us by Prof. dr Janusz Bień (University of Warsaw) to whom we express hereby our thanks.

⁵ Only such derived perfective verbs are quoted which appear at least once in the Frequency dictionary. For instance, current verbs as *wyspać się* and *zaspać* which belong to the lexicon of Polish native speakers do not appear at all.

ent specialisations of one perfective hypercategory⁶. Thus, the perfective aspect is an abstraction that subsumes numerous different specific versions.

<i>robić</i> 265	<i>zrobić</i> 268, <i>porobić</i> 2
<i>czekać</i> 184	<i>poczekać</i> 39, <i>zaczekać</i> 11, <i>przeczekać</i> 1, <i>odczekać</i> 4, <i>doczekać</i> 3
<i>styszeć</i> 181	<i>usłyszeć</i> 50, <i>postyszeć</i> 5
<i>pytać</i> 85	<i>zapytać</i> 83, <i>spytać</i> 40, <i>popytać</i> 1, <i>napytać</i> 1, <i>rozpytać</i> 1
<i>spać</i> 72	<i>pospać</i> 1, <i>przespać</i> 2, <i>przespać się</i> 5
<i>budzić</i> 39	<i>obudzić</i> 15, <i>zbudzić</i> 7, <i>przebudzić</i> 1, <i>rozbudzić</i> 1, <i>pobudzić</i> 1,
<i>pić</i> 47	<i>wypić</i> 27, <i>napić się</i> 26, <i>popić</i> 5, <i>spić się</i> 1, <i>rozpić się</i> 1, <i>upić się</i> 3
<i>jeść</i> 45	<i>zjeść</i> 27, <i>pojeść</i> 1, <i>dojeść</i> 1, <i>przejeść się</i> 1
<i>bawić się</i> 26	<i>pobawić się</i> 2, <i>zabawić się</i> 4

Table 1. Relative frequency of simple imperfective verbs and their derived perfective verbs

To give an illustration of this, we can take as an example the Polish simple imperfective verb *spać* (*to sleep*). When we are looking for a perfective partner of this verb we must take into account the context in which it is used and we have to choose one of the derived perfective verbs belonging to its derivational nest : *wyspać się* (‘sleep as long as you need’), *dospać*, ‘sleep some more’, *pospać*, ‘sleep for a while’, *przespać*, (1) ‘miss something while sleeping, or (2) ‘przespać coś’, ‘sleep away some period of time’, *zaspać*, ‘sleep too long and miss something’.

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 aspectual pairs (defined as two verbs of opposed aspect but sharing the maximal amount of common semantic features) are constituted only by an imperfective suffixed verb which 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 concept of hypercategory is based on the multiple inheritance or heterarchy: As shown above, the aspectual meaning of a verb used in a given context can be described as a bundle of several attributes. Thanks to inheritance, we do not have to build disjoint classes of aspectual meanings (as was the case with classes of Aktionsart verbs). In fact, aspectual meanings are linked irregularly to different superior nodes. Moreover, one and the same verbal lexeme may have different links depending on the context in which it is used. To give an example, this is often the case of *po-* prefixed verbs that can have several meanings (cf. Włodarczyk, A. & H. [26], II).

This approach allows us to take into account the lexical diversity in Slavic aspectual semantics and sheds light on the controversies about which features are

⁶ The term « allogram » proposed by Czochralski [6, p.43] was coined on the model of « allophone », it captures partly what we mean by hypercategory: “[...] Die abwechselnden perfektiven Partner des Imperfektiven Aspektsverbs sind Varianten, die in komplementären Distributionsverhältnis zueinander stehen.”

grammatical and which are lexical. 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 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. It is precisely 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, that we proposed to define the perfective aspect as a hypercategory.

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 [31] and are defined in the meta-theoretical framework we outlined for interactive semantic research (Włodarczyk, A. [27]⁷, Włodarczyk, A. & H. [27, 28]). 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 viewed 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 usage 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 usage of aspect. This allows to propose a formal definition of Aspect as follows.

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

5.1 Types of semantic situations

The relevance of semantic situation types in aspectology was made famous by Vendler’s verb classification [23] and many authors wrote contributions following his. As regards the Polish language the most actual and precise classifications was proposed by Roman Laskowski [13]. The classification which we used [24] differs from others in that it does not incorporate features concerning situation participants and roles they enact⁸ but only situation frames. This classification is based

⁷ This paper points at the use of Petri Nets mathematical model in modelling aspect in linguistics. This model was first introduced into linguistic aspect studies by Mazurkiewicz [14] and Mazurkiewicz & Koseska-Toszewa [15]. It is now more widely used in aspect modelling: Chang [4], Chang & al. [5], Narayanan [16].

⁸ We take participants into account in order to define modality. However, when describing one use of a verb form in its context we consider the whole utterance and thus the

on four only relevant features: three-dimensional space, time, progression and granularity (Table 2).

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

Table 2. Classes of semantic situations [24]

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

Internal aspect (situation analysis) concerns only processes (dynamic situations with progression).

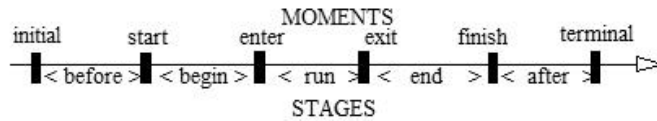


Fig. 1: Analysis of simple dynamic situations in moments and stages (constitutive parts)

The distinction 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 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

participants expressed as noun phrases are relevant for the semantic interpretation of the verb form.

different configurations with other parameters in order to give account of different usages of the perfective and imperfective verbs.

A situation characterised as 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

The description of the aspectual meaning of Polish verbs in context in the database showed that the initial moment is not expressed in utterances; what is relevant is the limit between the *before* stage and the *begin* stage of the situation expressed in an utterance: this is the first moment of the situation which we call the *start* moment. Although the terminal moment is not systematically explicitly expressed it is relevant in Slavic languages to take into account the after stage, about which it is possible to precise that it was put an end to it.

#1 *Otworzyłem okno. (I opened the window and it is still opened.)*

#2 *Otwierałem okno. (I opened the window but it has been closed back.)*

5.3 External Aspect: Control

What we call external aspect (« control » parameters) 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*⁹, *trans*¹⁰), 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. For example:

#3 *Pootwierałem wszystkie okna. (Lit. I opened all the windows one after the other.)*

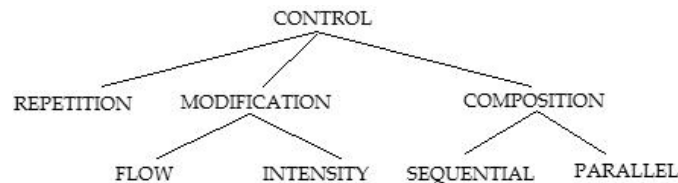


Fig. 2 Aspectual Parameters of Situation Control

⁹ The “off-and-on” flow modification concerns the unfolding of a situation intermittently.

¹⁰ We call “trans” the unfolding of the whole situation from start to finish (regardless of its stages)

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 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 db by Sauvet G).

6.1 From the work on the ontological tree of aspectual features

In the Aspect database, in order to annotate each aspectual specimen chosen in a corpus, we designed an ontological tree composed of the attributes and values (AV) described above. The aspectual value of each sample is described by a partial tree of attributes and values chosen in the common tree. As an example, fig. 3 is the partial tree used for the utterance:

#4 *Po kolacji Piotr znowu zabrał się do czytania. (After dinner, Peter resumed his reading.)*

```
ASPECT-*-ANALYSIS-*-{ANA}=run
      *-CONTROL--*-FLOWMODIF-*-FLOW-*-{MOD}=resume
ASPVALUE-*-ASPVAL-*-{VAL}=perfective
SITTYPE-*-{TYP}=ordProcess
```

Fig. 3 Partial tree for utterance #4

The complete tree was first designed in the Tree Assistant of the Semana software 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 the reason for it is 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. On the other hand, however, the moment we call *terminal* is sometimes explicitly expressed in an utterance as the end of the *after* stage when it is conceived as a new state occurring after the end of a process. Thus, only the moments (schematized by points on a line) indicating the border between the outer stages and the situation itself can be expressed directly in verbal expressions of aspect : the moment we call *start* constitutes both the ending point of the *before* stage and the beginning point 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 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 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 either remove them from the description or accept them (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 in order to give account of fine-grained semantic differences) or (2) when the same prefix is characterised by two or more different feature trees (polysemous prefix). It is a well-known fact that in Slavic languages the same aspectual meaning may be expressed by more than one morpheme (prefix or suffix), e.g., the so called ingressive Aktionsart is considered to have at least two markers in Polish: the prefix *po-* or the prefix *za-*. The linguistic material provided by large databases should help the linguist to determine to which degree two prefixes can be considered as synonymous. On the contrary, the material stored in the database can be used to study

the polysemy of each morpheme as soon as it is possible to compare all samples which have the same index, e.g. the linguist can compare all samples having the prefix po- and try to find out whether or not there is (are) some common semantic feature(s) between the different uses. This problem is crucial in Slavic aspectology; in the last years, it has been revisited from the point of view of cognitive linguistics but without relying on large databases [19].

6.3 Successive versions of the Aspect db

As we progressed by trials and errors in different versions of the Aspect db we obtained gradually improved statistical reports. In Table 3 column 1 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. As we limited the number of attributes 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 3. 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 :

```
ASP-*.-ANLS-*.-{ANA}=[ent|exi|fin|str|ini|term|beg|end|run|bef|aft|nan]
    *-CNTL-*.-FMOD-*.-FLOW-*.-{MOD}=[int|kp|res|stp|trans|par|seq]
    *      *      *-REP--*.-{CRE}=[dnb|indnb|OaO]
    *      *      *-ITS--*.-{ITS}=[inc|dec|stg|wea]
```

```
AVAL-*.-AVAL-*.-{VAL}=[imp|prf]
    *-MCMP-*.-{MCP}=[ip|pp|ii|pi|pip]
```

```
SIT-*.-{TYP}=[evt|oPr|rPr|sta]
```

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

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

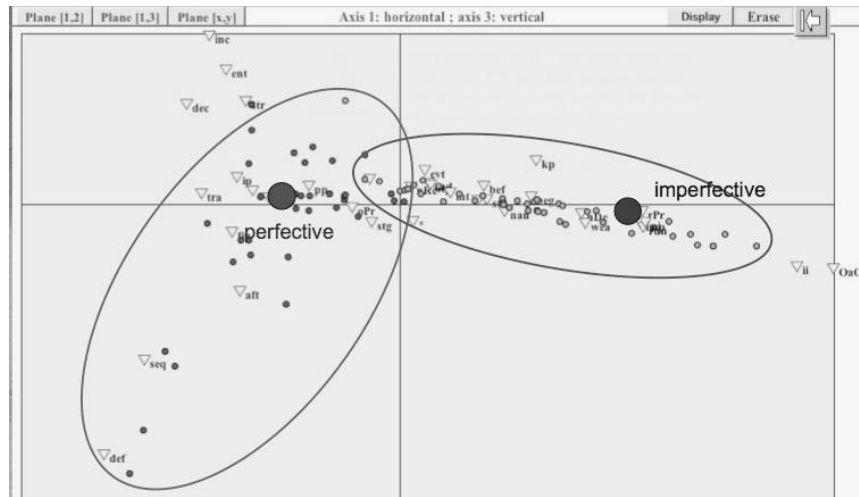


Figure 4. Two classes of values of attributes around the perfective and imperfective

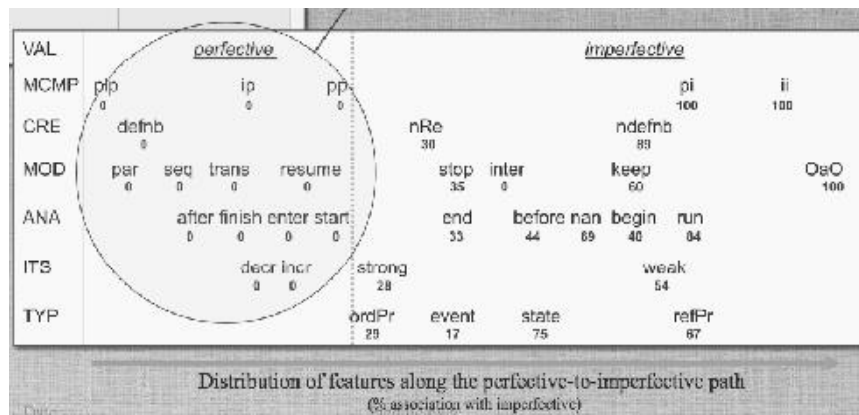


Figure 5. Polish Aspect: Correspondence factor Analysis

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 in aspectology 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 (i.e. ordinary process situations can be expressed both by imperfective and perfective verbs) 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. 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.

7 Experimental semantics as a basis for contrasting languages by alignment of ontological structures

Ontology-based and formalised 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 corpora (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 identical but also similar feature structures. This is very useful for translation methods because a text in language L1 is never absolutely identical to its translation in language L2.

Acknowledgment

The research on aspect using the Semana software has been conducted at CELTA (Paris-

Sorbonne University) since 2005 by H el ene and Andr e Włodarczyk (ontology and semantics of verbal aspect), Georges Sauvet and Andr e Włodarczyk (designers of the Semana software), Georges Sauvet (author of the analysis of my aspect database with KDD tools he implemented in Semana). Doctoral and master students also took part in this research.

Bibliography

- [1] Agrell, S. (1908). *Aspekt nderung und Aktionsartbildung beim polnischen Zeitworte*, Lund.
- [2] Bondarko, A. V. (1971). Grammaticeskaja kategorijej kontekst, Leningrad, Nauka.
- [3] Bondarko, A. V. (1971). Vid i vremja russkogo glagola, Moskva, prosvescenije.
- [4] Chang, N. (1997). “A Cognitive Approach to Aspectual Composition”, ICSI TR 97-034.
- [5] Chang, N., Gildea, D., Narayanan, S. (1998). “A Dynamic Model of Aspectual Composition”, Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society COGSCI-98. Madison.
- [6] Czochralski, J. A. (1975). *Verbalaspekt und Tempussystem im Deutschen und Polnischen, Eine konfrontative Darstellung*, Warszawa, PWN.
- [7] Jakobson, R. (1932). «Zur Struktur des russischen Verbums», *Charisteria Guilelmo Mathesio... oblata*, Prague, 74–84.
- [8] Jakobson, R. (1936). «Beitrag zur allgemeinen Kasuslehre (Gesamtbedeutung der russischen Kasus)», TCLP, VI, Prague, 240–288.
- [9] Karolak, S. (Ed.) (1995–1999). *Semantika i struktura slavjanskogo vida*, I 1995, II 1999, Wyd. Naukowe WSP, Krak w.
- [10] Karolak, S. (1997). “Arguments contre la distinction: aspect / modalit  d’action (Aktionsart)” in *Etudes cognitives*, Vol.2, SOW, Warszawa, 175–192.
- [11] Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., Woro czak, J. (Eds.) (1990). *S ownik frekwencyjny polszczyzny wsp łczesnej*, PAN, Instytut J zyka Polskiego, Krak w.
- [12] Kury owicz, J. (1977). *Probl mes de linguistique indo-europ enne*, Wroc w, Warszawa, Krak w, Gda nsk, WPAN Ossolineum.
- [13] Laskowski, R. (1998). “Uwagi o znaczeniu czasownik w” in *Gramatyka wsp łczesnego j zyka polskiego, Morfologia*, wyd. 2 zmienione, T. 1, PWN, Warszawa: 152–171
- [14] Mazurkiewicz, A. (1986). “Zdarzenia i stany: elementy temporalno ci”, *Studia gramatyczne bu garsko-polskie*, t. I, Instytut S owianoznawstwa PAN, Ossolineum, Wroc w, Warszawa, 7–22.
- [15] Mazurkiewicz, A., Koseska-Toszewa, V. (1991). “Sieciorowe przedstawienie temporalno ci i modalno ci”, *Studia gramatyczne bu garsko-polskie*, t. IV, Res Publica Press, Warszawa.
- [16] Narayanan, S. (1997), “Talking The Talk Is Like Walking the Walk: A Computational Model of Verbal Aspect ”, Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society COGSCI-97. Stanford: Stanford University Press
- [17] Padu eva, E. V. (1996). *Semanticeskije issledowanija*, Moskva

- [18] Piernikarski, C. (1969). *Typy opozycji aspektowych czasownika polskiego na tle słowiańskim*, Ossolineum, Wrocław.
- [19] Przybylska, R. (2006). *Schematy wyobrażeniowe a semantyka polskich prefiksów czasownikowych do-, od-, prze-, roz-, u-*. Językoznawstwo kognitywne, studia i analizy. Kraków.
- [20] 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/>.
- [21] Sémon, J.-P. (1979). "L'Acte itératif nommé et l'aspect", *II^e Colloque de Linguistique russe*, Paris avril 1977, 87–104, Institut d'études slaves, Paris.
- [22] Sémon, J.-P. (1986). "Postojat' ou la perfectivité de congruence, définition et valeurs textuelles", *Revue des Etudes Slaves*, T. 58/4, Institut d'Etudes Slaves, Paris.
- [23] Vendler, Z. (1967). "Verbs and Times", in Z. Vendler, *Linguistics and Philosophy*, Ithaca, New York: Cornell University Press.: 97–121 (revised version of Vendler, Z. "Verbs and Times", *The Philosophical Review*, 66 (1957), 143–160)
- [24] Włodarczyk, A. (2003). « Les cadres des situations sémantiques », *Etudes cognitives / Studia kognitywne V*, SOW, Polish Academy of Sciences, Warszawa.
- [25] Włodarczyk, A. (2009). "Interactive Discovery of Ontological Knowledge for Modelling Language Resources", MONDILEX Project's Warsaw Conference, Warszawa. <http://celta.paris-sorbonne.fr/anasem/papers/>.
- [26] Włodarczyk, A., Włodarczyk, H. (2001). « La Préfixation verbale en polonais I. Le statut grammatical des préfixes, II. L'Aspect perfectif comme hypercatégorie », *Etudes cognitives / Studia kognitywne IV*, SOW, Warszawa: 93–120.
- [27] Włodarczyk, A., Włodarczyk, H. (2003). « Les paramètres aspectuels des situations sémantiques », *Etudes cognitives / Studia kognitywne V*, SOW, Warszawa: 11–34.
- [28] Włodarczyk, A., Włodarczyk, H. (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: 389–408.
- [29] Włodarczyk, A., Włodarczyk, H. (2008). « The Pragmatic validation of Utterances », in *Etudes cognitives / Studia kognitywne VIII*, SOW, Warszawa: 117–128.
- [30] Włodarczyk, A., Włodarczyk, H. (2009). "Interactive Discovery of Ontological Knowledge for Modelling Language Resources", MONDILEX Project's Warsaw Conference, Warszawa.
- [31] Włodarczyk, H. (1997). *L'Aspect verbal dans le contexte en polonais et en russe*, Institut d'Etudes Slaves, Paris, 240 p.
- [32] Włodarczyk, H. (1998). "Wykładniki wartości informacyjnej wypowiedzenia w j. polskim i francuskim (aspekt, określoność, modalność)", *Congrès des Slavistes Cracovie 1998, Revue des Études Slaves* T. LXX/1: 53–66, Paris.

- [33] Włodarczyk, H. (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: 327–338 Paris.
- [34] Włodarczyk, H. (2008). « La place du temps dans la théorie cognitive de l'aspect (perfectifs d'achèvement et d'interruption en polonais et en russe) », *Le Temps construit, Mélanges offerts à Jean-Paul Sémon*, dir. Jean Breuillard, Institut d'Etudes Slaves, Paris: 109–131.
- [35] Xrakovskij, V.S. (1997). “Mul'tiplikativy i semel'faktivy (problema vidovoj pary)”, *Semantika i struktura slavjanskogo vida*, red. S. Karolak, Wyd. Naukowe, Kraków.

VIOLETTA KOSESKA-TOSZEWA¹
MAŁGORZATA KORYTKOWSKA²
ROMAN ROSZKO¹

¹Institute of Slavic Studies, Polish Academy of Sciences, Poland

²University of Łódź, Poland

CONTRASTIVE STUDIES AND SEMANTIC INTERLANGUAGE

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.

The approach presented in the Part 2 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.

Keywords: Semantic, Contrastive Studies, Interlanguage, Predicative unite, Experiencer, Agentive.

PART 1.

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 [23]. 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 ([24], [23], [6, 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 [21], 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 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 [21], 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 2nd 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 [20]). 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 of *Confrontative Bulgarian-Polish Grammar* [GKBP] [8] 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 [9]). 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 [8], [7]).

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” [14, p. 101]. The above quotation illustrates the multifarious applications of the English terms state and event in linguistics (see also [15]).

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 [17]. 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” [19, 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 [17].

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* [7].

6 Selected notions of the interlanguage with elements of its metalanguage

Collective quantitative quantification / collective quantification (concerns quantification₂). 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 quantification₂). 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 quantification₂). 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 «the door is closed», and a global state will be «the door is closed, the windows are open». 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 quantification₂). 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 quantitative multiple quantification.

Numerical quantification, see quantification₂

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

Precedence — succession relation. Succession relation. 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 «growing», 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_1 and quantification_2 . Operation of binding variables with a quantifier (numerical, universal, existential one, or iota-operator).

Quantification₁ / (=logical quantification) / **scope-based quantification**, see quantifying . Quantification_1 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 [18, p. 211–255] and definite description [19, p. 253–293], see [8]. 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 (ιX) «is a two-legged featherless being» 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* [8], 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.

Quantification₂ / **quantitative quantification (=numerical quantification)**. Quantification_2 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 quantification₂). 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 quantification₂

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

Scope-based existential quantification / existential quantification / existentiality (concerns quantification₁). 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 quantification₁, 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 quantification₁). 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 quantification₁). 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 quantification₁). 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 quantification₂). 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 quantification₁). 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 praesenti 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.

PART 2.

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*'¹ 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 [22], 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 1st actant was identified with the subject position, of the 2nd actant — with the direct object's position, of the 3rd 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 1st actant, which contradicts the definition of the latter position, assigning the 1st 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 1st actant's position is assigned to the phrase *I* in English and to the phrase *Vous* in French [22, 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 here even most briefly the consecutive theoretical propositions referring to that problem area [16], [4], [3]. Doubtlessly, the next breakthrough attempt to refer to it was the theory of so-called semantic cases proposed

¹ A broader review of theoretical positions on those issues can be found in [11].

by Ch. Fillmore [5]. 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. [5, 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 positions². These will be kinds of labels which are associated with a specific place at a certain type of possibly simple semantically predicate³. 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

² A detailed description of such a model and of its application to contrastive analysis is given in [11].

³ 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.

by paraphrasing, see e.g.: *Paweł wgniółł maskę samochodu. — Paweł zrobił coś, co spowodowało, że maska samochodu jest wgnieciona. / Skąta wgniotła maskę samochodu. — Stało się coś ze skąką* (see *Spadająca z góry skąta*), *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 skały ze zbocza 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)⁴. 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 ^S')*; Bulg. *x чувствава / вижда / чува / знае (че ^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 voluntatively, 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.:

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

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)⁵. 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), *dlatego nie przyszła*.
 Bulg. *Ана ми* (Exp) *обясни, защото не е дошла*.

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 voluntative stand. Hence the context of sentences with predicators opening an Agentive position may contain causative clauses with voluntative 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).

⁵ An extensive analysis of the material is given in [11].

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* – Bulg. *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 глед-*
ката 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 multifarious 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 structure⁶.

⁶ For analysis of nominalization process for sentences which realize a propositional argument, see [13]

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 place⁷, 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.*).

⁷ More broadly on processes which lead to content reduction see [12].

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*⁸, 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:

- Pol. *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 then 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:

- Pol. ***Wysyłanie skarg przez Adama szkodzi awansowi / wygranej / zdrowiu Piotra.***
- Adam szkodzi awansowi / wygranej / zdrowiu Piotra wysłaniem 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

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

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] Projekt (1984). Projekt gramatyki konfrontatywnej bułgarsko-polskiej i serbsko-chorwacko-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] Daneš, F. (1968). Sémantická struktura větneho wzorce. In *Otázky slovanské syntaxe II*, pages 45–49, Brno.
- [4] Daneš, F., Hlavsa, Z. (1973). Hierarhizace sémantické struktury věty. In *Československé přednášky pro VIII. Mezinárodní sjezd slavistů v Zahřebu*, pages 67–77, Praha. Lingvistika.
- [5] Fillmore, Ch. J. (1968). The case for case. In *Universals in Linguistic Theory*, Bach, E., Harms, R. T. (eds.), pages 1–88.
- [6] Heinz, A. (1978). *Dzieje językoznawstwa w zarysie*. Warszawa.
- [7] Koseska-Toszewa, V., Korytkowska, M., Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Dialog, Warszawa.
- [8] Koseska, V., Gargov, G. (1990). *Семантичната категория определе-ност/неопределеност, Българско-полска съпоставителна граматика, т. 2*. София.
- [9] 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.
- [10] Korytkowska, M. (2004). Wokół problemów opisu kategorii kauzatywności i sposobów jej realizacji (na przykładzie języka bułgarskiego i polskiego). *Slavia Meridionalis*, 4.
- [11] Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska, t. 5, Typy pozycji predykatowo-argumentowych*. SOW, Warszawa.
- [12] Korytkowska, M., Kiklewicz, 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*.
- [13] Korytkowska, M., Maldziejewa, W. (2002). *Od zdania złożonego do zdania pojedynczego. Dopuszczalność nominalizacji argumentu propozycjonalnego w języku polskim i bułgarskim*. Toruń.

- [14] Lyons, J. (1989 (1977)). *Semantyka 2*. Warszawa.
- [15] Miller, G., Johnson-Laird, Ph. (1976). *Language and Perception*. Cambridge – London – Melbourne.
- [16] Pauliny, E. (1943). *Struktura slovenskeho slovesa*. Bratislava.
- [17] Petri, C.A. (1962). Fundamentals of the theory of asynchronous information flow. In *Proc. of IFIP'62 Congress*, Amsterdam. North Holland Publ. Comp.
- [18] Rasiowa, H. (1975). *Wstęp do matematyki współczesnej*. Warszawa.
- [19] Russell, B. (1967). Denotowanie, deskrypcje. In *Logika i język*, pages 259–293, Warszawa.
- [20] Selinker, L. (1972). Interlanguage. *Iral*, 10(3):209–213.
- [21] Szulc, A. (1984). *Podręczny słownik językoznawstwa stosowanego*. Warszawa.
- [22] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris.
- [23] Weinsberg, A. (1983). *Językoznawstwo ogólne*. Warszawa.
- [24] Zabrocki, L. (1970). Uwagi o problemach spornych gramatyki kontrastywnej (konfrontatywnej). *Lingua Posnaniensis*, XIX:2–23.

Antoni Mazurkiewicz

Institute of Computer Science, Polish Academy of Sciences, Poland

FORMAL DESCRIPTION OF TEMPORALITY (PETRI NET APPROACH)

Abstract. In the paper two methods of representing temporal dependencies expressed in natural languages are given. The first one has been introduced by H. Reichenbach in 1948 is based on a linear representation of time, with events represented by points on the time scale. The second one is based on a net representation of states, events, and their succession introduced by C. A. Petri in 1962. The main difference between these two approaches consists in accepting by nets (i) partial ordering of events and states rather than their linear ordering, hence accepting their mutual independence, and (ii) a possibility of representing coexisting as well as mutually excluding states, hence accepting different histories in one model. Reichenbach's representation can be viewed as a particular case of the Petri net one. Both representations use graphical means for modeling temporal phenomena.

Keywords: States, events, ordering, temporality, histories, Petri nets.

1 Temporality description issues

The main difficulty in proper translation temporal and modal phrases (expressions) consist in an imprecise description of situations expressed. Additional difficulty is caused by the fact that different languages exhibit a variety of different means used for the same situations and there is a great variety of temporal situations expressible directly in one language and not expressible in another. Clearly, a faithful translation of temporal phrases is of a primary importance, hence there is an urgent need for a background of strict and reliable temporality description. The main issue discussed in this paper is how to describe temporality dependencies and which are necessary means to grasp and express a given temporal situation. There are several possible approaches to this questions.

- a. Explaining temporal situations formulated in a language by their detailed descriptions expressed in the same language (self-explaining)
- b. Expressing temporality by equivalence of phrases in different languages (the question: "what does it mean" is replaced by "how it is expressed in another language"). Then temporality is an abstraction induced by all temporarily equivalent phrases.

- c. Expressing temporality by an *inter-language* (an “in between” language), with meaning of temporality supposed to be known
- d. Create formal models of temporal situations and then compare how they are described in different languages.

In the present paper the last approach is discussed. The aim of the paper is to present formal models of temporality, creating background of understanding the temporal situations and meeting the following requirements.

- **Directedness.** The scope of the description possibilities should be limited to temporal situations only. The intention is to get rid of unnecessary linguistic phenomena that can obscure the essence of temporality features.
- **Completeness.** The required model should cover all possible temporal (and modal) situations, leaving no room for imprecise and intuitive interpretations or for a necessity of relying on some hidden assumptions.
- **Independency.** The model description language should not use the linguistic temporal means specific for different languages; instead, it should have its own formalism of description, not relying on specifications introduced by existing natural languages.
- **Simplicity.** The model structure should be simple enough to guarantee its proper understanding by languages users.
- **Applicability.** The model should be easy to be applied for possibly large number of situations that can be described in natural languages. Some temporal situations that can be distinguished in some natural languages may be not such in other languages; therefore, the required models should be capable to cover all of them.

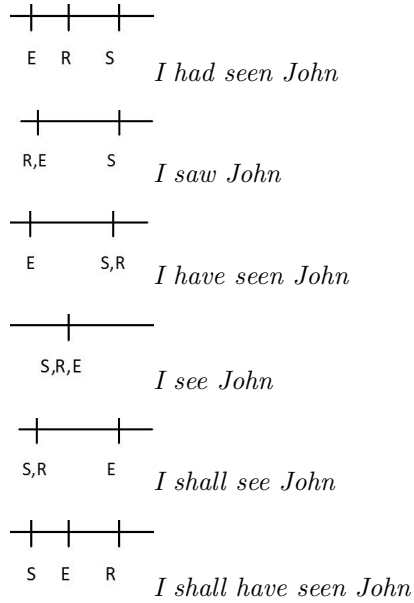
Below, two such formal models are presented and compared. Both of them use graphical representations of temporality. The first one is so-called *Reichenbach’s model*, formulated in his book *Elements of Symbolic Logic*, edited in New York, 1944, [3]; the second, so-called *net model*, called also *Petri net model*, formulated by Carl A. Petri [2] and described in [1].

2 Reichenbach’s representation of temporality

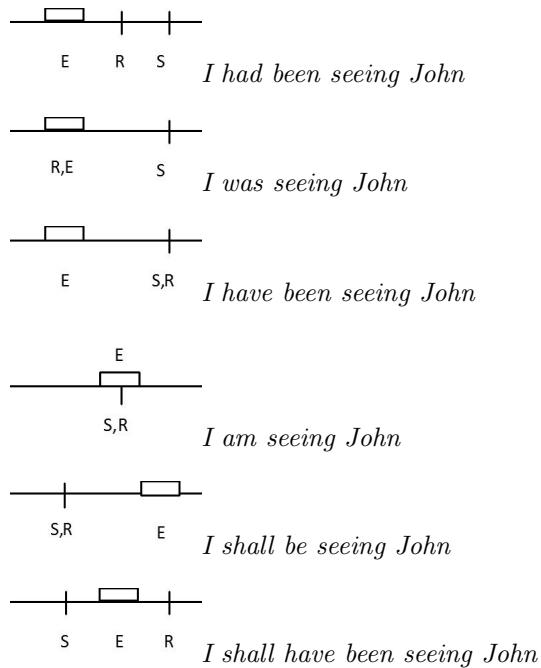
Reichenbach’s model of temporality phenomena was the first graphical representation of temporal relationships between a speaking subject and described objects occurring in sentences in natural languages. The basis for this model is a straight line, representing time scale, running from the left to the right, and some points on it, representing moments occurring in the reality described by the model of analyzed sentence. Among them, three points are distinguished: the point of utterance (corresponding to the moment of speech), the point of event (corresponding to the moment the statement is referring to), and the point of reference (the moment to which all other moments of the described situation are referred to).

The following Reichenbach’s schemata can serve as examples of using his model for explaining some temporal dependencies. Let consider simple variations based

on phrase “to see John”, expressed in different moments and referring to different moments. According to [3], we have the following descriptions:



For completeness, Reichenbach introduces an additional graphical symbol to indicate a time duration of some events, as it is shown below:



Actually, the diagrams shown above are not radically different from those given in the previous figure; the only difference is that point of event (E) is not a moment, but it is a period of time. In such a way it introduces a continuity of events, expressed by continuous tenses of English.

The Reichenbach's model can be summarized by the following table (remembering that events may be points as well as segments of a time line):

Time ordering	Tense
$\{e, r, s\}$	Present
$e \rightarrow \{r, s\}$	Present perfect
$\{s, r\} \rightarrow e$	Simple future
$s \rightarrow \{r, e\}$	
$s \rightarrow e \rightarrow r$	Future perfect
$\{s, e\} \rightarrow r$	
$e \rightarrow s \rightarrow r$	
$\{e, r\} \rightarrow s$	Simple past
$e \rightarrow r \rightarrow s$	Past perfect

3 Petri net approach

Reichenbach's temporal scheme is perfectly suited for English grammatical tenses description. For more general purposes, as e.g. for comparing grammatical means in different languages, more general framework is needed. Such a method for describing real temporal schemes on a very basic level, precisely defined in a formal (natural language independent) way has been formulated by C.A Petri [2] in 1962. Since then it is widely used for many purposes and interpreted in a number of different ways. From the name of its author, models using formalism introduced by Petri, are called Petri Nets. For linguistic purposes, the model presented below seems to fulfill requirements of temporality description. It is built from the following basic notions:

states, situations expressed in sentences, directly or indirectly

events, initiating or terminating states,

succession, a relation binding events with states initiated or terminated by them, represented graphically by circles, boxes, and arrows:

State: \bigcirc

Event: \square

Succession: \longrightarrow

States, describing situations we are talking about in everyday language, are properties of objects (or collections of objects). Examples of states are: “*the door is open*” or “*the door is closed*”. The characteristic feature of states is their extension in time – a state is a property holding during some amount of time. States can be

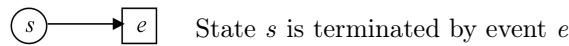
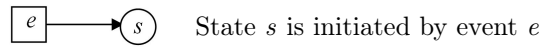
permanent, without beginning or ending (then lasting infinitely long), or temporary, lasting a finite amount of time.

Events are changes of situations; as such, they are momentary, taking no time – they can only occur in some moments. Example of an event is a change the state of the door from “open” to “closed” The characteristic feature of any event is that it happens either in the past or in the future with respect. to any chosen moment (saying “*e* is happening” we have in mind a collection of events, not a single one).

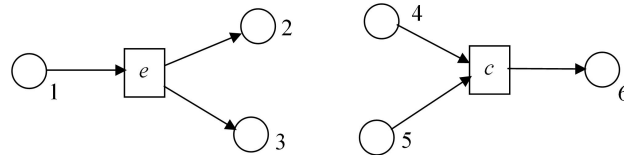
Succession is a relation between states and events establishing which events initiate (or terminate) which states. This relation determines a flow of time in the model in such a way that the beginning of any state always precedes (in time) its ending. The succession can be also treated as a causal relationship between elements described by nets.

States and events are fundamental concepts of the Petri nets theory; their causal (or temporal) ordering is the main issue discussed in terms of nets. In general, to describe real situations one needs a number of states and events, bound together by succession relation. Graphical representation of such a temporal scheme is a finite directed graph, with circles and boxes (as its nodes), joined with arrows (directed arcs). This graph is bipartite, i.e. any arrow leads either from a box to a circle or from a circle to a box; i.e. neither two boxes nor two circles are joined by an arrow.

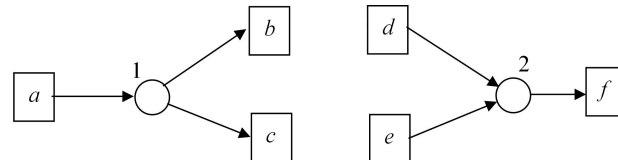
The basic constructs used in Petri nets are:



Nets arise by arbitrary combinations of the above constructs. Basic (and simple) combinations of them are:



Both diagrams describe actions *e* and *c*, the first action terminates state 1 and initiates two (coexistent) states 2 and 3, the second action terminates two (coexistent) states 4 and 5, and initiates state 6. The diagrams below describe other situations:



The first represents state 1 initiated by event a and terminated by exactly one of mutually excluded events, namely event b or event c , the second, state 2 initiated by exactly one of mutually excluded events d and e , and terminated by event f . In this way nets admit a possibility to deal with alternative actions. There is a similarity of states, events, and their mutual relationship to intervals, points, and their relationship on the number line. Namely, any event begins a state in the same way as a point starts an interval. Points begin some intervals which in turn are ending with some points. Any interval of the number line is terminated or initiated by a single point; on the other hand, any point can begin or end many intervals of the line.

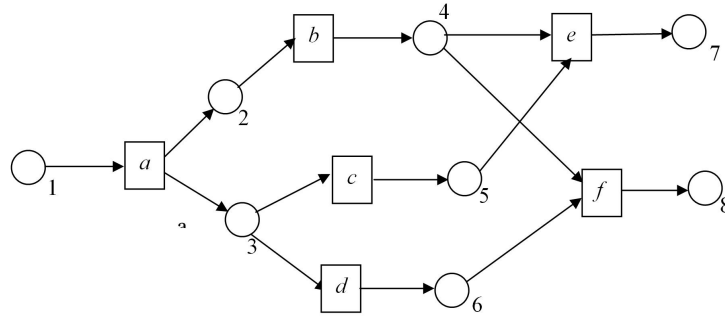
4 Histories

Nets describe temporal structures of pieces of reality, limited to states, events, and their mutual relationship. They can describe a single course of actions as well as a number of such courses, depending on different possibilities or conditions. Any specific course of actions defined within a net is a single *history* supported by the net. There can be one or more different histories supported by the same temporal scheme; histories can engage a part of the scheme or some repetition of its elements. In any case, any history must respect the succession relation between events and states defined by the net. For the time being, the temporal schemes without repetitions can be considered; nets with repetitions, i.e. containing cycles, will be discussed in the forthcoming parts of this study. For the needs of the present paper the following definition is sufficient: a history is a connected part of temporal scheme such that:

1. Any state of the history is initiated or terminated by at most one event of this history,
2. All states initiated or terminated by an event in the history belong to this history,

However, in a history a single event can initiate or terminate a number of states (as a single point on a number line can start or end a number of intervals).

In general, temporal structures can contain a number of different histories, representing various possibilities of the course of actions. It is reflected in the net model of such structures by presence of states ending (or beginning) by a number of events excluding each other. In a history singled out from such a structure, the termination or initiation of all its states is determined. In any history, two events can either precede each other, or can be independent –happening independently of each other; similarly, two states can either precede each other (i.e. the ending of one of them precedes beginning of the other), or can be coexistent – both of them exist at some interval of time. The best way to explain these relationships is to discuss an example of a complex temporal scheme. The diagram given below represents a net containing more than one history. It consists of 8 states, 6 events, and 15 arrows representing the succession relation.

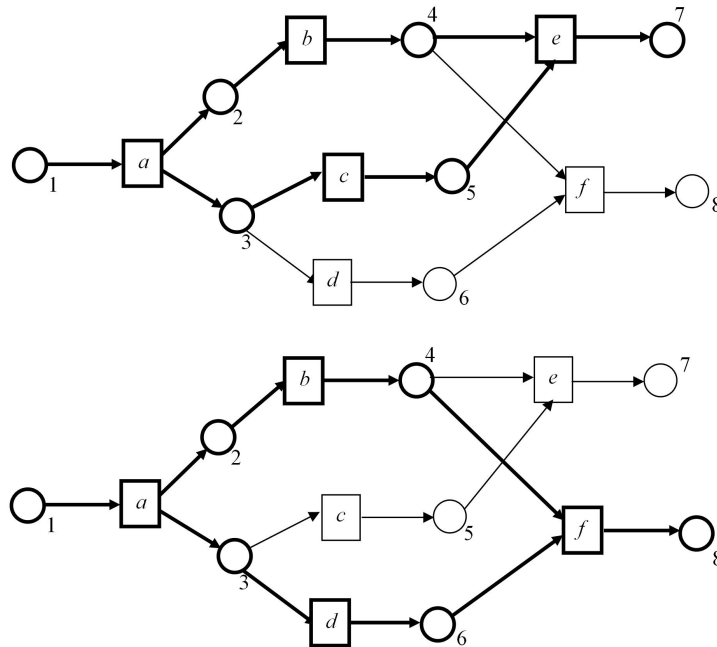


Event *a* terminates state 1 and starts two states, 2 and 3. States 2 and 3 are coexistent, as started with the same event. State 3 can be ended by two events *d* and *e* (mutually excluding each other). State 2 is terminated by event *b* initiating in turn state 4. If 3 is terminated by *c*, state 5 begins to exist, as initiated by *c*; otherwise (if 3 is terminated by *d*) state 6 begins to exist. States 4 and 5 are coexistent and terminated by event *e*. Similarly, states 4 and 6 are coexistent and terminated by a common event *f*. Events *c* and *d* exclude each other; hence, states 5 and 6 are not coexistent, since they are initiated by mutually excluded events and then also exclude each other. However, state 4 is coexistent with 5 as well as with 6. Coexistent states 4 and 5 are closed by event *e*, and coexistent states 4 and 6 are terminated by event *f*. Events *e* and *f* are excluding each other, since state 4 can be terminated by exactly one event: either *e* or *f*, but not by both of them. Consequently, states 7 and 8 are not coexistent.

This is an abstract explanation of the above net structure. To be more specific, assign the following meaning to states and event of the presented net. Namely, interpret it as a (fragment of) a real life reviewing procedure of a paper submitted for a publication, with states and events explained given in the tables below.

States		Events	
1	Preparing paper	a	End of preparing paper
2	Paper is ready	b	Start waiting for opinion
3	Preparing evaluation	c	Taking positive decision
4	Waiting for opinion	d	Taking negative decision
5	Opinion is positive	e	Sending paper to publisher
6	Opinion is negative	f	Rejecting the paper
7	Expecting publication		
8	Thinking about corrections		

Two diagrams below represents two possible histories contained in the above scheme. The first one is “optimistic”, resulting the paper acceptance, the second one – pessimistic, ending with rejecting the paper.



Analysis of the above histories in terms of the proposed interpretation is left for the interested reader.

Such an annotated net is a description of a temporal situation in a way independent of linguistic properties as well as of peculiarities of different languages.

5 State of utterance

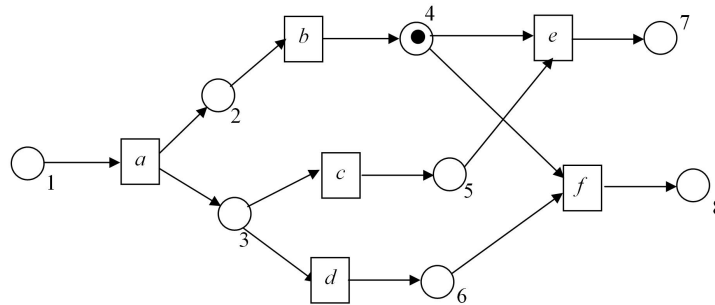
The main objective of net schemes presented here is a description of temporal properties of phrases in a similar way to Reichenbach's line sketched above. To explain a phrase expressing a temporal situation, one has to know objects (states and events) the phrase is referring to, and a state of an utterance subject. Their mutual combination determines linguistic means adequate to the described situation. The proper understanding of the phrase describing a given situation depends on proper choice of its net representation. Once the situation is characterized by a net and the state of utterance is given, phrases of different languages expressing this situation can be compared and analyzed, using the net scheme as a bridge joining different formal means specific for compared languages.

To this end, one has to construct a net comprising objects of utterance, i.e. to assign objects to states or events of the net, and to chose a state of utterance. Then the temporal meaning of the analyzed phrase is completely defined. Placing in a net scheme the state of utterance (i.e. choosing a state of the scheme where the speaker is situated) has an essential influence on the grammatical form of the analyzed sentence, similarly as it has been done using Reichenbach's schemata. In the net scheme, placing the state of utterance can distinguish actual history from any other

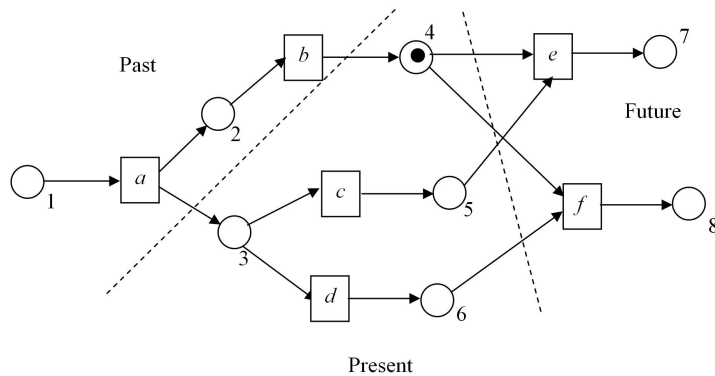
which is impossible in the accepted course of action. Namely, by introducing the state of utterance, the net scheme is split into

1. the present, past, and future of the history (histories), and
2. the possible and impossible objects of distinguished histories.

In graphical representation used here the place representing the state of utterance is marked by a dot. Below some examples of the net representation of some temporal situations are given. Consider the net discussed above with 4 as the state of utterance.

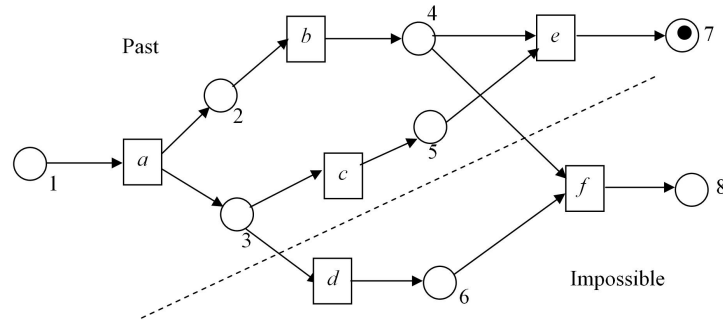


Then the whole net is partitioned into three parts:



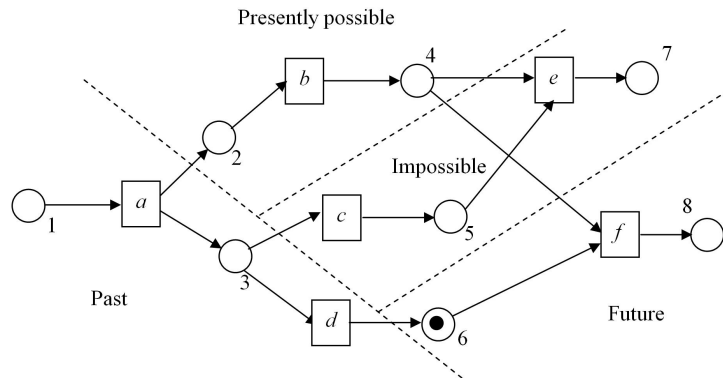
From the point of view of state 4 (the state of utterance) states 1, 2, and events *a*, *b* are in the past. Events *e* and *f* are in the future; the future is uncertain, since only one of the two can happen. It depends on the present, which is also uncertain; the speaker does not know whether state 3, or one of states 5,6 is holding now. Moreover, the speaker does not know which one of *c* and *d* will or already has happen. In other words, speaker at state 4 does not know which history is going on.

The next diagram shows the same structure, but with the state of utterance placed in state 7.



Then states 1, 2, 3, 4, 5 and events a , b , c , e are in the past, while states 6 and 8 will never take place, although they would be possible if d rather than c had been happen in the past.

Placing the state of utterance at place 6 we have the partition of the considered scheme into four parts: past, future, impossible, and possible at present:

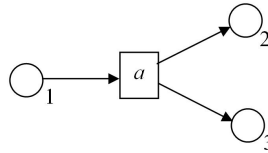


Term “presently possible” means here that, from the point of view of state 4 (of utterance), the speaker does not know whether it is now 2 or 4, but certainly one of them. About event b the speaker can be sure that either it has happen (and then 4 is now) or it has not happen yet (and, consequently, still is 2).

6 Enhancing nets

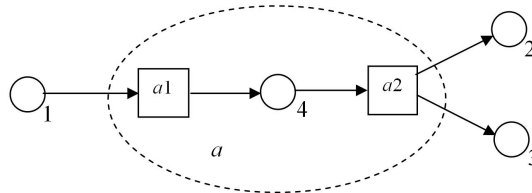
Two properties of net descriptions are worth to be mentioned. First, In order to a faithful description of situations, some additional event and states, not mentioned explicitly in their descriptions, should be inserted into the net. Inserting them into the net serves to proper sequencing of remaining states and events. Secondly, the net description can be made more or less precise, depending on the description purposes. Sometimes an event should be refined to more detailed structure, as it is shown in the following simple example. The net describes a very simple situation of a person (say John) who leaves his home and goes to his office. The following

states of John are taken into account: 1. John is at his home; 2. John is outside of his home; 3. John is on his way to his office. The single event binding the above states is leaving home (event a). Below the net describing the above situation is given.



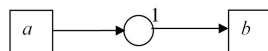
It says that the state “staying at home” (1) is terminated by action of leaving home (a) and two new states are initiated: “to be outside home” (2) and “to be on the way to John’s office”(3).

One can contest the qualification “leaving home” as a momentary event without any duration. Then one can refine the above scheme to the following:

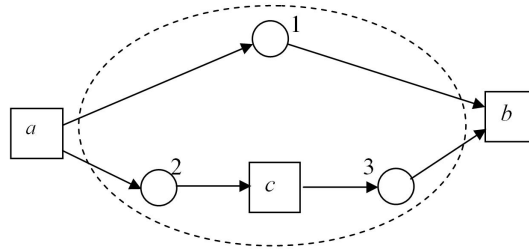


where event a (“leaving home”) has been split into two, more specific events, $a1$ (“begin of leaving home”) and $a2$ (“end of leaving home”) and a new state 4 (“action of leaving home”) has been introduced. In this way, “leaving home” lost attribute of being an event and became a state. It is a general situation: in order to be more specific, states and events can be refined, enhancing corresponding net. Thus, net descriptions can be enhanced by adding some new objects and by refining existent objects.

A similar possibility we have in case of states. Consider the scheme



State 1 can be viewed as e.g. “to be on holidays”, event a is the beginning of holidays, b is their end. To be more specific, one could state explicitly that the first part of holidays are to be spend at the Baltic sea (state 2), while the rest at home (state 3). Then enhanced net would look like that:



Event c in the above diagram separates states 2 and 3 and the nature of c is left unspecified. In effect of this transformation, the scheme has been enhanced by adding new elements.

Both transformations given above make nets more specific; they are called net *refinements*. There are possible transformations in the opposite direction, making nets less specific, to avoid some unnecessary details. Such transformations are called net *abstractions*. Both of them allows us to tailor net schemes exactly to the description needs.

7 Summary and conclusions

The aim of this paper is to present a Petri net formalism and to show how it can be used for defining temporal situations. Reichenbach's schemes have been thought to be used for the same purpose. It is clear that the net formalism covers the formalism of Reichenbach, treating points on a number line as events and intervals as states; however, Reichenbach's formalism does not cover independency of events, uncertainty of sequencing events and states, and various aspects of modality. Thus, the net formalism can be viewed as an essential extension of Reichenbach's one. The scope of this paper is limited to a presentation of the net formalism; it was not the intention of the present paper to analyze temporal situations from the linguistic point of view. Presentation of usage the net formalism to some specific linguistic phenomena is expected in forthcoming parts of this study.

Bibliography

- [1] Koseska-Toszewa, V., Mazurkiewicz, A. (1988) *Net representation of sentences in natural languages*, *Advances in Petri Nets*, LNCS 340, Springer Verlag: 249–259.
- [2] Petri, C. A. (1962). *Fundamentals of the Theory of Asynchronous Information Flow*, Proc. Of IFIP'62 Congress, North Holland Publ. Comp.: 249–259.
- [3] Reichenbach, H. (1944). *Elements of Symbolic Logic*, New York, 1944, McMillan Publ. Comp.
- [4] Reisig, W. (1985). *Petri Nets – An Introduction*, New York, Springer Verlag.

VIOLETTA KOSESKA¹, ANTONI MAZURKIEWICZ²

¹Institute of Slavic Studies, Polish Academy of Sciences, Poland

²Institute of Computer Science, Polish Academy of Sciences, Poland

NET-BASED DESCRIPTION OF MODALITY IN NATURAL LANGUAGE (ON THE EXAMPLE OF CONDITIONAL MODALITY)

Abstract. The intention of the present paper is to show how the Petri nets formalism can be applied for explaining not only temporal but also modal properties of sentences in natural languages. A special attention has been paid for distinguishing courses of actions with forking (that creates different, but coexistent courses) from branching (that creates different and mutually exclusive courses). It is argued that conditional sentences cannot be represented properly by means of logical implication; instead, for this representation the net description is proposed. Examples serve to show how Petri nets can be viewed as a universal tool (an intermediate language) for analyzing and comparing different natural languages.

Keywords: states, situations, events and histories, branchings and forks, implication and conditional, Petri nets.

1 Introduction

The semantic model of description of modality in a natural language can be based on the basic notions of Petri net theory, see [8] and [4]. The net-based representation of time and modality is a significant extension of Reichenbach's conception of tenses, and hence it is rather a generalization than negation of that conception. In other words, each temporal situation expressed using Reichenbach's schemata can be represented using nets, while not every situation expressible with nets can be represented in Reichenbach's model [8], [13].

In the net-based representation of an utterance, we talk about **states, situations, events and histories**. *Local states* represent certain momentary properties of objects being the subject of utterance; *global states* consist of states of all such objects. *Events* cause a change in the state of some object or several objects, which gives the net-based description a dynamic character, varying over time. The course of events expressed by an utterance forms a *history*, representing mutually interrelated dependence among states and events. In the net-based approach to description of such processes, the paradigm of a *state* is its continuance. Each states continues for some specific time. Two different states following one another are separated by

some *event* which begins the new state and ends the old one. The event, which represents a change, does not continue; it only occurs at a certain point of time. By a *situation* we shall mean here a certain fragment of reality which might encompass part of the past, the present and the future of the states of some objects. All utterances of a temporal character will refer to such situations.

Temporal utterances describe some situation, i.e. they talk about dependencies which appear in the temporal course of events and states. We will describe situations with help of Petri nets [...]. Analysis of an utterance must take into consideration the speaking subject's position with respect to the uttered situation. In Reichenbach's schemata, that position corresponds to a point on the timescale, in Petri nets it is a state of some object, namely the subject of utterance, from now on referred to as the observer. The position of the **observer** describing the situation will correspond to the so-called *moment of speech*. Due to the impreciseness of the latter term, in our description it has been replaced with the term "**state of utterance**". The state of utterance is a position occupied by the observer, i.e. the sender of information (or the *speaker* in terms of traditional grammar); hence the state of utterance determines all states possible in the present (the present situation), and indirectly also all states and events possible in the future and in the past. Knowing the (hypothetical or actual) course of the described events and states, we can draw conclusions regarding situations which take place in the discussed fragment of the changing reality.

It is worth adding that the above basic notions of the net-based description of time are also associated with other notions, which are important for the semantic interlanguage connected with the net theory:

Local state — state of certain special objects discussed in the utterance.

Global state — state of all objects determining the situation.

Accessible state — state reachable in the future with respect to the observer's situation.

A global state consists of the states of all objects in some situation, as opposed to a local state, which refers to one or several objects of that situation. For example, a local state for the objects "doors, windows" will be "the doors are closed", while "the doors are closed, and the windows are open" will be a global state. We can say that a global state is a special case of a local state: namely, it is local state that encompasses, as we have mentioned in the foregoing, all objects of the discussed situation, in opposition to a local state, which encompasses either one or some of them. 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; modal phenomena in a natural language reflect effects of the local character of states. This implies the need for the description methods to take into consideration the local character of states. According to the principles of net-based description, a given local state can be assigned a set of global states — namely, all the states which are compliant with that local state in the given fragment of the described reality.

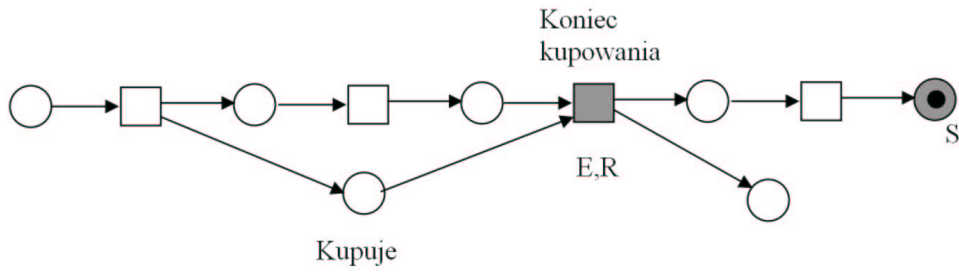


Fig. 2. Ona skończyła robić zakupy [She has finished doing the shopping]

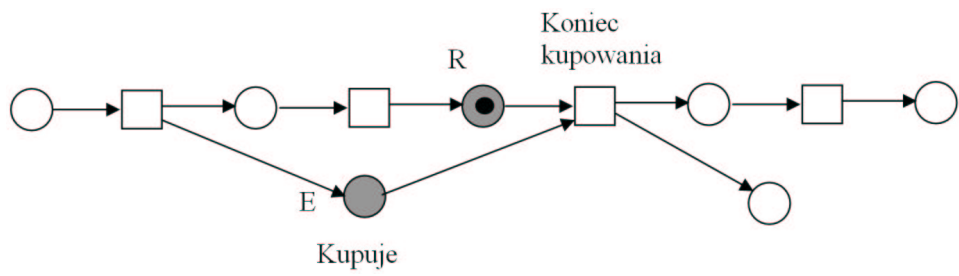


Fig. 3. Ona robi zakupy [She is doing the shopping]

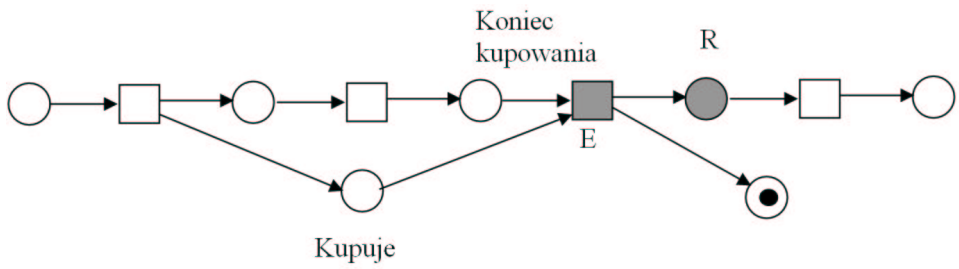


Fig. 4. Ona ma zrobione zakupy [She has the shopping done]

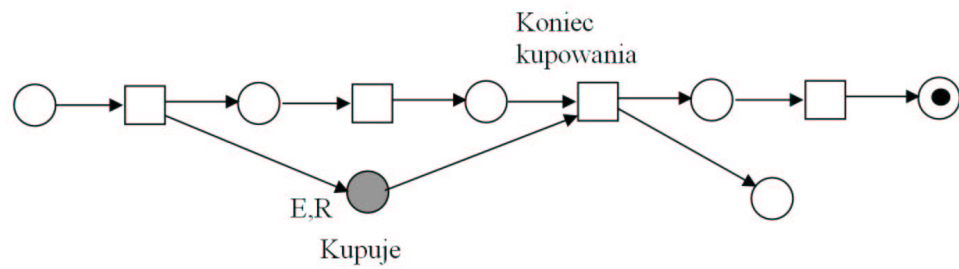


Fig. 5. Ona robiła zakupy [She was doing the shopping]

Ona robi zakupy is the Bulgarian *Тя назапува*. In that net, our speaking and her doing of the shopping are concurrent. In the fourth net, the analogue of the Polish *Ona ma zrobione zakupy* is the Bulgarian *Тя е напозарувала*. In that net, we refer to the state initiated by the event ending “kupowanie”, and we have there the perfectum of perfective verbs. In the fifth net, the analogue of the Polish *Ona robiła zakupy* is the Bulgarian *Тя е назапувала*. In that net, we refer to the state “kupowanie”, which was initiated before the state of utterance. We have there the perfectum of imperfective verbs. It should be noted that the first net contains an occurrence of Bulgarian aorist of perfective verbs used for denoting the event which occurred before the state of utterance. In the fourth net, in Bulgarian we have an occurrence of Bulgarian perfectum with the function of ascertainment. In the fifth net, perfectum has a resultative meaning; for details see Koseska in [6]. As the net description implies, in such nets two different elements can occur simultaneously, and hence two different states can coexist. Moreover, two events can be executable independently, and events can be executable during the existence of some state.

2 Branchings and forks

The descriptions and schemata given above referred to a single history, and the differences followed from different mutual positioning of the observer’s position, events and states. In reality, we sometimes speak about certain variants of the future and the past, and then the observer’s position must be positioned in some way with respect to these variants. In the net-based description, the description of such situations follows in a natural way from the specifics of the net-based description itself.

2.1 Branchings

A net may contain *branching*; if they exist, they represent different possibilities of its course. The branchings may be unconditional in the sense that the choice of one or another exit event is not conditional upon anything. To explain the idea of net-based description of temporal situations, below we present simple nets illustrating basic cause-effect relationships. In both net schemata (see Fig. 6, p. 70), there are three events, *a*, *b* and *c* connected with some common state 1. This state in Example (1) begins with event *a*, and ends with one of (mutually exclusive) events *b*, *c*. In Example (2), the state begins with one of (mutually exclusive) events *a*, *b*, and ends with event *c*.

Schema (1) expresses the situation where occurrence of event *a* is a necessary condition for occurrence of one of events *b*, *c* — without occurrence of *a*, occurrence of any of them is not possible. However, the schema does not imply that occurrence of *a* causes occurrence of *b*, because we can have a course of events where occurrence of *a* will cause occurrence of the event *c*, excluding *b*. Schema (2) describes a situation where occurrence of event *a* is a sufficient condition for occurrence of event *c*; the schema also implies that another such condition is occurrence of event *b*, with *a* and *b* being mutually exclusive. This is because we can have a

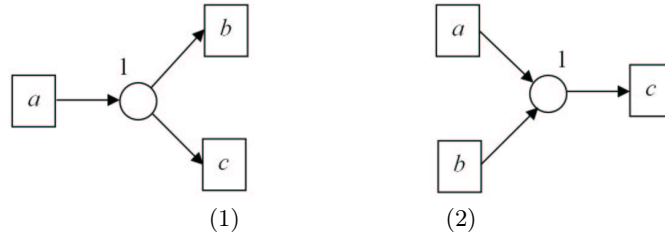


Fig. 6.

course of events where c is preceded by b , but event a does not occur, which means occurrence of event a is not a necessary condition for occurrence of event c .

2.2 Forks

Another schema of a situation, in some sense dual to branching, is the so-called *fork*, where one event starts (or ends) a number of co-existing states. Examples might include e.g. end of a railway journey, which ends both the state of travelling and the state of remaining in a railway car; or beginning of a sickness, which also begins the state of fever. The simplest examples of forks are presented in Schemata (3) and (4) (see Fig. 7). In Schema (3), event a ends state 1 and begins two coexisting states 2 and 3, which represent the beginnings of two independently running histories. In Schema (4), event a begins state 3 and ends two coexisting e states 1 and 2, which representing the final states of two independently running, but not mutually exclusive, histories.

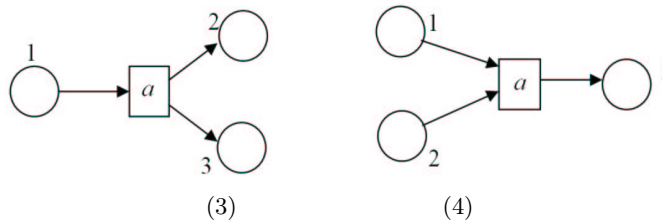


Fig. 7.

In the situation presented in Schema (3), states 2 and 3 are consequences of state 1; state 1 is a necessary condition for occurrence of any of states 2 and 3. In Schema (4), occurrence of both states 1, 2 is a necessary condition for occurrence of state 3; if any of them fails to occur, event a cannot occur either, and hence state 3 cannot begin.

Summing up, by a fork in the net we mean a *situation* where one event can begin or end more than one state, and is characteristic for a parallel course of one or more components of the system, not colliding with each other.

By a branching in the net we mean a *situation* where one state can begin or end with more than one event. A branching is a characteristic feature of nets describing a situation with the possibility modality and corresponds to what has been characterized in logic using the functor “it is possible that”. A branching represents a choice among a few mutually exclusive possibilities. A branching and a fork in the net are the main sources of the possibility modality in natural language sentences. Two or more events directly following some state are a phenomenon characteristic for a branching, while two or more states following one event are characteristic for a fork.

In the above examples, the states of utterance have not been marked, because the cause-effects relationships do not depend there on one or another position of that state [7].

3 The possibility modality

The problems connected with sentences involving the possibility modality require pointing out a certain fact which is important for understanding the essence of the theory we are using. Understanding “conflicts”, or “branchings”, as a synonym of “negation” would be a gross misunderstanding — and we have met with such remarks in the discussions among linguists over the net-based description method and its applications to studies of modality in a natural language. The essence of **conflict** is choice between two (or more) mutually exclusive possibilities, while negation is a logical functor, and as such has a static character; however, resolution of a conflict (choice) has a dynamic character (like each net-based description), and affects the subsequent course of things. This remark applies especially to the description of conditionality in a natural language.

In case of a conflict, the history presented with the net can run in different ways, depending on the circumstances or the choice made. Such a choice determines one of a few possible, but mutually exclusive, continuations of action. Using a net, we can describe future and past consequences of choices already made in the past or those which can be made in the future. The possibility of a branching in the net offers means for describing conditional sentences, see [8].

Most of the works describing the semantic structure of conditional sentences base it on the so-called conditional, and deliberate on the relationship between the conditional and logical implication, see: [15]. The problems of implication and conditional have been the subject of numerous papers and discussions in logic, which concentrated mainly on the so-called implication paradox. The problem emerged when implication was read using the words *if...then*, see [14]. In a natural language, the above expression as a rule has a broad range of different meanings. The problem reduces to the question whether logical implication can be “read” as the conjunction *if...then*.

3.1 Implication, conditional, or something else?

Most of the authors involved in the discussion agree that the truth of a natural language sentence does not necessarily depend on the truth of the succedent and

the antecedent, as is the case for logical implication. Ajdukiewicz points out the fact that in each natural language the meaning of the conjunction *if...then* is only close to, but different from the meaning associated by contemporary logic with the notion of implication, and that certain natural language sentences become false after replacing the implication symbol with the conjunction *if...then*. This refers to the theorems, writes Ajdukiewicz, which are connected with the fact that implication is true whenever either its antecedent is false or its succedent is true. This is because in logic implication is false only if the antecedent is true, and the succedent — false [1].

The divergences between the meaning of implication and the meaning of a conditional from a natural language gave rise to the question whether it is appropriate to analyse reasoning in a natural language using the notion of implication. The problem of the implication paradox led to disputes on the limits of applicability of logical methods to natural language studies [11]. In connection with difficulties in interpreting a sentence of the type : *If p, then q*, the linguistic literature on that subject also put forward a question whether it is appropriate to analyse reasoning in a natural language using the notion of implication [3], [2]. For example, A. Bogusławski refuses to equal conditional sentences with either material or strict implication in any sense or mode [3].

It is worth stressing that most of the scholars acknowledges that there is a “dynamic relationship” between p and q , i.e. the two components of a conditional. The most appropriate approach in the net-based description seems to be representation of the conditional “if – then” by the cause-effect relationship (Petri,62). Subject to the reservation that we do not list all meanings of *if...then*, “and indirectly connections between p and q , i.e. between events or states of things referred to, respectively, by conditional sentence p , or antecedent of the conditional, and by the main clause q , or its succedent”, Pelc lists several ways of interpreting the conditional *if p, then q*, see for example:

1. Causal relationship, e.g. *If you eat too many carbohydrates, then you will grow fat*;
2. Sign relationship, e.g. *If he has rash, then he is sick with scarlet fever*.
3. Special cases of a universal relationship, one of which is formal logical implication [15, p. 272].

The **quantifier *only*** occurs in the semantic structure of an expression rather than in its surface structure, where it can be “incomplete”. In our opinion, in a natural language we do not have to do with the expression *if p, then q*, but rather with an expression with the meaning: *only if p, then q* or *if only p, then q*. A proof for the fact that the “only” quantifier always occurs in the semantic structure of such a natural language sentence, though it does not necessarily appear in the surface structure of that type of sentence, is negation of the logical expression *if p, then q*, see the sentence: “*If it’s raining, I’ll take an umbrella*” and its logical negation “*it is raining and I won’t take an umbrella*”. The truth of the above implication is guaranteed by the following situations:

- (a) It's raining and I'm taking an umbrella
- (b) It isn't raining and I'm not taking an umbrella
- (c) It isn't raining and I'm taking an umbrella.

When in a natural language we say “*If it's raining, then I'll take an umbrella*”, our intention is to exclude possibility (c), i.e. we really have in mind the formulation “Only if it rains, then I'll take an umbrella”. This sentence uttered in a natural language is true in the following situations:

- (a) It's raining and I'm taking an umbrella,
- (b) It isn't raining and I'm not taking an umbrella,

i.e. it describes the logical equivalence “p if and only if, when q”. Let us note that, formally, “I'll take an umbrella only if it's raining” means the same as “if I take an umbrella, then it's raining”, and it is true in the following situations:

- (a) I'll take an umbrella and it's raining,
- (b) I won't take an umbrella and it's raining,
- (c) I won't take an umbrella and it isn't raining.

The (logically) correct equivalent formulation of “I'll take an umbrella only if it's raining” is “I'll take an umbrella when it's raining and only when it's raining”. In reality, it is not the speaker's intention to talk about the situation of somebody who can see that it isn't raining, that the sun is shining, and who is nevertheless taking an umbrella. Accordingly, the natural situation is completed by the sentence: *It isn't raining and I'm not taking an umbrella*. However, it is also a negation of the semantic structure of the sentence: *if I take an umbrella, then it's raining*, i.e. *I'll take an umbrella only if it's raining*. The latter sentence can occur without the surface unique quantifier: *only*, see the sentence: “If it rains, I'll take an umbrella”, when by default we have “*only*”.

In Petri nets, a **history** describes a sequence of transformations of states through occurrence of events; in each history, the relation between states and events is a cause-effect relation [13]. Without defining precisely, what we understand by cause and effect, we can assume that we know how to understand a cause-effect relation. It is a temporal relation. The states representing a condition for event occurrence (appearing “before” the event) represent its **cause**, and the states following from them (appearing “after” the event) are their **effect**. In the net theory, connections of states and events are underlain by a relationship corresponding to the cause-effect relation rather than to the notion of implication. The cause for occurrence of some event in a Petri net is the occurrence of all states constituting the causes of that event, and the effect of an event is the occurrence of all states constituting the effects of that event, see schemas (3) and (4) concerning a necessary condition and a sufficient condition [12].

Example. Let us examine the sentence: *If it rains tomorrow, I'll take an umbrella*, and the net corresponding to that sentence, presented in the figure below. The example is rather expanded due to further considerations, which are outside the

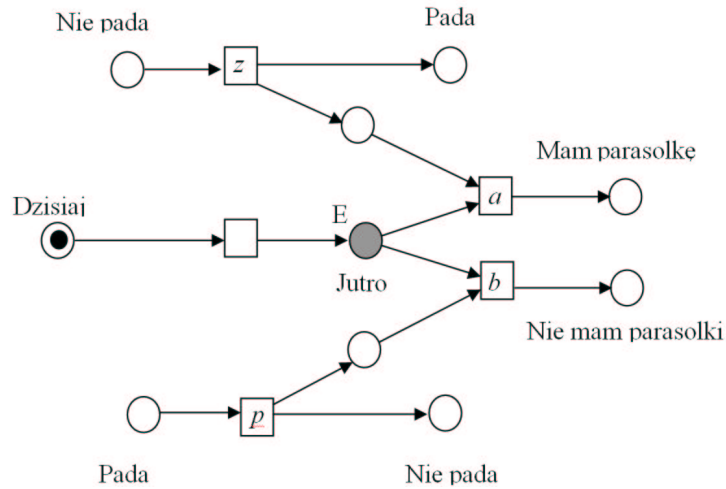


Fig. 8.

scope of the present problem area; it will then serve as an explanation for other language phenomena too.

Legend:

- Nie pada = It isn't raining
- Pada = It's raining
- Dzisiaj = Today
- Jutro = Tomorrow
- Nie mam parasolki = I don't have any umbrella
- Mam parasolkę = I have an umbrella

In the figure we have a net describing a situation which explains the meaning of the utterance *If it rains tomorrow, I'll take an umbrella*. The net describes two mutually exclusive histories:

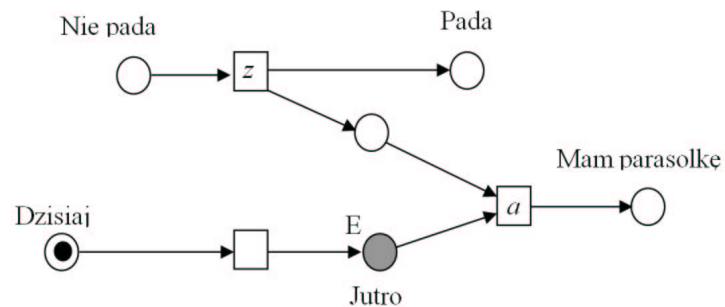


Fig. 9. History 1 ("It's raining")

and

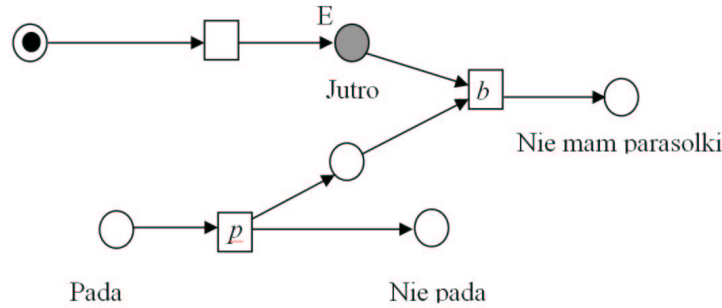


Fig. 10. History 2 (“It isn’t raining”)

In the net we have marked the state of utterance (“today”), the state of decision-making (E), as well as events z and p (it starts and stops raining), and a and b (I’m taking and not taking an umbrella). Moreover, the states “it’s raining” and “it isn’t raining”, which according to the laws of logic exclude each other, are also marked; we assume that the states of raining and not raining change cyclically (alternate). Let us note that occurrence of event z (it starts raining) as well as of event p (it stops raining) is independent of the state of utterance: it can occur either before or after, or else during the state of utterance. The decision on taking an umbrella is made under the influence of those events; if event z occurs in the history, then I have an umbrella and it’s raining; if p occurs, then I don’t have an umbrella and it isn’t raining. Moreover, let us note that the change of the state *dziś* (today) to the state *jutro* (tomorrow) is effected by an event independent of the states *pada* (it’s raining) and *nie pada* (it isn’t raining). The above net shows both states and events appearing explicitly and those appearing implicitly, as well as possible relations among them. It is, in our opinion, a good representation of the semantic structure of the conditional sentence: *If it rains tomorrow, I’ll take an umbrella.*

Nets describe transformations of states by events and their mutual relations, determined in the net theory by the cause-effect relation. The cause-effect relation is always a temporal one. Hence the net-based description of conditionality allow us to use the notions of state, event and cause-effect relationships. The net-based interpretation of conditionality refers to previous states, previous events, as well as states and events following the former as a result of the cause-effect relation. Logical implication is an indispensable tool of formal deduction, leading always from true premises to true conclusions, but it says nothing about cause-effect dependencies — and they are exactly what we have to do with in the conditional sentences of a natural language.

4 Nets and interlanguage

An interlanguage, necessary for juxtaposing different languages, in particular Polish and Bulgarian, within the time and modality categories is a semantic tool, see [4]. Petri Nets, through their universality and independence of natural languages, are a perfect candidate for an interlanguage. This theoretical tool reveals language phenomena sometimes overlooked by linguists. Proponents of the net-based description discover on the example of conditional modality more and more of new possibilities provided by the net-based description of natural language, which often fundamentally diverge from the tradition. The net-based description of modality and time allows us to capture the most important semantic features of various types of modality, such as conditionality, imperceptiveness, or “hypotheticality”. In this paper, we have captured conditionality through the constructions of net branching and the cause-effect law, connecting states and events. The modal and temporal problems concerning conditionality discussed here are a novelty with respect to the problems connected with conditionality already discussed in the Polish-Bulgarian contrastive grammar, see [7].

Bibliography

- [1] Ajdukiewicz, K. (1956). Okres warunkowy a implikacja materialna, In *Język i poznanie*, t. 2. *Wybór pism z lat 1945–1963*, Warszawa.
- [2] Banyś, W. (1989). *Theorie semantique et SI...ALORS. Aspects semantico-logiques de la proposition conditionnelle*.
- [3] Bogusławski, A. (1986). Analiza zdań warunkowych a problem funkcji semiotycznych, In *Studia semiotyczne XIV–XV*, Warszawa, pages 215–224.
- [4] Bulgarian-Polish Contrastive Grammar (1990–2007). Sofia – Warszawa.
- [5] Косеска-Тошева, В., Гаргов, Г. (1990). *Семантичната категория определеност/неопределеност, Българско-полска съпоставителна граматика*, том 2, София.
- [6] Koseska-Toszewa, V., Korytkowska, M., Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Wydawnictwo Akademickie Dialog. Warszawa.
- [7] Koseska-Toszewa, V., Maldzieva, V., Penchev, J. (1995). *Gramatyka konfrontatywna bułgarsko-polska, Tom 6. Część 1. Modalność. Teoretyczne problemy opisu*, Warszawa.
- [8] Koseska-Toszewa, V., Mazurkiewicz A. (1988). Net Representation of Sentences in Natural Languages. In *Lecture Notes in Computer Science 340. Advances in Petri Nets*, Springer-Verlag, 249–266.
- [9] Korytkowska, M., Roszko, R. (1997). *Gramatyka konfrontatywna bułgarsko-polska. tom 6. Część 2. Modalność imperceptywna*, Warszawa.
- [10] Korytkowska, M., Roszko, R. (2006). *Gramatyka konfrontatywna bułgarsko-polska. Tom 7. Semantyczna kategoria czasu*, SOW, Warszawa.
- [11] Kotarbiński, T. (1957). *Wykłady z dziejów logiki*, Łódź.
- [12] Mazurkiewicz, A., Koseska, V. (1991). Sieciowe przedstawienie temporalności i modalności w zdaniach języka naturalnego. In *Studia gramatyczne bułgarsko-polskie, t. IV, Modalność a inne kategorie językowe*, pages 7–25, Warszawa.

- [13] Mazurkiewicz, A. (1986). Zdarzenia i stany: elementy temporalności. In *Studia gramatyczne bułgarsko-polskie. Tom I, Temporalność*, pages 7–21, Wrocław.
- [14] Quine, W. (1955). *Mathematical Logic*, Cambridge, Mass.
- [15] Pelc, J. (1986). Jeżeli, to. In *Studia semiotyczne XIV–XV*, pages 271–287, Wrocław.
- [16] Petri, C. A. (1962). Fundamentals of the Theory of Asynchronous Information Flow. In *Proc. of IFIP'62 Congress*, North Holland Publ. Comp., Amsterdam.

KATARZYNA DROŹDŹ-ŁUSZCZYK, ZOFIA ZARON

Institute of Applied Polish Language Studies, Warsaw University, Poland

SEMANTIC INTERRELATIONS BETWEEN THE WORDS *MISTRZ* AND *UCZEŃ*

Abstract. The paper presents the results of research into relation between the lexemes *mistrz* (master) and *uczeń* (apprentice or student).

Some phrases (*mój mistrz*, *mój uczeń*) and syntactic features suggest a kind of convertive relation. Both of the lexemes are — for example — a three-actant: *ktoś*, *czyjś*, *w jakiejś dziedzinie wiedzy (/sztuki)* (someone, someone's, in a field of science (/art)). This feature is not sufficient to make a convertive relation. In Polish Language the *mistrz-uczeń* relation is a unilateral one. The relation depends on the apprentice who selects a master. The master confirms this only if *uczeń* (apprentice) is taught by the master and if the master is regarded by the apprentice as a skilled, knowledgeable person.

Keywords: master, apprentice, student, relationship, meaning.

The issues presented in the paper were collected during the research on the project *Synchronic and Diachronic Research on Contemporary Proper Names*. The project has been carried out since December 2008 by an eight person team including prof. St. Dubisz, prof. Z. Zaron, dr J. Chojak, dr K. Dróżdź-Łuszczczyk, dr P. Sobotka, dr M. Stępień, C. Heliasz (MA), M. Horodeńska-Ostaszewska (MA). The aim of the research is to create a multi-dimensional study of common proper names, therefore the main paper is composed of the following parts:

1. semantic one in which there are mainly explanations of meanings corresponding to real meaning definitions; it also deals with semantic relationships (synonyms, conversions, antonyms) and information on lexemes creating lexical nest (i.e.: words being in word formation relationship with the lexeme in question);
2. inflectional one which contains information about the inflectional gender (gender classes were adopted after Z. Zaron [10]) and, perhaps, on alternations in stem endings;
3. syntactical one which is composed of information on: (i) syntactical requirements (valence) of the lexemes in question, (ii) syntactical agreements with adjectives and verbs, and (iii) collocations and information on the limits in connectivity with the described units;

4. paradigmatic one which gives information on the stylistic marking of a lexeme and its classification in a certain variety of language and also on homonymy in other language varieties.
5. etymological one which gives not only the information on the origin of a lexeme and its semantic changes but also information on inflectional changes or the stylistic register. We will also point out to a dictionary which contains the given lexeme as the first one.

The final result of the research will be lexicographical (mini)portraits of proper names.

Our text came into being in relation with the research we have been carrying out, therefore to a certain extent its co-authors are also all the above-mentioned participants of the project. The analysis referring to *mistrz* (master) and *uczeń* (apprentice, student) has been carried out by the authors of the present text.

* * *

In the present paper we have decided to present word portraits not even in the understanding of Apresyan but mainly to show the interrelations between the lexemes *mistrz* and *uczeń*. These interrelations have been intriguing philosophers (Scheller, Tischner) and sociologists (Czapigo). The latter ones show a strong tendency for posing dramatic questions, e.g.: does the master-apprentice/student relationship nowadays, in the times when role-models are questioned, make any sense whatsoever. Our response to that is tackling this problem in the present paper.

If one should want to discuss the interrelations between the words *uczeń* and *mistrz*, one should also make it clear that the relationship does not hold for all meanings of the words in question given by dictionaries of Polish. And there is a great deal of them available. *Uniwersalny słownik języka polskiego* (Universal Dictionary of Polish) [5] gives as many as five meanings for the word *mistrz* (master) and three for the word *uczeń*¹ (apprentice/student). Without disagreeing with the authors of the dictionary on the rightness of their decisions (as this is not an aim of this text), let us focus on the meanings in question. According to USJP the word *mistrz* denotes (cf. K-Ó, p.677):

- /1/ a person excelling in certain skills, with an expertise in a certain field — cf. *mistrz pióra, akwareli* (a master of writing, a master of watercolour) (pointing to the field of arts by pointing the object), *mistrz w sztuce kulinarnej* (a master of culinary arts), *mistrz propagandy* (a master of propaganda);
- /2/ a person regarded by others as a guide, a paragon in a certain field of knowledge — cf. *mój mistrz, mistrz duchowy* (my master, spiritual master);
- /3/ a person who turned out to be the best, by winning in a sports competition of great importance (or the title given to such a person) — cf. *mistrz w bie-gach przez płotki, mistrz szachowy, mistrz Europy* (a hurdles champion, a chess champion, a European Champion)

¹ ISJP for the word *mistrz* gives twice as many entries (10), cf. [2: I A. ... Ó:273]; for the word *uczeń* it gives two [2: II P. ... Ż:887].

- /4/ a person who holds a diploma entitling to run a workshop — cf. *mistrz kamieniarski*, *mistrzowie cechowi* (a master mason, guild masters) (in my opinion outdated);
- /5/ a person who has qualifications for supervising the work of subordinate workers — cf. *mistrz murarski*, *mistrz zmiany* (master bricklayer, shift master) (still present in the labour code in the register of occupations).

The word *uczeń* denotes according to USJP [5: T-Ż:200] a person who:

- /1/ learns usually at primary or secondary school, in other words a person for who learning is compulsory, cf. *uczeń pierwszej klasy*, *gimnazjum*; (a first grade *gimnazjum* student);
- /2/ does apprenticeship in a certain profession or learns it (not necessarily from one teacher); in the case of *uczeń₂* the most important aspect is acquiring practical skills in a certain area;
- /3/ a person who is a supporter and / or a follower of someone's teaching or artistic work (por. *uczeń wielkiego filozofa* (a student of a great philosopher)).

As one can see by briefly examining the above-mentioned definitions not every *mistrz* (master) noted by USJP must have an apprentice or a student (*uczeń*), and similarly not every *uczeń* must have a master (*mistrz*). When one uses the phrase *mistrz kamieniarski* (master mason) or *mistrz zmiany* (shift master) there is no need to determine the existence of an apprentice (*uczeń*) or apprentices. Similarly syntactic and semantic factors do not result in understanding *mistrz₃* as someone who must have a student or apprentice (*uczeń*); what is more important here, is the fact the word denotes that with the use of objective criteria one person has been chosen from a group of people as the best in a certain area. When one uses *mistrz₁* the information that someone is a master of painting (*mistrz pędzla*) does not imply the statement that the person has an apprentice or a student, and what is more it does not also imply that there are other people dealing with this area. However, it is obviously not excluded. In the case of *mistrz₁* what matters is his or her skills displayed in broadly understood artistic output. It is quite contrary in the case of *mistrz₂* (further referred to in the paper as simply *mistrz*); what is important in here is this person's knowledge and the fact that the person shares it with someone else who receives it. *Mistrz₂* will always be someone else's master or a master for someone and according to someone. However, this someone will not be every *uczeń* given by USJP but only *uczeń₃*.

Therefore, we can talk about the existence of a specific semantic relationship only in reference to *mistrz₂* and *uczeń₃*. Apparently, this relation is more complex than it may seem. Philosophers express this notion by writing that the relationship between a master and an apprentice/student is based on acknowledging by the latter that he or she has common values with the master; the thinkers also say that the master is a paragon for the apprentice and that both the apprentice and the master may not be conscious of their own existence.

Also we, slightly despite all deconstructionist beliefs, have adopted the *mistrz – uczeń* relationship for the object of our study. Obviously, we are interested not only in social, philosophical or ethical elements of this relationship, but in ones which

are truly reflected in language and may be verified linguistically. The crux of the matter for us is, of course, the question about the character of the *mistrz – uczeń* relationship.

The structures *mój mistrz*, *mistrz Nowaka* czy *mój uczeń*, *uczeń Kowalskiego* (*my master*, *Nowak's master* or *my apprentice*, *Kowalski's apprentice*) in a natural way bring to my mind the thought of conversion. This suggestion is confirmed by the fallacious structure: **niczyj mistrz* (nobody's master) (even though, as it seems, it is possible to say: *On nie jest niczym mistrzem!* (He is nobody's master!).

Similarly, it is impossible to point out to a person who can be called an apprentice (*uczeń₃*) and who is not someone's student. This can be seen from another perspective: *mistrz₂* is obligatorily someone's master and implies the existence of someone (pointed in the genitive position) who is his or her apprentice or student (*uczeń₃*). Similarly *uczeń₃* is also unconditionally someone's student and implies the existence of someone (who is pointed to in the genitive position) who is his or her master (*mistrz₂*). This syntactic and semantic relationship between the words *mistrz* and *uczeń* is not undermined by such examples as the following one:

- (1) *Moim mistrzem był też Adam Hoffmann, choć nigdy (formalnie) nie byłem jego uczniem.* (Adam Hoffmann was also my master, even though I was never (officially) his student)

The sentence is not internally contradictory because the key to its understanding is the adverb *formalnie* (officially) as it refers to *uczeń₁* and not to *uczeń₃* who is in relationship with *mistrz₂* which does not result from the fact of 'being taught by ...'

If *uczeń₃* is someone's student (*mistrz₂*) and *mistrz₂* is someone's (*uczeń₃'s*) master, then maybe it is not worth asking about the convertive nature of *mistrz₁* and *uczeń₃* relationship. It is not difficult to predict what our answer to that question is. We have addressed this problem because we suspect that being someone's master and being someone's student or apprentice is a weak argument in favour of a convertive nature of this relation. It can be verified only by examining its language usage.

Therefore, we have assumed:

firstly that we acknowledge as syntactically primary these usages in which *mistrz₂* and *uczeń₃* occur in predicate position as predicatives;

secondly, we shall examine semantic interrelations for type (2) and (3) contexts:

- (2) *Kowalski jest mistrzem Nowaka.* [Kowalski is Nowak's master]
 (3) *Nowak jest uczniem Kowalskiego.* [Nowak is Kowalski's student]

Thirdly, we understand the notion of conversion similarly to Yuriy Apresyan [1], therefore we believe that this relation must refer to units which fulfil the following criteria:

1. opening at least two syntactical positions,
2. belonging to the same part of speech,
3. referring to the same occurrence (situation),

4. having ‘reversible’ syntactical structures²

Hypothetically, we assume that *mistrz*₂ and *uczeń*₃ fulfil Yuriy Apresyan’s conditions referring to the relation of conversion because the two units: /1/ are nominal proper names; /2/ open two syntactic slots (*someone’s* and *in a certain area*); /3/ have appropriately ‘reversed’ syntactic units: *X jest mistrzem Y-a* i *Y jest uczniem X-a* (*X is Y’s master* and *Y is X’s student*) and also /4/ they can be characterized by the same occurrence.

But if *mistrz*₂ and *uczeń*₃ are really examples of conversion, then the test for checking contradictions should also prove it. Cf.:

- (4) **Jeśli Kowalski jest mistrzem Nowaka, to nieprawda, że Nowak jest uczniem Kowalskiego.* (**If Kowalski is Nowak’s master, then it is not true that Nowak is Kowalski’s student.*)
- (5) ?? *Jeśli Nowak jest uczniem Kowalskiego, to nieprawda, że Kowalski jest mistrzem Nowaka.* (?? *If Nowak is Kowalski’s student, then it is not true that Kowalski is Nowak’s master*)

The obligatory condition for the relationship of conversion is internal contradiction between the segments of an implication (example 4). It should take place in both directions and therefore also in the case of an implication with a changing sentence sequence (cf. example 5). One can say that example 5 is unfortunate, unclear and even strange, yet one probably cannot say that it is internally contradictory. Such a possibility is also confirmed by (correct!) sentences of the following type:

- (5b) *Nowak jest uczniem Kowalskiego, ale Kowalski nie jest mistrzem Nowaka.*
(*Nowak is Kowalski’s student but Kowalski is not Nowak’s master.*)
- (5c) *Nowak jest (co prawda) uczniem Kowalskiego, ale mistrzem Nowaka jest Abeceński (a nie Kowalski).* (*Nowak is (however) Kowalski’s student but Nowak’s master is Abeceński (and not Kowalski).*)

And even though at first glance it may seem that *mistrz* and *uczeń* fulfil Apresyan’s conditions of conversion the above-mentioned examples show that for *mistrz*₂ and *uczeń*₃ such a relationship is not that obvious.

If this is not a relationship of conversion then what does the interrelation between *mistrz* and *uczeń* depend on?

We believe that the *mistrz-uczeń* relationship works only in one direction — as it is the student (*uczeń*) who creates and chooses his or her master (*mistrz*) and it is the student who can say the following: *Moim mistrzem jest Y* (*Y is my master*) or *Jestem uczniem Y-a* (*I am Y’s student*). The master can only confirm this fact. And if so, then not always and only in the situation when *uczeń*₃ is simultaneously taught by the master (*uczeń*₃ is also *uczeń*₁ and *mistrz*₂ is his teacher). The master

² *These two necessary positions change places (yet not semantic functions), cf. e.g.: X ma Y i Y jest X-a. czy Ojciec kupił od sąsiada samochód [za 15.000zł]. i Sąsiad sprzedał ojcu samochód za 15.000 zł(X has Y and Y belongs to X or Father bought a car from a neighbour [for 15,000 zł] and The neighbour sold the car to the father for 15,000 zł.).*

will never earnestly say about oneself: *Jestem mistrzem Iksińskiego* (I am Iksiński's master) (however, he or she will easily say³: *jestem Iksińskiego nauczycielem* [I am Iksiński's teacher]).

It is only Iksiński or some other third party (a sender from the outside of the relationship in question) that may express the following judgment:

Y jest moim mistrzem; Y jest mistrzem Iksińskiego. (Y is my master; Y is Iksiński's master.)

We have already mentioned that philosophers (and lexicographers) have associated the word *master* (*mistrz*) with being superior or even proficient in a field of knowledge or creation in which he or she is his student's guide. Intuitively we are ready to accept this. However, such a judgment is not confirmed by language use. The information about being the best, excelling or even proficient in a certain field (or in everything he or she does) does not necessarily lie within the meaning range of *mistrz*. Please compare the lack of contradiction in example (6):

(6) *Kowalski, Iksiński i Nowak zajmują się pewną (tą samą) dziedziną. Iksiński jest jej najwybitniejszym znawcą, ale to Kowalski, nie Iksiński jest mistrzem Nowaka.* (Kowalski, Iksiński and Nowak deal with (the same) area. Iksiński is the best expert in the field but it is Kowalski and not Iksiński who is Nowak's master.)

However, this statement does not exclude that despite the general opinion and even despite facts one may believe that Kowalski is better than Iksiński, cf.:

(6a) *Kowalski, Iksiński i Nowak zajmują się pewną (tą samą) dziedziną. Iksiński jest jej najwybitniejszym znawcą, ale zdaniem Nowaka Kowalski jest lepszy od Iksińskiego i jego wybrał na mistrza.* (Kowalski, Iksiński and Nowak deal with a certain (the same) area. Iksiński is the best expert in the field but according to Nowak Kowalski is better than Iksiński and that's why he has chosen him for his master.)

And this information about the subjectivity of judgment (expressed also formally by the structures *zdaniem X-a*, *według X-a*, *dla X-a* [according to X, in X's opinion, X believes that]) should, in our opinion, be reflected in the definition of *mistrz*. What is more, it also is an argument for the fact that the *mistrz-uczeń* relationship works only in one direction: the choice, making someone his or her master depends only on the apprentice or a student and on the fact whether he or she believes this person is worthy of this title. Philosophers (and lexicographers) say that the person ought to be exemplary. Yet linguistically speaking 'being exemplary' or 'being a paragon' is just another myth, which is proved by the following examples:

(7) *Kowalski nie jest doskonały, nie jest wzorem naukowca, mimo to Kowalski jest moim mistrzem.* (Kowalski is not perfect he is not an ideal scientist and even though Kowalski is my master)

³ Unless being Iksiński's teacher is a reason to be proud of.

- (8a) *Jest dla mnie wzorem (do naśladowania) i moim mistrzem. (He is my role model and my master.)*
- (8b) *Jest moim mistrzem i wzorem do naśladowania. (He is my master and a role model)*
- (8c) *Jest dla mnie wzorem do naśladowania, choć nie jest moim mistrzem. (He is my role model even though he is not my master.)*

It seems that the deciding factor in choosing someone for a master or in following this person's teaching is the belief of the student in unique skills or special knowledge that the person possesses. On these grounds the student acknowledged X as his master. It is rather unlikely that the student is aware of the fact that X is not the best expert in the field of their both interest and despite that fact he or she still thinks of him or her as a master or a guide in that area (*mistrz₂*). Cf.:

- (9a) **Kowalski jest dla Nowaka mistrzem w pisaniu reportaży, chociaż Nowak wie, że Kowalski nie umie pisać reportaży. (*Kowalski is Nowak's master in reportage writing even though Nowak knows that Kowalski does not know how to write a reportage).*
- (9b) *Kowalski jest dla Nowaka mistrzem w zbieraniu materiału do reportażu, chociaż według Nowaka Kowalski nie umie pisać reportaży. (Kowalski is Nowak's master in collecting reportage data even though according to Nowak Kowalski does not know how to write a reportage.)*
- (9c) *Kowalski jest beznadziejny zarówno w zbieraniu materiału do reportażu, jak i w ich pisaniu, mimo to dla Nowaka Kowalski jest mistrzem reportażu. (Kowalski is hopeless both in data collecting and writing reportages and despite these facts in Nowak's opinion Kowalski is master in reportage writing.)*

Internal contradiction (9a) results from the fact that according to student's judgment the master should be an expert in the field of their interest, while it is not the case. In the case of (9b) there is no contradiction because the judgments refer to two different fields and in (9c) we can see judgments of two different people — the student and a third party and that is why this sentence is not contradictory.

So far, we have been undermining the theory that the relationship between the words *mistrz* and *uczeń* is convertive. We know that it is a one-way relationship — it is the student who is an active party of the relationship and it is on him that the existence of causality relation between *mistrz* and *uczeń* depends on. *Mistrz* does not have to know that he is a master for someone because *uczeń₃* does not have to be taught by him or her (just as *uczeń₁* does not have to be taught by a teacher). *Uczeń₃* learns by himself from the person who he or she has chosen as a paragon in the field of his or her interest. This is why Tadeusz Różewicz could write in his poem "The Survivor" without any redundancy the following line: "Szukam nauczyciela i mistrza" (I seek a teacher and a master).

And as *mistrz* does not teach (he or she is not a teacher) *uczeń₃*, then *uczeń* may be looking for paragons among people who he or she does not know in person by choosing for his master someone who is no longer among the living. Cf (10) *Moim mistrzem jest też Max Scheller* (Max Scheller is also my master).

It is possible only on the condition that the student believes that his or her master has acquired expert knowledge or skills in the field of his interest (or that the person is better than others); and this is why he or she chooses this person for a guide in this field. That is also why the person decides to follow into master's footsteps by continuing his or her work or method. A person ceases to be a master when the student or apprentice loses this belief.

As R. Matuszewski perversely wrote in his *Alfabet*:

“W niejednym związku mistrza i ucznia trudno ustalić, kto kogo bardziej potrzebuje — mistrz bez ucznia jest przecież tylko emerytem!”. (‘There are many master-student relationships in which it is difficult to determine who needs who more — a master without a student is only a pensioner!’)

Pensioner is probably not the happiest word but it shows that a master without a student would be ‘nobody’s master’ because he or she would not be a master after all. As Michał Heller puts it [6]:

“wiedzy nigdy nie gromadzi się dla siebie, powinna nie tylko oświecać tego, co ją gromadzi, ale jakoś promieniować na innych” („one does not accumulate knowledge only for oneself; it should illuminate not only the one who accumulates it but it should be spread to others’).

Conclusions

- *mistrz₂-uczeń₃* determines a 3 actant structure: *ktoś, czyjś, w jakiejś dziedzinie wiedzy (/twórczości)* [someone, someone’s, in a certain field of science/art.].
- the occurrence of *mistrz₂-uczeń₃* relationship depends only on the student; it is he who makes someone his master: it takes a student to be a master.
- the master (*mistrz*) does not have to know that he or she is a master for someone. It is different with the word *uczeń* (*student*).
- in his student’s opinion the master is a good (the best) guide / teacher in a field of their own interest.
- *uczeń₃* may gather knowledge directly from the master (*mistrz*) by being his student (*uczeń₁*), but not necessarily. He or she may also take up and follow master’s work / idea without being taught by him. (This is where a doubt arises on whether in this situation he is master’s student or maybe a follower.)

Our research shows (we sincerely do hope so) that the convertive relationship between *mistrz₂* and *uczeń₃* does not take place because Apresyan’s condition of the ‘inverted’ structure has been fulfilled only formally. The only situation in which such the statement is justified is the situation of a third-party statement with a sender who is an outsider and who superficially judges the relationship between the student and his teacher.

Bibliography

- [1] Apresjan, J. D. (1980). *Semantyka leksykalna. Synonimiczne środki języka*, Wrocław.

- [2] Bańko, M. (red.) (2000). *Inny słownik języka polskiego*, PWN, Warszawa.
- [3] Bednarek, A., Grochowski, M. (1993). *Zadania z semantyki językoznawczej*, Toruń.
- [4] Czapigo, D. (2007). *Mistrz, uczeń, uniwersytet*. In *Societas/Communitas* 1(3), pages 75–88.
- [5] Dubisz, St. (red.) (2008) *Uniwersalny słownik języka polskiego*, t.2. and t.4, PWN, Warszawa.
- [6] Heller, M. (2009). *Jak być uczonym*, Kraków.
- [7] Różewicz, T., *Ocalenie*.
- [8] Skorupka, St. (red.) (1984). *Słownik frazeologiczny języka polskiego*, Warszawa.
- [9] Tischner, J. *Etyka a historia. Wykłady*. Tom II *Dzieł Zebranych*, wykład X: *Mistrz i uczeń*.
- [10] Zaron, Z. (2004). *Aspekty funkcjonalne kategorii rodzaju*, Puńsk–Warszawa.

DARJA FIŠER¹
TOMAŽ ERJAVEC²

¹Department of Translation, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

²Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

SEMANTIC CONCORDANCES FOR SLOVENE

Abstract. The paper presents the annotation of a Slovene language corpus at the semantic level. Manual annotation was performed in two cycles with an automatically generated semantic lexicon according to the wordnet model. The analysis of the results shows that nearly all polysemous words in the corpus can be assigned a sense from our 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.

Keywords: corpora, semantic annotation, semantic lexicon, Slovene language.

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 a corpus are connected with their meanings specified in a semantic lexicon. Semantic concordances are a 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. They 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 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 the 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 [19]. 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 [2] which are used in tasks such as information extraction and question answering [20]. An example of a corpus with semantic role annotations is PropBank [18]. In the project described in this paper we follow the second paradigm, which is the annotation of polysemous words with one of their senses [13]. This type of resources, such as SemCor¹ and MultiSemCor², have mostly been developed within the Senseval initiative and are used for automatic word sense disambiguation and machine translation [11].

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 project³ 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 license.

The jos100k corpus [5] is a 100,000 word Slovene corpus containing sampled paragraphs from the Slovene reference corpus FidaPLUS.⁴ The corpus is annotated

¹ <http://multisemcor.itc.it/semcor.php>

² <http://multisemcor.itc.it/>

³ <http://nl.ijs.si/jos/>

⁴ <http://www.fidaplus.net/>

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

sloWNet⁵ 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 [7] and is aligned to all wordnets for other languages that use Princeton WordNet identifiers for concept representation. The creation process consisted of three stages [9]:

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 included via 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: Equivalents 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 parallel corpus while the latter come from Wikipedia. sloWNet mostly contains nominal synsets, although there are some verbal and 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 [8].

4 The annotation process

The main goal of this project 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 the coverage of the senses it contains compared to the senses represented in the corpus and thereby evaluate the developed lexicon in a practical semantic task 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*

⁵ <http://nl.ijs.si/sloWnet/>

semantic annotation principle [15] 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 [12]) 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*, on the other hand, 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 [16].

In addition, we followed the *joint approach* of coordinated wordnet validation, refinement and corpus annotation as proposed by Agirre et al. [1] 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 validated and finalized each annotation.

As Figure 1 shows, the annotation procedure consisted of several stages: first, the 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.

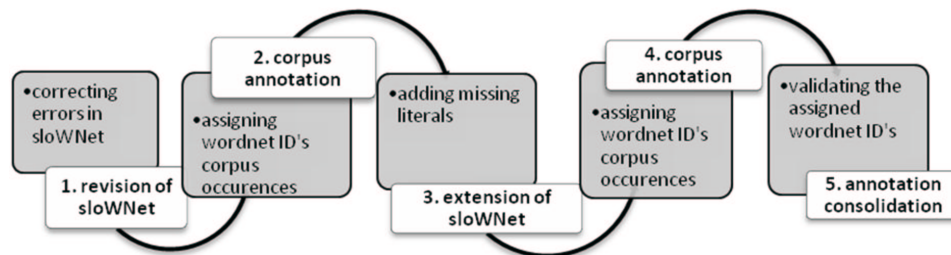


Fig. 1. The annotation procedure

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 where they have more than one sense and appear in the jos100k corpus with a frequency of 30 or higher. There were 100 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). While the most extracted nouns 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). This yields a total of 5,384 tokens with a manually assigned sense, which means that on average there are about 54 annotation examples for each noun included in the annotation process.

It was 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 seemed likely that the lexicon was still missing some senses for nouns which are frequent in the corpus but have very few senses or are even monosemous in the initial version of sloWNet, which is why these nouns needed to be carefully examined as well (e.g. *člen*, freq. 57 appeared 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 100 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 already existed 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 belonged to a multi-word expression which does not yet exist in sloWNet, it was extracted and was annotated as part of a multi-word expression and the expression added to sloWNet by the annotator.

4.2 Annotation guidelines

In order to facilitate the annotation process and to ensure a greater consistency of annotations, annotation guidelines were 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 [10], 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 studying an occurrence of the target word in context they determined which

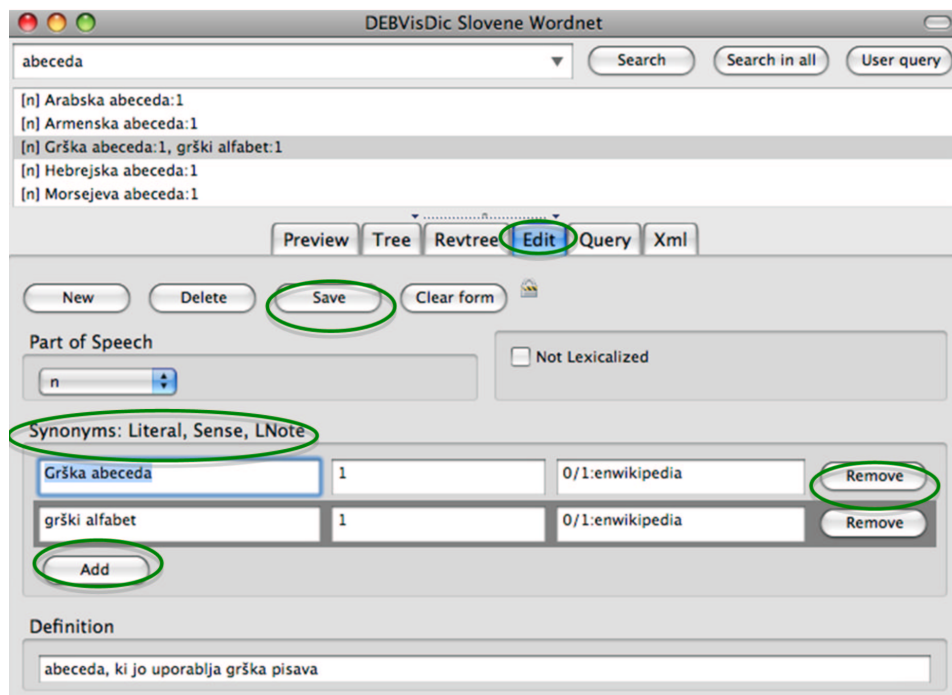


Fig. 2. Revision of synsets in DEBVisDic with highlighted editing features

synset was the most appropriate for it and annotated it with the corresponding synsensed id from wordnet (see Figure 3).

A	C	D	E	F
<i>n</i>	pomen	<i>Opomba</i>	<i>levi kontekst</i>	<i>beseda desni kontekst</i>
13			poto Triglava , je za	knjigo Triglav , Sveta gora Sl
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 !

Fig. 3. Annotation of the corpus in MS Excel

The goal of the annotation was to assign a sense to all occurrences of the target words. If more than one sense seemed appropriate despite best efforts to disambiguate them, the annotators were directed to choose the most basic sense. If an occurrence of the target word belonged to a multi-word expression (MWE), it was annotated with that sense and marked as a MWE. In case the target word was (part of) a proper name that does not exist in wordnet, the word was flagged as a (part of a) proper name.

If the appropriate sense could not be found in either sloWNet or PWN, the word was left unannotated and flagged as an OOV item. Most of 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

While the annotation is still undergoing some minor revision, such as suspicious and incongruent sense assignments, we report here on the current state of the semantic concordances over jos100k.

5.1 The extent of wordnet revision

The analysis of the first annotation cycle 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 was multi-word expressions, which could not be added automatically due to the limitations of the wordnet generation method.

Finally, the revised wordnets were merged and checked in the second annotation cycle by the third annotator who made sure that no hypernyms of the target words were left unannotated and that no untranslated synsets were used for corpus annotation.

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 were annotated by two different annotators in the first cycle is 3520. Less than one per cent of them were 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 were therefore re-examined and resolved by the third annotator in the second annotation cycle.

The annotators used 389 different senses in total, 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* (Eng. *court*) 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 word 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*). Words with a low number of assigned senses (3 or 4, such as *člen*, Eng. *Article* or *oče*, Eng. *father*) have an agreement exceeding 90%. Also, the level of agreement decreases with the increase of target word frequency in the corpus. This confirms our initial hypothesis that highly frequent and polysemous words would be difficult to annotate.

As the inter-annotator agreement was rather low, we checked whether annotators agreed on the most frequent sense for a given word. 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 almost in all the 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*. 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 clear-cut and are very fine-grained, which is a common criticism of the resource as a sense repository for practical applications.

Once the first cycle of annotation was completed, the second cycle began. The files containing double annotations were given to the third annotator who compared the assigned senses and chose the best one. She also checked any notes made by the first two annotators. At this stage, the initial set of 77 annotated nouns was extended to the entire annotation set which is now complete.

Table 1 gives the basic statistics over the reviewed and validated annotation set. Each of the 100 nouns has, on average, about 50 occurrences in the corpus. The annotators assigned over 500 different synsets to this set, i.e. over 5 senses per noun. The figures are slightly higher compared to the first annotation cycle because annotations for 23 words that were missing from the first cycle were added at this stage. Five of the annotated words are monosemous (e.g. *muzej* Eng. *museum*), while the most polysemous word is *čas* (Eng. *time*) which has 15 senses. Finally, 71 tokens in the corpus were left unannotated. 46 of them are proper names, or parts of proper names not present in PWN, and for a further 25 tokens no appropriate synset could be found, e.g. for *voda* (Eng. *water*) in *voda na [nekogaršnji] mlin / water on [somebody's] mill*, a Slovene idiom.

Tokens	5,384
Literals	100
average tokens/literal	53.8
min tokens/literal	30
max tokens/literal	346
Synsets	502
Average synsets/literal	5.4
min synsets/literal	1
max synsets/literal	15
proper names	46
no synset	25

Table 1. Annotation statistics

Although MWEs were not explicitly selected for annotation, a surprisingly large number of focus nouns turned out to be part of MWEs which had, or could sensibly have their own literals in the wordnet. Table 2 gives the number of instances tagged as MWEs, almost 10% of the overall tokens, of which almost half had to be annotated with an approximate synset. Altogether, MWEs were tagged with 170 synsets, a third of the overall total.

MWE tokens	471
MWE tokens with approximate synset	223
MWE tokens with appropriate synset	248
MWE synsets	170

Table 2. Multi-word expressions

6 Conclusion

The paper has presented the first attempt to semantically annotate a corpus for Slovene. In the project, 100 high-frequency nouns from the jos100k corpus were assigned wordnet senses. The validation of sloWNet with corpus annotations has shown that most core senses that were required to annotate the corpus had already been present in sloWNet whereas the same is not true for peripheral senses and especially for multi-word expressions which had to be added by the annotators in many cases. Multi-word expressions were especially difficult, as in almost half of the cases no exactly appropriate sense could be found in wordnet. 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 annotation tasks. 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 is 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 [21]. 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% [14].

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 was compensated by a second annotation cycle in which an experienced team of 4 annotators checked and consolidated the differences between the original two annotators. Their work hopefully provided much more reliable and consistent data that will be more useful in further research. But even the experienced annotators encountered significant problems in determining the best sense for each token at the second annotation cycle, often involving lengthy discussions, and inconclusive decisions. The most problematic words to annotate were hard-to-distinguish senses and culturally-specific expressions that are not included in the Princeton WordNet which was used as the backbone of the Slovene wordnet.

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 [17] and automatically by Bruce and Wiebe [4] who achieved a 10% improvement on the results.

Notwithstanding the difficulties of the annotation, the result is the first Slovene corpus that is annotated at the semantic level. In the future, we wish to extend the annotation set to multi-word expressions that were already present in Slovene wordnet but were excluded from this annotation cycle, and to all high-frequency monosemous words. The annotated corpus will be freely available for linguistic analysis or as a training set for applications in human language technologies, while sloWNet is already publicly available under the Creative Commons license.

Bibliography

- [1] 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.
- [2] Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. *Proceedings of ACL'98*, (pp. 86–90). Montreal.
- [3] Bentivogli, L., Forner, P., & Pianta, E. (2004). Evaluating cross-language annotation transfer in the MultiSemCor corpus. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva.

- [4] 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.
- [5] Erjavec, T., & Krek, S. (2008). The JOS Morphosyntactically Tagged Corpus of Slovene. *Proceedings of LREC'08*. Marrakech.
- [6] Fellbaum, C. (2002). On the Semantics of Troponymy. In R. Green, C. Bean, & S. Myaeng (Eds.), *Relations*. Dordrecht: Kluwer.
- [7] Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, London: MIT.
- [8] Fišer, D., & Erjavec, T. (2008). Predstavitev in analiza slovenskega wordneta. *Proceedings of IS-LTC'08* (pp. 37–42). Ljubljana.
- [9] Fišer, D., & Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. *Proceedings of TSD'08*. Brno.
- [10] 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.
- [11] Kilgarriff, A. (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language. Special Issue on Evaluation*, 12 (4), 453–472.
- [12] Kilgarriff, A. (1998). SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. *Proceedings of LREC'98*. Granada.
- [13] Landes, S., Leacock, C., & Tengi, R. I. (1998). Building Semantic Concordances. In C. Fellbaum (ed.), *WordNet* (pp. 199–216). Cambridge, Massachusetts: MIT Press.
- [14] Mihalcea, R., Chklovski, T., & Kilgarriff, A. (2004). The Senseval-3 English lexical sample task. *Proceedings of ACL/SIGLEX Senseval-3*.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31 (1), 71–105.
- [19] 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.
- [20] Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003). Using Predicate-Argument Structures for Information Extraction. *Proceedings of ACL'03*. Sapporo.
- [21] Veronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. *Programme and advanced papers of the Senseval workshop*. Herstmonceux Castle.

PETER ĎURČO¹
RADOVAN GARABÍK²
DANIELA MAJCHRÁKOVÁ²
MATEJ ĎURČO³

¹Univerzita sv. Cyrila a Metoda v Trnave, Trnava, Slovakia

²Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

³Austrian Academy Corpus, Austrian Academy of Sciences, Vienna, Austria

CONTRASTIVE DICTIONARY OF GERMAN AND SLOVAK COLLOCATIONS¹

Abstract. In the article we discuss ongoing work concerning a confrontational German-Slovak collocation lexical database. The database consists of two parts, a section of German collocations with Slovak equivalents and a section of Slovak collocations. Intended size of the database is several hundred words of different parts of speech (nouns in the first phase of the project) for each of the languages, together with their collocation profiles. The database uses MediaWiki engine and a wiki-based approach to article editing and collaborative work of a team of lexicographers.

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) [24, 25]. Clearly, since then the whole language underwent immense changes in almost all of its parts, starting with the whole

¹ 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. Parts of this article have been published in [16].

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 German and 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 [12]. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partical collocations.

We exclude regular systematic terminological and proprial collocations from the database, leaving only irregular systematic collocations (idioms, phrasemes), regular text-collocations (e.g. *zimná rekreácia*) and fixed text-systematic collocations (e.g. *krájať nadrobno*, *hovoriť úsečne*).

2 Obtaining collocation profiles

To obtain Slovak collocation profiles from a large lemmatized corpus we are using the sketch engine² [19] — a corpus tool which generates word sketches, i. e. corpus based summaries of grammatical and collocational behaviour of a word. Disadvantages of the 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, MI-score, MI3, log likelihood, min. sensitivity and salience to recognize word co-occurrences [20].

In our lexical database, the Slovak 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.

² <http://www.sketchengine.co.uk/>

The German collocations are obtained from the IdS corpus [9], DWDS corpus [8] and the Wortschatz-Portal [27]. We use the most frequent words from [17], updated by the data from [18] (currently the only one contemporary frequency list of German words, created out of a specialized corpus reference texts). Unfortunately, there is no word sketch created for German language, however we are working on a preliminary version. We then add Slovak equivalents to the German collocation in the database.

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.

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 7.2.

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.

4 Building Slovak collocations

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.

5 Slovak equivalents of German collocations

German section of the database consists of German language collocations with their equivalents in Slovak. During the construction of the database, we observed several common patterns in the equivalency[5]:

Monosemic German words are reflected in monosemic equivalency – in the whole collocation profile, there is consistently one Slovak equivalent. Examples of such words are Schutz – ochrana (protection), Reise – cesta (journey); Schüler – žiak (pupil); Sommer – leto (summer).

effektiver Schutz	efektívna <i>ochrana</i>
Schutz gegen Inflation	<i>ochrana</i> proti inflácii
jmdm. Schutz gewähren	poskytnúť niekomu <i>ochranu</i>

Table 1. Example Slovak collocation equivalents for a monosemic German word *Schutz*

Polysemic German words cover several meanings and consequently they are assigned several Slovak equivalents — sometimes even one German meaning is translated into several (closely related) Slovak words.

einfacher Satz	holá <i>veta</i>
entscheidender Satz	1. rozhodujúca <i>veta</i> , 2. rozhodujúci <i>set</i>
ermäßigter Satz	znížená <i>sadzba</i>
der zweite Satz des (Volleyball) spiels	druhý <i>set</i> hry (volejbalu)
Satz Autoreifen	<i>sada</i> pneumatík
Satz des Kaffees	<i>usadenina</i> z kávy
Satz Fische	<i>násada</i> rýb
Satz von Anordnungen	<i>súbor</i> predpisov
mit einem jähen / schnellen Satz an der Tür sein	rýchlym <i>skokom</i> byť pri dverách

Table 2. Example Slovak collocation equivalents for a polysemic German word *Satz*

Monosemic foreign German words (*Fremdwörter*, *Lehnwörter*) are usually reflected by monosemic Slovak equivalents, thanks to the common Greco-Latin heritage often of the same etymology. Examples:

Reform – reforma (reform);
 Symbol – symbol (symbol);
 Student – študent (student).

Polysemic foreign German words cover several meanings, and are usually assigned several Slovak equivalents, thanks to etymological divergence there are often several possible Slovak words (usually a loanword and its Slovak equivalent). Examples: Reaktion – reakcia, odozva, odpoveď, chemická reakcia, protipôsobenie (reaction); Situation – situácia, stav, pomery (situation); Transport – transport, preprava, doprava (transport); Qualifikation – kvalifikácia, posúdenie (qualification).

Polyequivalence — there are words, whose equivalents are not just a single translation, but a set of several synonyms that are sometimes freely interchangeable, but sometimes not. We assign the Slovak equivalents according to their typical usage in the Slovak language. E.g., German word *Sinn* (sense, meaning, point, idea, consciousness, feeling, . . .) can be translated by *význam*, *zmysel*, *cit*, *pochopenie*, *mysleň*, but the collocations of the equivalents are somewhat rigid and not all of them are interchangeable.

im engeren Sinn des / eines Wortes	v užšom <i>zmysle</i> slova
keinen Sinn in etwas sehen	v niečom nevidieť žiadny <i>význam</i> / <i>zmysel</i>
jmds. Sinn ist nur auf dieses eine Ziel gerichtet	niekoho <i>mysleň</i> je napriamená len na jeden cieľ
frohen Sinnes sein	byť veselej <i>mysle</i>

Table 3. Example of Slovak collocation equivalents diversity for the German word *Sinn*

The same applies for idiomatic word usage. E.g. the German word *Stunde* (hour) has more or less straightforward meaning, translated by Slovak *hodina*, but when the German collocations cover the idiomatic usage, we have to use corresponding Slovak idioms.

in einer schwachen Stunde	v slabej <i>chvíli</i> , <i>chvílke</i> - *v slabej <i>hodine</i>
die richtige Stunde abwarten	vyčkať správny <i>okamih</i>
in einer stillen Stunde	vo chvíľke <i>pokoja</i>
zur richtigen Stunde	v pravý <i>čas</i>

Table 4. Example of idiomatic usage of the German word *Stunde*

6 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 [13, 15]. These two projects share

the same database and the same MediaWiki installation, and (to an extent) use the same methods and guidelines regarding the collocation profiles. The databases are distinguished on the logical level, by marking each entry as belonging to one of the Slovak part of speech collocation categories **Slovak Nouns**, **Slovak Adjectives**, **Slovak Verbs**, **Slovak Particles** or to the **German collocation** category. The rest of the pages (not belonging to either category) are system, user or administrative pages, or user discussions.

The database macrostructure is simple — all the entries are equal, each entry corresponds to one MediaWiki page, we are using neither subpages nor redirects. 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 could be used to distinguish between German and Slovak entries.

7 Structure of an entry

Overall structure of an entry is identical for both German and Slovak parts of the database. They differ in the language of section titles, where we use German terms for the German entries and Slovak terms for Slovak entries. In the following text, we describe both version together, putting German section names first, followed by Slovak ones.

An entry page consists of three main sections: *Bedeutung* or *Významy* (Meanings), *Kollokationen* or *Kolokácie* (Collocations), *Links* or *Externé odkazy* (External links). While the structure of *Bedeutung* and *Links*, or *Významy* and *Externé odkazy* is the same for all the parts of speech and these sections do not have any substructure, the structure of *Kollokationen* or *Kolokácie*, the most important section, is more complicated [14].

7.1 Bedeutung, 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; second, trying to define and distinguish meanings is traditionally a very cumbersome process, 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.

7.2 Kollokationen, 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 subsections, each for a specific collocation combination (see Fig. 1, 2, 3, 4).

The subsections' 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*, *Sub1Akk* or *Sub1Aku*, *Sub1Lok*, *Sub1Ins*). We are ignoring the Slovak vocative controversy by conflating (semantic) vocatives with the nominative case — fortunately, none of the nouns chosen for the collocation dictionary is from the set of those few Slovak words that have a morphological vocative.

The other part of the subsection name reflects describes the neighbouring word part of speech, so it can be one of *Sub2*, *Verb*, *Attr* or *Atr* (another noun, verb, attribute). *Attr* or *Atr* subsumes adjectives, pronouns, particles and 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 + Sub1Gen* indicates that the subsection contains collocation of verb and base noun in genitive (not necessarily in this order).

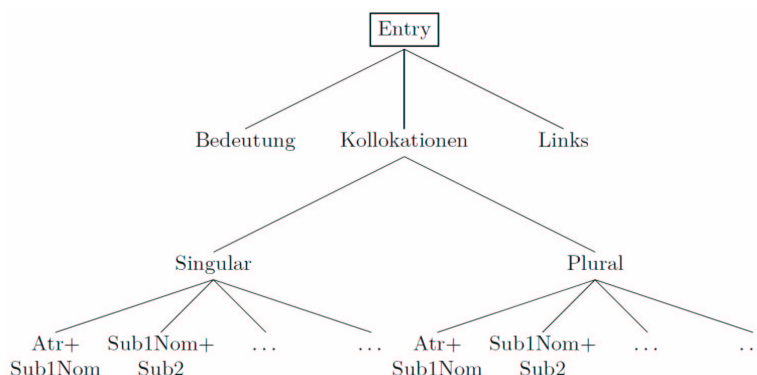


Fig. 1. Entry structure diagram for German nouns

7.3 Links, 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

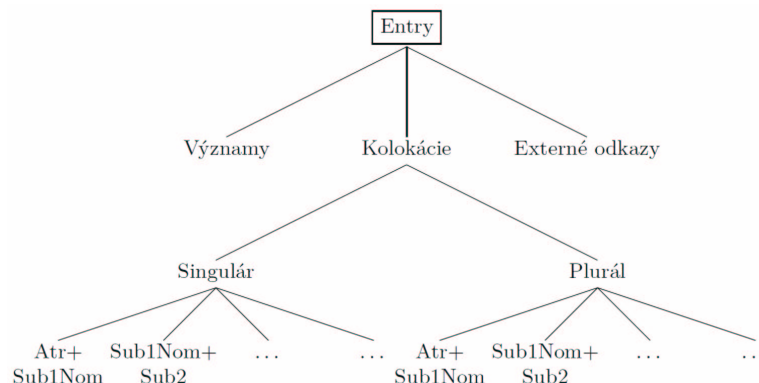


Fig. 2. Entry structure diagram for Slovak nouns

Sub1	Sub2	Verb	Atrr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Sg Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Sg Akku	Sub1Akku+Sub2	Sub1Akku+Verb	Sub1Akku+Attr
Pl Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Pl Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Pl Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Pl Akku	Sub1Akku+Sub2	Sub1Akku+Verb	Sub1Akku+Attr

Fig. 3. Matrix for the entry structure of a German noun

Sub1	Sub2	Verb	Atr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Sg Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Sg Aku	Sub1Aku+Sub2	Sub1Aku+Verb	Sub1Aku+Attr
Sg Lok	Sub1Lok+Sub2	Sub1Lok+Verb	Sub1Lok+Attr
Sg Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Attr
Pl Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Pl Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Pl Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Pl Aku	Sub1Aku+Sub2	Sub1Aku+Verb	Sub1Aku+Attr
Pl Lok	Sub1Lok+Sub2	Sub1Lok+Verb	Sub1Lok+Attr
Pl Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Attr

Fig. 4. Matrix for the entry structure of a Slovak noun

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.

Slovak language macros. 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 [22] published at the L. 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 [21]. 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.

German language macros. In the German section, the entries use only a single macro to link to all of the external sources — `{{links-de|...}}` links to several German online dictionaries: dict.cc German-English dictionary [11] (includes full morphology paradigms), the LEO German-English dictionary [10], DWDS monolingual dictionary [8] and Zoznam German-Slovak dictionary [26].

8 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 [7] for automated processing in various programming languages. However, we settled on using a slightly different approach — WikipediaFS [6], a fuse-based [23] 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.

9 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 collocation is written with the base in its corresponding case/number, there is only one exception — in the Slovak database, the combination *Atr + Sub1Nom* is so frequent that we decided to omit the base if in nominative, when it immediately follows the attribute. Auxiliary particles/pronouns are sometimes rearranged, to fit the syntactical requirements of the base (this applies mainly to the reflexive pronouns *sa, si* in combination with infinitives). German reflexive verbs that take the subject in dative are marked with a special qualifier (Dat.), e.g. *sich (Dat.) die Augen reiben; sich (Dat.) die Augen verderben; sich (Dat.) die Augen wischen*.

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.

In the Slovak database, each 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.

In the German database, collocations in each subsection are organized in a two column table, with German collocates on the left and Slovak equivalents on the right. We are using standard MediaWiki table syntax, starting each table with a following code snippet (preamble and table header):

```
{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
```

Each collocation is in one table row, the rows are separated by a |- string (a vertical line followed by a hyphen-minus), in each row there is a German collocate, followed by a string || (two vertical lines), followed by a Slovak collocate.

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).

In the Slovak section, there are on average 173 collocations per entry — the distribution of entry sizes is depicted on Fig. 7. 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.

In the German section, the average number of collocations per entry is 195.5 (see Fig. 8), the median is 150. The entry with the fewest number of collocations is *September*, with 21 collocations, the entry with the highest number is *Kind* (child), with 703 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. The subjective differences are even more pronounced when comparing two different languages, where also the language competence of the lexicographer plays its role (if they are not native speakers of the language), and also different methodology of obtaining collocation profiles.

```

==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. 5. Fragment of a Slovak collocation entry, word *pomer*

10 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 Slovak database contains collocation profiles of 190 nouns and 38 particles, the German database contains 280 collocation profiles, all of them are nouns.

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

```

===Attr + Sub1Nom===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|beigegebene Abbildungen || priložené obrázky
|-
|unzüchtige Abbildungen || nemravné / oplzlé obrázky
|-
|verschiedene Abbildungen || rôzne obrázky
|-
|zahlreiche Abbildungen || početné obrázky
|}

===Sub1Nom + Sub2===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|Abbildungen aller Art || rôznorodé / rozmanité zobrazenia
|-
|Abbildungen durch Linsen || zobrazenia prostredníctvom šošovky
|-
|Abbildungen in Band xy || obrázky vo zväzku xy
|-
|Abbildungen im Text || obrázky v texte
|-
|Abbildungen oder Darstellungen || zobrazenia alebo znázornenia
|-
|Abbildungen von Funden / Gegenständen / ... || obrázky nálezov / predmetov / ...
|-
|Beschreibungen der Abbildungen || popisy obrázkov
|-
|Zeichnungen oder Abbildungen || kresby alebo obrázky
|}

===Sub1Nom + Verb===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|Abbildungen enthalten etwas || obrázky obsahujú niečo
|-
|Abbildungen geben Vorstellung / Zeugnis / ... || obrázky poskytujú
|predstavu / svedectvo / ...
|-
|Abbildungen veranschaulichen etwas || obrázky znázorňujú niečo
|-
|die Abbildungen zeigen jn, etwas || obrázky ukazujú niekoho, niečo
|}

```

Fig. 6. Fragment of a German collocation entry, word *Abbildung*

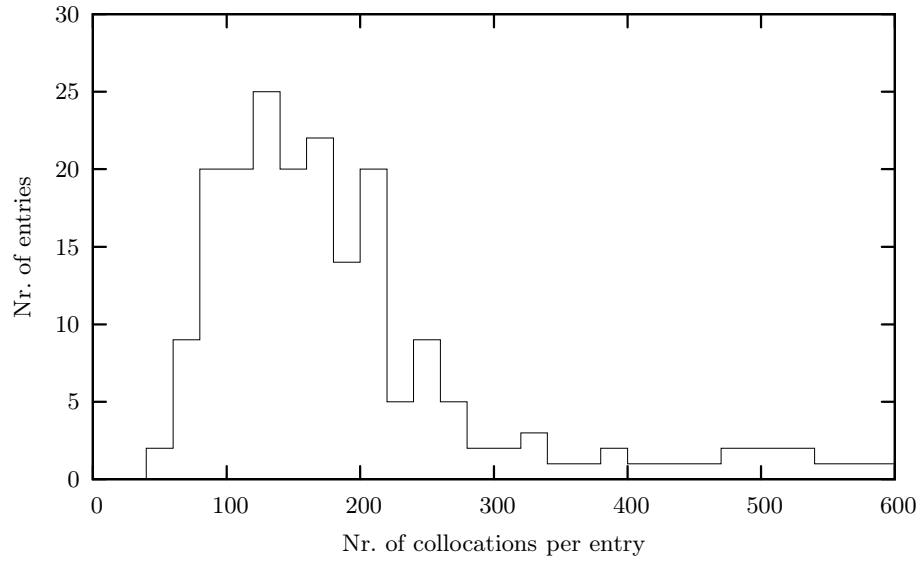


Fig. 7. Distribution of number of collocations per noun in the Slovak section, bin size = 20

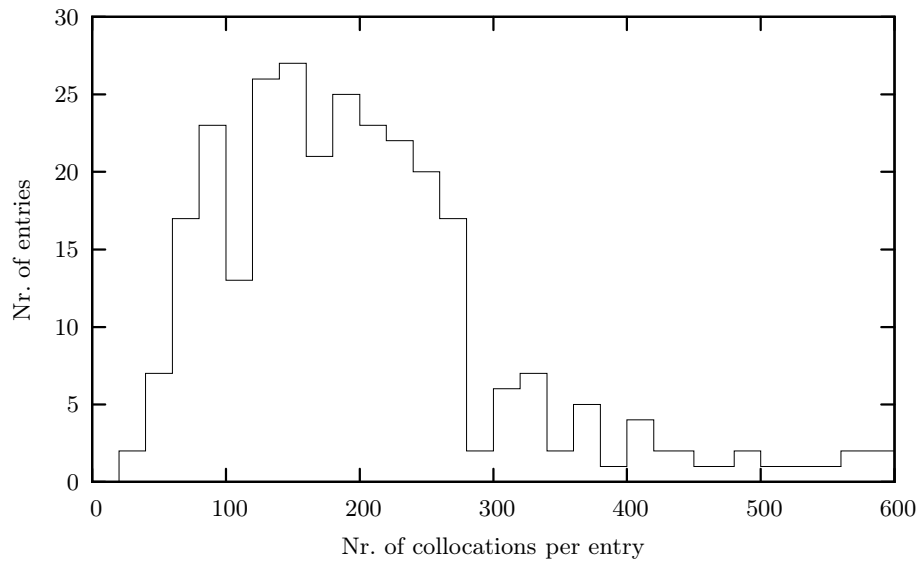


Fig. 8. Distribution of number of collocations per noun in the German section, bin size = 20

already existing Slovak language 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.

The bilingual German-Slovak collocation database will offer excellent possibilities for contrastive linguistic studies, and will be similarly useful for Slovak speakers in learning German as a foreign language as well as for German speakers in learning Slovak.

Bibliography

- [1] <http://www.google.com>.
- [2] <http://www.ask.com>.
- [3] <http://www.yahoo.com>.
- [4] <http://www.cuil.com>.
- [5] Banášová, M. (2008). Polysemie und Polyäquivalenz der Kollokationen im Deutsch-slowakischen Kollokationswörterbuch. In Ďurčo, P. (Ed.), *5. Kolloquium zur Lexikographie und Wörterbuchforschung. The Fifth International Colloquium on Lexicography/Feste Wortverbindungen und Lexikographie/Fixed word combinations and Lexicography.*, Bratislava, Slovakia. (in press).
- [6] Blondel, M. WikipediaFS. <http://wikipediafs.sourceforge.net/>. Retrieved 2009-06-08.
- [7] Botwiki. <http://botwiki.sno.cc/wiki/Manual:Frameworks>. A wiki for documenting and testing bots. Retrieved 2009-06-08.
- [8] Das Digitale Wörterbuch der deutschen Sprache des 20. Jh. <http://www.dwds.de>.
- [9] Das Portal für die Korpusrecherche in den Textkorpora des Instituts für Deutsche Sprache. <http://www.ids-mannheim.de/cosmas2/>.
- [10] Deutsch-Englisch Wörterbuch, Ein Online-Service der LEO GmbH. <http://dict.leo.org>.
- [11] dict.cc, English-German Dictionary. <http://dict.cc>.
- [12] Ďurčo, P. (2007). 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.
- [13] Ďurčo, P. (2007). 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.

- [14] Ďurčo, P. (2007). Zásady spracovania slovníka kolokácií slovenského jazyka. <http://www.vronk.net/wicol/images/Zasady.pdf>. Online documentation.
- [15] 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.
- [16] Ďurčo, P., Garabík, R., Daniela, M., & Ďurčo, M. (2009). Dictionary of Slovak Collocations. In Koseska-Toszewa, V., Dimitrova, L., Roszko, R. (Eds.), *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography.*, (pp. 128–137)., Warsaw, Poland. Institute of Slavic Studies, Polish Academy of Sciences.
- [17] Glaboniat, M., Muller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch. Gemeinsamer Europäischer Referenzrahmen*. Berlin, Germany.: Langenscheidt.
- [18] Jones, R. L. & Tschirner, E. (2005). *A Frequency Dictionary Of German*. Routledge.
- [19] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The sketch engine. *Information Technology*, 105.
- [20] Majchráková, D. & Ďurčo, P. (2009). Compiling the First Electronic Dictionary of Slovak Collocations. To be published.
- [21] Slovak National Corpus. <http://korpus.juls.savba.sk>.
- [22] Slovenské slovníky. <http://slovník.juls.savba.sk>.
- [23] Szeredi, M. Filesystem in Userspace. <http://fuse.sourceforge.net/>. Retrieved 2009-06-08.
- [24] Tvrđý, P. (1931). *Slovenský frazeologický slovník*. Trnava: Spolok sv. Vojtecha.
- [25] Tvrđý, P. (1933). *Slovenský frazeologický slovník. Druhé doplnené vydanie*. Praha and Prešov: Nákladem Československej grafickej unie, úč. spol.
- [26] Web slovník. <http://webslovník.zoznam.sk>.
- [27] Wortschatz Universität Leipzig. <http://www.ids-mannheim.de/cosmas2/>.

LUDMILA DIMITROVA¹

VIOLETTA KOSESKA-TOSZEWA²

¹Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

²Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland

CLASSIFIERS AND DIGITAL DICTIONARIES

Abstract. The paper discusses some problems related to entry classifiers in digital dictionaries. Information technologies offer great possibilities to linguists and lexicographers for the development of various dictionaries, especially for bi- and multilingual digital dictionaries. The authors' point of view is based on their experience from the development of the first Bulgarian-Polish Digital Dictionaries. The dictionaries are being developed 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, coordinated by L. Dimitrova and V. Koseska. The experimental version of the Bulgarian-Polish electronic dictionary is prepared in WORD-format and consist approximately 20 thousand dictionary entries. The dictionary is used for creation of the lexical database (LDB) that is an entry point to the relational database (RDB) of the first Bulgarian-Polish online dictionary. The structure of the LDB allows synchronized and unified representation of the information for Bulgarian and Polish, which is a step towards the creation of online Polish-Bulgarian dictionary in the future.

Keywords: digital dictionary, entry classifiers, digital corpus, semantics and contrastive studies.

1 Introduction: Basic advantages of the digital vs paper dictionary

Information technologies offer great possibilities to linguists and lexicographers for the development of various dictionaries, especially for bi- and multilingual digital dictionaries. The following remarks are based on our experience from the development of the first Bulgarian-Polish Digital Dictionaries.

First let us mention briefly the basic advantages of the digital vs paper dictionary. The preparation of the paper dictionary is a continuous process (it takes several months or even years) and the dictionary remains unchangeable after publication, i.e. the paper dictionary is a static collection of dictionary entries. The creation of a digital dictionary is also a continuous process in time, but the collection of words

can be continuously expanded. New dictionary entries can be added or their content can be enriched by addition of supplementary information about the headword (grammatical, etymological), of examples (for clarification of usage), of phrases and combinations, etc. The digital dictionary is a dynamic collection of dictionary entries, which provides a dynamical structure of the dictionary entry per se. This characteristic allows:

- a relatively easy adaptation of the lexical database, which the collection of words in a dictionary actually is, to a new (improved) model of dictionary entry and its enrichment with new information, for example the addition of the word-forming group of the headword, etc.
- a refinement of the system of classifiers, used for structuring the dictionary entry in order to describe optimally the headword.
- use of the digitally-presented information for the creation of a new (or different type of) digital dictionary, for example two monolingual digital dictionaries (explanatory or terminological) in two different languages can be used to produce a new bilingual dictionary (although in practice that is non-trivial);
- when necessary – last but not least – correction of various mistakes.

2 Problems and challenges

One of the main problems of the development of digital dictionaries is the *choice of classifiers* of the dictionary entry. Whenever the development of a system of bilingual digital dictionaries, serving as a basis for a system of multi-lingual dictionaries in perspective, is concerned, there arises an issue of *unification of the classifiers* in the dictionary entry. This is an *issue of harmonisation of the classifiers for various languages*, whose solution has to present a *unified selection of classifiers and a standard form of their presentation*. In a broader sense the issue of unification of classifiers in the dictionary entry *approaches the issue of a new part-of-speech classification* keeping in mind the specifications of a digital dictionary.

3 Classifiers

It is accepted that classifiers carry different morphosyntactic and/or semantic characteristics of the words (in particular, the dictionary entry). They split the set of words according to properties. Most often the classifier connects the word with its respective part of speech, depending on the class, to which the word belongs. But the classifier can show specific features of the word, such as gender, number, tense, etc. Tense is a meaning of the form, but has not been fully defined, see the examples about aorist (*аорисм* in Bulgarian) and imperfectum (*имперфект*).

At the current stage of research the part-of-speech classification in a natural language continues to be under discussion because it is not consecutive. It is based on different criteria (morphological, syntactic or “narrow” semantic) which are reduced only to the separation of grammatical categories. Thus the part-of-speech classification is different not only depending on language but is also significantly

different in certain languages. This fact made us consider the unification of the part-of-speech classification at least in the two Slavic languages in our study, see [15]. In order to accept a common for these languages, i.e. a standard type of part-of-speech classification we start a discussion on these issues in this article. At the same time we offer new arguments on this issue on Bulgarian and Polish material using F. Slawski's *Bulgarian-Polish Dictionary* [17] as well as examples from machine translation from English to Polish and from English to Bulgarian.

So far the meaning of the forms has been the Achilles' heel of the description, dictionaries and corpora, both mono- and bilingual. That is why we shall focus our attention on some entries in the Bulgarian-Polish Dictionary depending on the form's meaning and its differentiation from a given meaning.

Examples

Let us have a look at the following examples of dictionary entries which do not explain anything in the dictionary. It is not clear whether they concern form or meaning. Neither is it clear what the meaning of this form is.

Example 1. Entry with headword "aorist"

а̀̀ри́ст, -и *m gram. aoryst m*

This entry with headword the verbal form "aorist" does not make clear what kind of aorist is meant. In Bulgarian aorist can be formed from perfective and imperfective verbs, for instance, *написа* and *писа*. In the sentence *Той написа интересна книга.* the form *написа* is a perfective aorist. But the form *писа* in *Той писа тази книга 5 години.* is an imperfective aorist.

Perfective aorist determines an event that has happened before the state of speaking and reserves a place for a unique quantifier in the sentence's semantic structure [11], [13].

Imperfective aorist means a configuration of states and events that have happened before the state of speaking and reserves a place only for a unique quantifier in the sentence's semantic structure [11], [13], [15].

In order to describe the two different meanings of aorist we suggest the following two new dictionary entries:

а̀̀ри́ст от свършен вид, -и *m gram.* – единично събитие настъпило преди състоянието на изказването. (A unique event that has happened before the state of speaking.)

This meaning is conveyed by Polish perfective praeteritum [11]. For example:

Той боледува от грип.

On chorował na grype.

а̀̀ри́ст от несвършен вид, -и *m gram.* – единично квантифицирана конфигурация от състояния и събития, извършваща се преди състоянието на изказването. (A unique-quantified configuration of states and events that have happened before the state of speaking.)

This Bulgarian meaning is conveyed by Polish imperfective praeteritum [11]. For example:

В четвъртък ходих пеша до центъра на града.
 W czwartek chodziłam pieszo do centrum miasta.

Example 2. Entry with headword “imperfect”:

Имперфект *m gram. Imperfectum n*

Just as in the case of aorist, we have no information that in Bulgarian this form (if form is meant here) is formed from imperfective as well as perfective verbs. We have no information about the difference in the meaning of the two. The imperfective imperfect serves to determine configurations of states and events that have happened and lasted before the state of speaking. The form here in contrast to the imperfective aorist (which is connected with a unique quantifier), reserves a place for all quantifiers (existential, universal, although rare, unique) [16]. In this case our suggestion about the new entry with headword “imperfective imperfect” is the following:

Имперфект от несвършен вид, -и, *m gram.* Многозначно квантифицирана конфигурация от състояния и събития, настъпили и траещи преди състоянието на изказването — по значение съответства полската форма praeteritum от несвършен вид. (Multiply-quantified configuration of states and events that have happened and lasted before the state of speaking – by meaning corresponds to Polish imperfective praeterium.)

Той понякога намирал време за разходка.
 On od czasu do czasu znajdował czas na spacer.
 Той понякога боледувал от грип.
 On czasem chorował na grype. (See [15])

Concerning the alternative “имперфект от свършен вид” (perfective imperfect) we must note that it occurs very rarely and only in special modal, conditional contexts, such as: *Пийнеше ли* (perfective imperfect), *вдигаше* (imperfective imperfect) *много шум около себе си*.

Example 3.

Let us consider the entry:

минал *part. adi* przeszły, zeszły, ubiegły; **миналата година** dwa lata temu; **-о време** *gram.* Czas przeszły.

Here we have another type of problems. There are three Polish forms *przeszły, zeszły, ubiegły* that correspond to the Bulgarian form *минал* (‘past’). As in the case of aorist and imperfect it is not clear what is meant — meaning or form of past tense.

If a meaning is meant, it is not clear what past tense is meant. If however a form is meant, it must be mentioned that this is a form with multiple meanings.

We already mentioned ([5], [6]) that a single form can have multiple meanings and they naturally vary in number across the various languages. This is a problem whose solution would allow the creation of a new **L₂-L₁** dictionary from a **L₁-L₂** dictionary. How do we invert a Bulgarian-Polish dictionary entry so that it represents a Polish-Bulgarian dictionary entry? It is obvious that the elimination

of shortcomings among the entries of a given L_1 - L_2 bilingual dictionary, eliminating the impossibility of a new ordering of information with the scope of obtaining an inverted L_2 - L_1 bilingual dictionary, requires a reconsideration of the representation of the relation “form-meaning” in the dictionary.

An automated inversion of the dictionary is possible and easy to implement only when the relation “form-form” is considered. But then the inverted dictionary is quite poor and its cognitive value quite weak.

In order to keep all different meanings we suggest for discussion the option where each meaning is shown with the same form but enumerated, for example:

- минал** (1) – przeszły
- минал** (2) – zeszyły
- минал** (3) – ubiegły

In other words the form is indexed and appears in the list as many times as its different meanings.

Another example from the Bulgarian-Polish Dictionary – the dictionary entry for headword *май*:

- май** (1) *m* maj; първи май — pierwszy maja
- май** (2) *adv.* Chyba, prawie, zdaje się, prawdopodobnie

Maybe in this case it is necessary to list this form a third time so that its third Polish meaning *prawie* corresponding to Bulgarian *почти* (‘almost’) is listed as well.

- май** (3) *adv.* prawie

A short look at the Explanatory Dictionary of Bulgarian [2] shows us the following two ways to describe homonymy.

1. when the forms are different parts of speech, the difference in meaning is shown by indexing the different meanings
 - малко**¹ *нарч.* ...в ограничено или недостатъчно количество...
 - малко**² *ср.* Наскоро родено или излюпено същество...
 or it is implied by listing the respective part of speech.
 - май** *м.* Петият месец на годината...
 - май** *част.* За изразяване на предположение....
2. when the forms belong to the same class, the different meanings are indexed
 - мина**¹ *ж.* ... рудник
 - мина**² *ж.* ... снаряд
 - мина**³ *ж.* ... израз на лицето

The usage of indexing for each meaning of a form (as in the above examples (2)) would allow the Bulgarian-Polish dictionary to be “inverted” and thus to obtain automatically a Polish-Bulgarian digital dictionary. Whenever a bilingual digital dictionary is being compiled, in the beginning the most common words/forms (parts of speech) are selected in a given digital corpus of L_1 language. Then this frequency dictionary is completed with the translated correspondences from L_2 language. We must mention here that besides frequency the forms may be selected according to a certain topic which contains them and which they describe. In other words the dictionary may be compiled according to topics (something

like topic and frequency). 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, coordinated by L. Dimitrova and V. Koseska, the first Bulgarian-Polish digital dictionaries are being developed. The experimental version of the Bulgarian–Polish electronic dictionary is prepared in WORD-format and consist approximately 20 thousand dictionary entries. This dictionary is used for creation of the lexical database (LDB) that is an entry point to the relational database (RDB) of the first Bulgarian-Polish online dictionary [9]. We remark here that the suggested LDB structure of Bulgarian-Polish dictionary entry is suitable for automated generation of a Polish-Bulgarian dictionary entry. The structure of the LDB allows synchronized and unified representation of the information for Bulgarian and Polish, which is a step towards the creation of online Polish-Bulgarian dictionary in the future.

4 Some comparative remarks on the classifiers of the verbs

The comparison of the Bulgarian and Polish material [7] requires an explanation, which is important for the part-of-speech classifiers in the dictionary entries of the cited bilingual electronic dictionary. It is a common practice to list as a headword in the dictionary entries the infinitive of the verb. In Bulgarian the infinitive has disappeared and has been functionally replaced by the *da*-construction, which connects the particle *da* to the present tense forms. In this respect Bulgarian is more similar to other Balkan languages (Modern Greek, for example), but differs from Polish where the infinitive is preserved. This is an important example for the requirement of distinguishing a form from its function and meaning. The present tense form in this case does not have “present tense”-meaning. In the Bulgarian verb entries it is accepted to list as headword the 1st person singular form of the present tense.

One of the important classifiers of the verbal form which must be included in the dictionary entry refers to the transitivity or intransitivity of the verb. In our opinion the tendency of including more classifiers in the dictionary entry which we consistently follow, leads us to confirm the necessity of a classifier reflecting transitivity or intransitivity of the verb [8]. It is a different question what this classifier should reflect. According to the tradition in the older Bulgarian and Polish grammars, transitivity and intransitivity used to be considered as a phenomenon related to the voice of the verb (active or passive). In Polish and Bulgarian the verbs which form the passive participles are called transitive. They stand in contrast to the intransitive verbs which do not form such participles. A fact which we must stress here is that the Polish transitive verbs are always followed by the accusative case of nouns or adjectives. This fact is important for the comparison of the dictionary entries in Polish and Bulgarian, because Bulgarian lacks a nominal declination, while Polish is a typical synthetic language. The classifier “aspect” of a verb is universally accepted. However we must stress also that the “aspect” classifier is obligatory in the dictionary entry for a Slavic language. The aspect in Slavic languages is a well-formed grammatical category whose meaning expresses events — perfective aspect, and states — imperfective aspect, where we interpret “event” and “state” as described in the net description of temporality in a natural language

[14], [16]. On aspect and the problems of its classification see [12] for an overview of the different interpretation of aspect in the linguistic schools and the treatment of this category as word-forming, morphological, lexico-grammatical, grammatical and semantical. We must stress that the connection of the “aspect” category to temporality depends on the interpretation of “aspect” category. If we assume that “aspect” is a semantic category, the question about its relation to the semantic category “temporality” is inevitable. According to some linguists, “aspect cannot be treated separately from tense” [10], according to others the tenses are meanings independent from the meaning of the “aspect” of the verbal form [1]. Based on Bulgarian language material we see how important are the aspectual-temporal relation in the language. This leads us to the conclusion that the forms and meanings of time, especially with respect to Bulgarian, are a key problem that must affect the dictionary entry in every bilingual dictionary, which contains Bulgarian. It must be stressed that the Bulgarian language differs typologically from the other Slavic languages. It is an analytic language, and not synthetic (like the rest of the Slavic languages), has no cases in its nominal system (except some vestiges of vocative), but has many tense forms as well as well-formed category “aspect”. In this respect Bulgarian resembles a lot more English or Romance languages (French or Italian) than the other Slavic languages. In other words, the “aspect” problem opens the question about the “temporal” classifier in the dictionary entry: whether to include a “temporal” classifier and how to present it. This question must be answered in more detail later.

5 Suggestions

- Our suggestions can be grouped around the mode of form classification and the mode of writing the meanings of verb tense forms (two types with exact definition that can be “translated” in a formal language, for example, Petri nets). We take a step back so to say from the “form-meaning” principle and limit ourselves to the “form-form” principle in bilingual dictionaries.
- We suggest the headword form in the dictionary entry of the digital dictionary to be indexed according to the number of meanings, and each different meaning to be related unambiguously to the form. In this manner most meanings of the form can be encompassed. Such a description might require more classifiers but it is obvious that the greater number of classifiers provides a more adequate translation correspondence.
- We plan to use the CONCEDE model for dictionary encoding that respects the guidelines of the Text Encoding Initiative (TEI) Dictionary Working Group. The CONCEDE project [18], supported by the EC under INCO-Copernicus program, developed a formal model for LDBs. The LDBs using a common tagset for the six Central and East European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene were developed in accordance with the guidelines of the TEI Dictionary Working Group. In the framework of the project the first LDB for Bulgarian, based on encoding standards established by the TEI, was developed.

6 Bulgarian experience

Traditional grammatical classifications for Bulgarian

Traditional Bulgarian grammar for instance recognizes three main grammatical classifications:

- Semantic-grammatical – depending on the most general common meaning and on the grammatical properties words are ordered in classes, called parts of speech:
 - Nouns (a general terminological meaning of objects with common grammatical categories – gender, number, definiteness/indefiniteness),
 - Adjectives (have something in common in their lexical meaning, which is “indication, property, quality” of an object,
 - Verbs (common lexical meaning is “action or state” of a person/objects with common grammatical categories “tense”, “person”, “number”, “mood”, “voice”),
 - Numeral,
 - Pronouns,
 - Adverbs
 - Prepositions,
 - Conjunctions,
 - Interjections,
 - Particles,
- Morphological classification – according to the criterion “Open-class words or closed-class words”:
 - Open class words are nouns, adjectives, numerals, pronouns and verbs,
 - Closed class words are adverbs, prepositions, conjunctions, interjections and particles.
- Syntactic (functional) classification – depending on whether the word functions independently in the sentence or not:
 - Independent are nouns, adjectives, numerals, pronouns, verbs, and adverbs,
 - Dependent are prepositions, conjunctions, and particles. The interjections are excluded.

Lexical specifications for Bulgarian in MULTEXT-East

The semantic-grammatical classification of the Bulgarian wordforms was used during the development of lexical specifications for the Bulgarian language in the EC project MULTEXT-East [3], [4]. In the MULTEXT-East project multilingual parallel (Orwell’s 1984) and comparable (fiction and newspapers) corpora for six East-European languages - Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene - were developed and a lexicon was compiled for each corpus and language.

The lexicons have been prepared in the form of lexical lists where each line contains one entry in the following form:

word-form <tab> lemma <tab> morphosyntactic description

Morphosyntactic description (MSD) contains encoding lexical specifications of the corresponding word-form (“word-form” represents an inflected form of the lemma). When the the wordform (inflected form) coincides with its main form (lemma), then the entry “lemma” is replaced by “=”.

The MULTEXT-East project has provided harmonised lexical specifications for the six East-European MTE languages and English. The specifications are presented as sets of attribute-values, with their corresponding codes used to mark them in the lexicons. The core features were determined (these features are shared by the most of the languages) and this provided the comparability of the information encoded in the lexicons across the MULTEXT-East languages. Except these “general properties” so-called language-specific features were defined, which describe language-specific morphosyntactic phenomena.

Bulgarian MSD

Here we shall briefly present the Bulgarian wordform MSD because these can provide useful information about digital bilingual Bulgarian-lang2 (digital bilingual dictionaries with Bulgarian language) as possible classifiers in the dictionary entry in regard to applications of digital dictionaries in machine translation systems, e-learning, etc.

MSD is defined as a linear string of symbols, representing the morphosyntactic descriptions, the positions of a string are numbered 0, 1, 2, etc. in the following way:

- the symbol at position 0 encodes part of speech;
- each symbol at position 1, 2, n, encodes the value of one attribute (person, gender, number, etc.);
- if an attribute does not apply, the position in the string contains a hyphen “-”.

Some examples of Bulgarian MSDs:

барабан = Ncms-n (Noun, common, masculine, singular, no-definit)
 барабани барабан Ncmp-n (Noun, common, masculine, plural, no-definit)
 барабани барабаня Vmia2s (Verb, main, indicative, aorist, 2nd person, singular)
 барабани барабаня Vmia3s (Verb, main, indicative, aorist, 3rd person, singular)
 барабани барабаня Vmip3s (Verb, main, indicative, present, 3rd person, singul)
 барабани барабаня Vmm-2s (Verb, main, imperative, 2nd person, singular)

май = Ncms-n (Noun, common, masculine, singular, no-definiteness)
 май = Qgs (Particle, general, simple)
 май мая Vmm-2s (Verb, main, imperative, 2² person, singular)

мина = Ncfs-n (Noun, common, feminine, singular, no-definiteness)
 мина = Ncft (Noun, common, feminine, count)

малки малко Ncnp-n (Noun, common, neutral, plural, no-definiteness)
 малки малък A--p-n (Adjective, plural, no-definiteness)
 малките малко Ncnp-y (Noun, common, neutral, plural, yes full_article)
 малките малък A--p-y (Adjective, plural, yes full_article)

7 Examples of machine translation

Let us have a look at some examples of machine translation, randomly picked from a web-page with an original text in the English language, which offers translation to Bulgarian, Polish and other languages. The lack of morphosyntactic descriptions that contain encoded (according to the standard) lexical specifications of the word-forms and the lack of adequate classifiers (or any classifiers) in the database (or in the digital dictionaries), used in the machine translation system, leads to the following translation mismatches:

First example

Original English text:

His play/direct partnership with the Scottish Chamber Orchestra has been particularly fruitful, and as well as touring extensively with the orchestra *he has recorded a disc featuring Mozart's G major and D minor piano concertos.*

Machine translation in Bulgarian:

Неговата игра/преки партньорство с шотландски камерен ансамбъл е било особено ползотворно, а както и още по обстойно с оркестър *той е записано диск, с участието на Моцарт G големи и малки D пиано concertos.*

Comment:

(For the sake of comparison — English translation (as far as it is possible) of the Bulgarian text:

His game/direct partnership with a Scottish Chamber Orchestra has been particularly beneficial, and as well as more extensively with an orchestra *he was recorded a disc with the participation of Mozart G major and minor D piano concertos.*)

The errors in the machine translation of the sentences in the examples can be grouped as follows:

first, wrong choice of lexical meaning for the translation:

play = изпълнение ↔ игра = game

direct = ръководи, дирижира ↔ пряк = direct, straight; immediate

fruitful = плодотворно ↔ ползотворно = beneficial

featuring = включвайки ↔ участието и тхе партиципацион

second, lack of concordance between pronoun (as subject) and the verb form in the sentence:

he | той (pronoun, **masculine**)

recorded | записано (participle, **neutral**).

Machine translation in Polish:

Jego *grać / bezpośredniej współpracy* ze Scottish Chamber Orkiestra była szczególnie owocna, jak również szerokie tournée z orkiestrą *ma zapisane dysk* zawierający Mozarta G- dur i d – moll koncerty fortepianowe.

Comment:

The errors in this sentence are:

play is translated as a verb infinitive due to lack of classifiers, in this case the English *play* is a noun, not a verb.

ma zapisane — rodzaj nijaki is related to *dysk* — rodzaj męski the participle *zapisane* is neutrum and is not in accordance with the masculine noun *dysk*.

Second example

Original English text:

Piotr Anderszewski was born in Warsaw to *Polish-Hungarian parents*.

Machine translation in Bulgarian:

Пьотр Anderszewski е роден във Варшава с полския-унгарски родители.

(For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: Piotr Anderszewski was born in Warsaw with the Polish-Hungarian parents.)

Comment:

Lack of concordance between qualifier and word that it qualify in the translation of *Polish-Hungarian parents* | с полския-унгарски родители.

Machine translation in Polish:

Anderszewski urodził się w *Warszawa* – *Węgier do Polski rodziców*.

Comment:

The error here is triggered by the preposition *to*, to which only one meaning is given (from... Hungary to Poland). The English phrase *Polish-Hungarian parents* is not quite logical. Rather it should say “parts of Polish and Hungarian origin” or “Hungarian mother and Polish father”.

Furthermore, *Warszawa* instead of *Warszawie* — lack of casus locativus form.

The errors in this sentence are:

“play” is translated as a verb infinitive due to lack of classifiers, in this case the English “play” is a noun, not a verb.

Third example

Original English text:

An exclusive artist with Virgin Classics since 2000, Anderszewski’s first disc on the Virgin label was Beethoven’s Diabelli Variations, a work which had already fascinated him for a decade. An exclusive artist with Virgin Classics since 2000, Anderszewski’s first disc on the Virgin label was Beethoven’s, a work which had already fascinated him for a decade.

Machine translation in Bulgarian:

Един изключителен артист с Вирджински класика от 2000 г. насам, Anderszewski първия диск на Богородица етикет е на Бетовен Diabelli варианти за работа, която вече е очарован му за едно десетилетие.

Comment:

(For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: One exceptional artist with Virginia classic since 2000, Anderszewski first disc of Virgin Mary label is of Beethoven Diabelli work versions, which is already fascinated to him for a decade.) His strong identification with this work went

on to become the subject of a film by Bruno Monsaingeon (creator of documentaries on Sviatoslav Richter and Glenn Gould).

Machine translation in Polish:

Artysta na wyłączność z Virgin Classics od 2000 roku, Anderszewski pierwszy dysk na etykiecie Dziewicy było Beethovena Diabellego wariacje na pracę, która fascynowała go już od dekady.

Comment:

1. Casus genetivus for Anderszewski in the sentence is missing
2. “było” is neutrum and is not in accordance with the masculine noun “dysk”.
3. “work” is translated as a “prace”, the right translation is “tvorba”
4. in the phrase “wariacje na pracę, która fascynowała go” predicate is missing, correct: “jest to dzieło, albo jest to utwór, który go fascynował...”

Fourth example

Original English text:

The 2008-09 season will see Anderszewski *giving recitals at (as points at)* Carnegie Hall, *Chicago's Symphony Center (Chicago of the Symphony Center)*, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London.

Machine translation in Bulgarian:

В сезон 2008-09 ще видите Anderszewski *като точки в (as points at)* Карнеги Хол, *Чикаго на Симфония център (Chicago of the Symphony Center)*, Уолт Дисни Концертната зала в Лос Анджелис и Роял Фестивал Хол, Лондон.

Comment:

For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: During the 2008–09 season you will see Anderszewski as points at Carnegie Hall, Chicago of the Symphony Center, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London.

Machine translation in Polish:

W sezonie 2008–09 *będzie zobaczyc* Anderszewski podając motywów w Carnegie Hall, Chicago of the Symphony Center, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London.

Comment:

In the Polish translation “można” is missing, correct *można będzie zobaczyć*. “Motyw” is not correctly translated, “recital” is meant instead.

Fifth example

Original English text:

Currently he lives in Paris and Lisbon.

Machine translation in Bulgarian:

Currently he lives in Paris and Lisbon.

В момента той живее в Париж и Лисабон.

“Successful” translation correspondences.

Machine translation in Polish:

Aktualnie mieszka w Paryżu i *Lisbona* (correct *w Lizbonie*).

Comment:

Casus locativus for *Lisbon* in the sentence is also missing.

Briefly, in Polish we observe the following mistakes:

- wrong gender,
- lack of cases,
- incorrect translation of tenses – see above the lack of “można”,
- incorrectly translated prepositions,
- incorrect translation of lexical meanings (motyv — recital).

There is not a single correctly translated sentence in the Polish text, in contrast to Bulgarian, but that is due to the analytical character of English and Bulgarian, whereas the Polish cases pose an additional difficulty to the translation software.

8 Conclusion

The dictionary entry classifiers must reflect the specifics of the compared languages, for example the transitivity/intransitivity classifier is important for the syntax of both languages, but is much more important on the morphologic-syntactic level for Polish, a synthetic language, in contrast to Bulgarian, an analytic language. As mentioned before, the Polish transitive verbs require an accusative case for their object.

We must also distinguish between forms and the meanings of the forms in the dictionary entries. In traditional grammatical descriptions this distinction is missing, which creates intolerable errors in the description of the respective language. This is especially important for the aspect characteristic of the verbs in Slavic languages, where the category “aspect” is not only semantic but also grammatical.

We must stress again that we should not fear the greater quantity of dictionary entry classifiers in the electronic dictionary. On the contrary, this is an advantage of the electronic over the printed dictionary. The increase of the number of classifiers of the headwords in the entry will make machine translation more adequate and enrich electronic dictionaries. A dictionary with more classifiers will be significantly more useful to the user. We believe that it is necessary to establish a possibility to obtain the inverse dictionary automatically. With traditional bilingual dictionaries this is impossible because of the polysemy of forms. Using the contemporary process theory (Petri nets theory) we suggest that dictionary entries related to time in a natural language render the content as well as the form. The content must reflect the main elements of time: the event, the state and the configuration of events and states (see above *Example 1 and 2*; [16]).

Bibliography

- [1] Андрейчин, Л. (1944). *Основна българска граматика*. София. (In Bulgarian).

- [2] Andreichin, L., Georgiev, L., Ilchev, St., Kostov, N., Lekov, I., Stoikov, St., Todorov, Tsv. (1997) Bulgarian Explanatory Dictionary. 4th revised edition, prepared by D. G. Popov. Nauka i Izkuvstvo Publishing House, Sofia. (In Bulgarian).
- [3] Dimitrova, L. (1998). Lexical Resource Standards and Bulgarian Language. In: *International Journal Information Theories & Applications*, Vol. 5, No. 1, 27–34.
- [4] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada, 315–319.
- [5] 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, 25–34.
- [6] Dimitrova, L., Koseska-Toszewa, V. (2008). Some Problems in Multilingual Digital Dictionaries. In: *International Journal Études Cognitives*, Vol. 8, SOW, Warszawa, 237–255.
- [7] Dimitrova, L., Koseska-Toszewa, V. (2009). Bulgarian-Polish Corpus. In *International Journal Cognitive Studies – Études Cognitives*, Vol. 9, SOW, Warszawa, (In: this volume).
- [8] Dimitrova, L., Koseska-Toszewa, V., Satoła-Staškowiak, J. (2009). Towards a Unification of the Classifiers in Dictionary Entries. In: R. Garabík (Ed.), *Metalinguage and Encoding scheme Design for Digital Lexicography*. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 48–58.
- [9] Dimitrova, L., Panova, R. Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: R. Garabík (Ed.), *Metalinguage and Encoding scheme Design for Digital Lexicography*. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 2009, 36–47.
- [10] Иванчев, С. (1971). *Проблеми на аспектиалността в славянските езици*. София. (In Bulgarian)
- [11] Koseska-Toszewa, V. (2006). Bułgarsko-polska gramatyka konfrontatywna, t. VII. Semantyczna kategoria czasu. SOW, Warszawa.
- [12] Koseska-Toszewa, V. (forthcoming). Form, its meaning, and dictionary entries.
- [13] Koseska-Toszewa, V., Mazurkiewicz, A. (1988). *Net representation of sentences in natural languages*, Advances in Petri Nets, LNCS 340, Springer Verlag, 249–259.
- [14] Koseska, V., Mazurkiewicz, A. (2009). Net-Based Description of Modality in Natural Language (on the Example of Conditional Modality). In: V. Shyrovkov, L. Dimitrova (Eds.), *Organization and Development of Digital Lexical Resources*. Proceedings of the MONDILEX Second Open Workshop, Kiev, 2–3 February 2009. 98–105.
- [15] Koseska-Toszewa, V., Roszko, R. (2008). Remarks on Classification of Parts of Speech and Classifiers in an Electronic Dictionary. In: L. Iomdin, L. Dimitrova (Eds.), *Lexicographic Tools and Techniques*. Proceedings of the MONDILEX First Open Workshop, Moscow, Russia, 3–4 October 2008, 80–88.

- [16] Mazurkiewicz, A. (2008). A Formal Description of Temporality (Petri net approach). In: L. Iomdin, L. Dimitrova (Eds.). *Lexicographic Tools and Techniques*. Proceedings of the MONDILEX First Open Workshop, Moscow, Russia, 3–4 October 2008, 98–108.
- [17] Sławski F., (1987). Podręczny słownik Bułgarsko-Polski z suplementem. Warszawa.
- [18] CONCEDE: <http://www.itri.brighton.ac.uk/projects/concede/>

LUDMILA DIMITROVA¹
VIOLETTA KOSESKA-TOSZEWA²

¹Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

²Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland

BULGARIAN-POLISH CORPUS

Abstract. The paper shortly describes the first Bulgarian-Polish digital corpus. The corpus is collected with the main purpose to ensure the selection of the entries for the first experimental electronic Bulgarian-Polish dictionary. The texts were collected concurrently and do not have a connection with national monolingual or other corpora. This bilingual corpus supports the lexical database of the first experimental online Bulgarian-Polish dictionary. The use of digital multilingual corpora in contrastive studies is briefly discussed.

Keywords: multilingual electronic corpus, parallel and comparable corpora, corpus annotation, bilingual digital dictionary, Bulgarian, Polish.

1 Introduction

The great achievements of the information technologies offer wide possibilities to natural language processing, especially for developing and using of digital corpora, mono- and multilingual. Here we want to mention some well-known multilingual corpora that were created in recent decades in the field of corpus linguistics, such as the MULTEXT corpus [9], initially in seven West European languages (Dutch, English, French, German, Italian, Spanish and Swedish, with more in later editions, including Bambara, Catalan, Kikongo, Occitan and Swahili); the MULTEXT-East annotated parallel and comparable corpus [2], initially in six Central and East European languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian, plus English as a “hub” language, in later editions including Croatian, Lithuanian, Resian¹, Russian and Serbian); the ParaSol, a parallel and aligned corpus of Slavic and other languages (so-called Regensburg Parallel Corpus) [13] Italian-German parallel corpus, a collection of legal and administrative documents written in Italian and German, due to the equal status of the both languages in South Tyrol [7]; Hong Kong bilingual parallel English-Chinese corpus of legal and documentary texts [6], etc.

¹ Resian is a distinct dialect of Slovenian spoken in the valley Resia in Italy, close to the border with Slovenia.

MULTEXT-East (MTE for short) is an extension of the project MULTEXT, one of the largest EU projects in the domain of the language engineering prepared useful language tools and resources. The MTE project has developed a multilingual corpus, in which three languages: Bulgarian, Czech and Slovene, belong to the Slavic group.

The MTE model for corpora design and development is being used in the design of the first Bulgarian-Polish corpus.

2 Multilingual Corpus and Contrastive Studies

Parallel corpora are bilingual in the least and this fact distinguishes them fundamentally from monolingual corpora. The language material collected in parallel corpora must be comparable. Here we would like to pay special attention to a couple of compulsory rules which parallel corpora texts have to obey (unlike monolingual corpora texts).

1. Language material in parallel corpora, unlike the one in monolingual corpora, has to be at the synchronous level and must reflect the current state of the two (or more) languages. The synchronous level of language material should not be mixed up with its diachronic level. Mixing of the two levels — synchronous and diachronic — leads to wrong conclusions due to the comparison of incomparable language facts. In order to work with texts that are comparable, we have to select works by the same authors, for example Stefan Žeromski, Ryszard Kapuściński (authors who worked in 20 century).
2. Keeping in mind the richness and diversity of natural languages, we point out that the selection of texts in a parallel corpus is essential, especially for linguistic purposes. In order to be able to draw representative facts about the compared languages, the texts must be of different linguistic styles and to include:
 - selected literary works — it is preferable that they be written by world-famous representatives in the literature in both (or all) languages;
 - works in children's literature, the primary lexica of which is selected so that children can understand it — texts in live language, intended for children: books, short stories and fairy tales;
 - texts in administrative style — such are texts of varied character, for example, protocols from European Parliament sessions, programs of the European Commission, texts from electronic newspapers and journals that reflect current events in the European Union (as the language used in them is contemporary), professional terminology of the field of international law or ecology (laws and regulations), etc.

The language material in the Bulgarian-Polish parallel corpus belongs to different styles. We include literary works (paying special attention to children's literature) as well as texts in administrative style.

3. The aforementioned affects commonly accepted methodological requirements related to comparative linguistic studies whose theoretical bases are not known enough yet.

- 3.1. For a long time, contrastive linguistics was the “Cinderella” of common theoretical linguistics. It was accepted that the comparison of two or more languages is a task for the so-called applied contrastive linguistics, in which a given language is described through the means of another. This method pointed out some typical phenomena in one of the languages, while ignoring the description of the other. Such type of linguistic developments were used in translation and teaching a language through another, as well as in machine translation, but with rather weak results, especially from the point of view of the translation quality.

Applied comparative studies were not satisfactory, as they included descriptions of only some of the facts of the studied languages. For this reason, in the 70s, applied contrastive linguistics was highly criticized by representatives of structuralism in linguistics. Unlike applied contrastive linguistics, however, theoretical contrastive linguistics developed semantics in natural languages through contrasting two or more languages in the direction from meaning to language form, rather than from a language form in one language to a language form in another, as was the tradition in applied linguistics.

- 3.2. The postulate to create a semantic interlanguage, which would be independent from the contrasted languages and through which the study and comparison of two or more languages will give results equal to both languages, was accomplished for the first time in the multi-volume academic Bulgarian-Polish (and respectively Polish-Bulgarian) contrastive grammar [10], [11].

The semantic-cognitive method for language comparison presented in this grammar allows us to select language material in parallel corpora in such a way as to eliminate the possibility of comparison of incomparable language facts.

- 3.3. We acknowledge the disadvantages of comparison of language material from translation literature. In order to eliminate some of them, in our parallel corpus we focused on translations from Bulgarian to Polish as well as translations from Polish to Bulgarian. We chose and processed translation materials from other languages in Bulgarian as well as Polish, for example, from English, French, etc.

The thus-selected language material is relatively complete and can reflect well the contemporary state of the two studied languages, which makes the parallel Bulgarian-Polish corpus adequate without making it too large by quantity of word-forms.

In annotating our parallel corpus, we will keep in mind the fact that, if we want, for example, to describe the category definiteness in Bulgarian in which this category has morphological character (definite article), while the same content in Polish is expressed lexically, not morphologically. Such an approach is essential for human as well as machine translation [4].

- 3.4. In the literature on the subject, traditionally there was a distinction between the morphological and syntactic level of the language, without a comparison with its lexical level. This did not allow the complex study of semantic

language phenomena, which could be described only through an approach of “meaning of the word-form” from one language to the other, rather than “word-form” in one language to “word-form” in the other.

This problem is broader and applies also to classifiers in bi- and multilingual dictionaries, in which traditions in classification of dictionary entries are not satisfactory.

We hope that the parallel Bulgarian-Polish corpus will be of great use in machine translation in both languages as well as regular (human) translation. We believe that this will be possible thanks to the attention we pay to semantic phenomena in the two languages and we can achieve on the basis of theoretical comparative (contrastive) linguistic studies.

3 Description of the Bulgarian-Polish corpus

We have started to collect the first Bulgarian-Polish corpus in the framework of the joint research project “Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary” between Institute of Mathematics and Informatics, Bulgarian Academy of Sciences and Institute of Slavic Studies, Polish Academy of Sciences, coordinated by L. Dimitrova and V. Koseska. This bilingual corpus supports the lexical database (LDB) of the first experimental online Bulgarian-Polish dictionary [5].

The Bulgarian-Polish corpus consists of two corpora: a parallel and a comparable. All collected texts in the corpus are texts published in and distributed over the Internet and was downloaded from existing digital libraries.

A detailed description of the corpus is provided for clarification to the user. The description includes: language, author, title, words in the text, and if available, year of creation, publication place, year and publishing house, translator, year of translation, source and original format of the text, etc.

Currently the corpus contains about 5 mln wordforms, among them 3 mln in parallel texts, that represent mostly modern Bulgarian and Polish literature (the second part of the XXth century).

Microsoft Word was used to count words in the texts.

3.1 The Bulgarian-Polish parallel corpus

The structure of the parallel corpus groups texts according to content. Every group contains two parts (respectively three if the original language is different from the languages in the corpus).

The Bulgarian-Polish parallel corpus includes two parallel sub-corpora a *pure* and a *translated*:

1) A ***pure* Bulgarian-Polish parallel corpus** consists of original texts in Polish — literary works by Polish writers and their translation in Bulgarian, and original texts in Bulgarian — short stories by Bulgarian writers and their translation in Polish.

The description of literary work is as follows:

BG Bulgarian: Станислав Лем, *Solaris*. Translated by Андреана Радева. Отечество, София, 1980.

PL Polish: Stanisław Lem, *Solaris*. Wydawnictwo Literackie, Kraków, 1961.

About 60 925 Bulgarian words, 56 654 Polish words,
// EN: Stanislav Lem, *Solaris*.//

BG Bulgarian: Стефан Жеромски, *Пепелища* (част 1: глави 1, 2, 3). Translated by Димитър Икономов, 1956 г.

PL Polish: *Popioły* (part 1, chapters 1, 2, 3), (1902–1903). Editor Zbigniew Goliński, Czytelnik, Warszawa 1988.

About 24 843 Bulgarian words, 21 908 Polish words.
//EN: Stefan Żeromski, *The Ashes*.//

The description of short story:

BG Bulgarian: Емилиян Станев, *Лакомото мече*.

PL Polish: Emilijan Stanew, *Łakomy niedźwiadek*. Translated by Violetta Koseska. 688 Bulgarian words, 605 Polish words.

//EN: Emiliyan Stanev, *The Gluttonous Little Bear*.//

BG Bulgarian: Светослав Минков, *Сапунени мехури*.

PL Polish: Swetoslaw Minkow, *Wińki mydlane*. Translated by Violetta Koseska. 1267 Bulgarian words, 1104 Polish words.

//EN: Svetoslav Minkov, *Soap Bubbles*.//

2) A *translated Bulgarian-Polish parallel corpus* consists of texts in Bulgarian and in Polish of brochures of the EC, documents of the EU and the EU-Parliament, published in Internet; Bulgarian and Polish translations of literary works in third language (mainly English).

Some excerpts of the description of the bilingual translated parallel corpus follows:

Literary works:

BG Bulgarian: Антоан Дьо Сент Егзюпери, *Малкият принц*. 11 936 words.

PL Polish: Antoine de Saint-Exupery, *Mały Książę*. 11 911 words.

FR French: Antoine de St Exupery, *Le petit Prince*.

BG Bulgarian: Джордж Оруел, *1984*. 87235 words.

PL Polish: George Orwell, *1984*.

EN English: George Orwell, *1884*.

3.2 The Bulgarian-Polish comparable corpus

The Bulgarian-Polish comparable corpus includes texts in Bulgarian and Polish: excerpts from newspapers and textual documents, shown in internet, excerpts from several original fiction, novels or short stories, with the text sizes being comparable across the two languages. Some of the Bulgarian texts are annotated at “paragraph” and “sentence” levels, according to CES [8].

Literary works examples:

BG Bulgarian: Димитър Талев, *Железният светилник*, 1952.

Words: 126 801

//EN: Dimitar Talev, *The Iron Oil Lamp*. 1952.

BG Bulgarian: Димитър Талев, *Преспанските камбани*, 1954.

Words: 225 164

//EN: Dimitar Talev, *The Bells of Prespa*, 1954.

BG Bulgarian: Димитър Димов, *Тютюн*. Пето издание. Български писател, София, 1964.

Words: 11 5745.

//EN: Dimitar Dimov, *Tobacco*./

BG Bulgarian: Димитър Димов, *Осъдени души*. Български писател, София, 1970.

Words: 91 724.

//EN: Dimitar Dimov, *Doomed Souls*./

BG Bulgarian: Ж. Желев, *Фашизмът*. Университетско издателство “Св. Климент Охридски”, София, 1991.

Words: 93 706.

//EN: Zhelyu Zhelev, *The Fascism*./

PL Polish: Stanisław Lem, *Dzienniki gwiazdowe*. 1957.

Words: vol. I: 86 829, vol. II: 78 019.

//EN: Stanislaw Lem, *The Star Diaries*//

PL Polish: Ryszard Kapuściński, *Imperium*. 1993.

Words: 4 565

//EN: Ryszard Kapuściński, *Imperium*//

3.3 Corpus annotation

Corpus annotation is the process of adding information, linguistic or structural, to a text corpus ([8], [12]). One common type of annotation is the addition of labels, tags that indicate the word class to which words in a text belong. This is the so-called part-of-speech tagging (or POS tagging). Apart from POS tagging, there are other types of annotation, e.g. structural annotation which corresponds to different structural levels of a corpus or text. Written texts contain a number of different structural forms — divisions. Novels have a complex hierarchy and are divided into parts and chapters, newspapers are divided into sections, reference works — into articles, etc. The most common division in this hierarchy is the paragraph.

Some texts in the ongoing version of the Bulgarian-Polish corpus are annotated at paragraph level. We use the standard markers — at the beginning of a paragraph `<p>` and `</p>` at the end of a paragraph.

This annotation allows texts in the two languages (Bulgarian/Polish and *vice versa*) to be aligned at paragraph level in order to produce aligned bilingual corpora. “Alignment” means the process of relating pairs of words, phrases, sentences or paragraphs in texts in different languages which are translation equivalent. One may say that “alignment” is a type of annotation performed over parallel corpora.

The <p> level allows us to draw a broader context in the two languages. This means that we get the opportunity — thanks to the broader context — to study more precisely the meanings of word-forms in both languages.

This approach is more correct — we are not comparing “word” with “word”, we compare word-forms in a broader context (level <p>), which allows us to obtain the word’s meaning.

Some examples of texts of the Bulgarian-Polish parallel corpus, marked at paragraph level follow.

(1) An excerpt of *The Ashes* (vol. 1, part 1 *In the forest*), a novel by Stefan Żeromski

Bulgarian:

<p>Кучетата млъкнаха. Веднага след това друг глас, по-близко до Рафал, отговори еднократно по същия начин.</p>

<p>Младият ловец лежа още малко на земята, позеленял от яд. Но после изведнъж скочи на крака, изтупа снега от себе си и потърси пушката в храстите. Избърса очи и скачайки като сърна през младите елички, полетя надолу.</p>

Polish:

<p>Psy ucięły. Zaraz potem drugi głos, bliższy Rafała, odpowiedział jednokrotnie tym samym sposobem.</p>

<p>Młody myśliwiec jeszcze przez czas pewien leżał na ziemi, pękając ze złości: Po chwili jednak zerwał się na równe nogi, strzepnął śnieg z siebie, odszukał w krzakach pojedynkę. Wytrzeł oczy i, na podobieństwo sarn skacząc przez choiny, pomknął na dół.</p>

//English:

The dogs got quiet. Immediately after another voice, closer to Rafal, answered once in the same way.

The young hunter lay a little longer on the ground, green with anger. But then he immediately jumped to his feet, brushed the snow off himself and looked for the gun in the bushes. He wiped off his eyes and, jumping like a deer through the young firs, hurled himself down.²

(2) Excerpt of texts of the EU document

http://europa.eu/abc/european_countries/languages/bulgarian/index_en.htm?_bg

<p>Европейци, единни в многообразието</p>

<p>Европейският съюз е семейство от демократични европейски страни, решени да работят заедно за мир и благоденствие. Той не е държава, която трябва да замести съществуващите държави, но не е и поредната международна организация. Всъщност Европейският съюз е уникален. Неговите държави-членки са учредили общи институции, на които са отстъпили част от суверенитета си, така че решенията по специфични въпроси от общ интерес да могат да се вземат по демократичен начин на европейско равнище.</p>

² Stefan Żeromski. *The Ashes*, vol.1, part 1: *In the forest*. Editor Zbigniew Goliński, Czytelnik, Warszawa, 1988, p. 15.

http://europa.eu/abc/european_countries/languages/english/index_en.htm?_en

<p>Europejczycy zjednoczeni w swojej różnorodności</p>

<p>Unia Europejska (UE) to rodzina demokratycznych państw europejskich, których celem jest wspólna praca na rzecz pokoju i dobrobytu. Nie jest państwem powstałym w miejsce istniejących krajów, jest jednocześnie czymś więcej niż inne organizacje międzynarodowe. Jest to organizacja jedyna w swoim rodzaju. Państwa Członkowskie przekazują stworzonym przez siebie wspólnym instytucjom część swoich kompetencji, tak aby decyzje w określonych sprawach będących przedmiotem wspólnego zainteresowania mogły być podejmowane w sposób demokratyczny na poziomie europejskim.</p>

http://europa.eu/abc/european_countries/languages/english/index_en.htm?_en

<p>Europeans united in diversity</p>

<p>The European Union (EU) is a family of democratic European countries, committed to working together for peace and prosperity. It is not a State intended to replace existing states, but it is more than just another international organization. The EU is, in fact, unique. Its Member States have set up common institutions to which they delegate some of their sovereignty so that decisions on specific matters of joint interest can be made democratically at European level. </p>

4 Conclusion

We want to note here that the parallel Bulgarian-Polish corpus will enrich and uncover some unstudied features of the two languages. The corpus will be useful to linguists-researchers for research purposes alike, for instance in contrastive studies of Bulgarian and Polish languages.

Besides, the bilingual corpus can be used in education, in schools as well as universities in foreign-language instruction.

Furthermore, the corpus has applications into development of LDB that supports a Bulgarian-Polish online dictionary. The advantage of processing a bilingual parallel corpus is to obtain context specific information about syntactic and semantic structures and usage of words in given language or languages.

Bibliography

- [1] Bulgarian Grammar. Тилков, Д., Стоянов, Ст., Попов, К. (Eds.) (1993). Граматика на съвременния български книжовен език. Том 2 / Морфология. Издателство на БАН. София. (In Bulgarian).
- [2] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada. 315–319.

- [3] Dimitrova, L., Koseska-Toszewa, V. (2008). Some Problems in Multilingual Digital Dictionaries. In: *International Journal Études Cognitives*. 8, SOW, 237–254.
- [4] Dimitrova, L., Koseska-Toszewa, V. (2008). The Significance of Entry Classifiers in Digital Dictionaries. In: L. Iomdin, L. Dimitrova (Eds.). *Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, Russia, 3–4 October 2008*, 89–97.
- [5] Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: R. Garabík (Ed.), *Metalinguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009*. 36–47. ISBN 978-5-9900813-6-9.
- [6] May F., Xu X. (2002). An evaluation of an online bilingual corpus for the self-learning of legal English. http://langbank.engl.polyu.edu.hk/corpus/bili_legal.html
- [7] Gamper, D. (1999). Primary Data Encoding of a Bilingual Corpus. <http://titus.uni-frankfurt.de/curric/gldv99/paper/gamper/gamperx.pdf>.
- [8] Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, 463–70.
- [9] Ide, N., and Véronis, J. (1994). Multext (multilingual tools and corpora). In COLING'94, pages 90–96, Kyoto, Japan.
- [10] Koseska-Toshewa, V., Gargov, G. (1990). Bãlgarsko-polska sãpostavitelna gramatika, t. 2, Semantičnata kategorija opredelenost / neopredelenost, Sofiya.
- [11] Koseska-Toszewa, V., Korytkowska, M., Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Warszawa: Wydawnictwo Akademickie Dialog.
- [12] Leech, G. (2004). Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. <http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm>
- [13] ParaSol corpus: http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/

DIANA BLAGOEVA

Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

ELECTRONIC CORPORA AND BULGARIAN NEW-WORD LEXICOGRAPHY

Abstract. The article substantiates the need to apply a corpus-based approach when describing neological lexis in a lexicographic way. It introduces the specialized lexicographic electronic corpus, used as an empirical basis for the compilation of the latest academic Bulgarian dictionary of new words. The procedure for semi-automatic extraction of neologisms from an electronic corpus applied in compiling the lexical frame of this dictionary is described in the article.

Key words: neologisms, new-word lexicography, corpus linguistics.

1 Introduction

The exceptional intensification of neologization processes in Bulgarian lexis in the period after 1989 brought forth the need for an opportune and adequate lexicographic recording of new lexical and phraseological units. This issue finds its partial solution only in the general and specialized Bulgarian dictionaries from the period discussed. Newly-coined and newly-borrowed lexical units, which are widely used and have been completely integrated within the system are subject to inclusion into monolingual dictionaries, and due to this a good deal of newly-formed units of a lower level of usualization have not yet been recorded lexicographically. Moreover, the efficient recording of neologisms in general dictionaries is hampered by the complex methods involved in publishing some lexicographic works (the academic Dictionary of Bulgarian Language in many volumes in particular). Yet, due to their specific nature, terminological and phraseological dictionaries, dictionaries of foreign words, abbreviation dictionaries, etc., could only encompass separate fields of the whole range of neological lexis.

An optimal method of lexicographically recording neologisms is the creation of differential neological dictionaries, which give an account of the innovative phenomena related to lexis for a specific space of time. The information these dictionaries supply is the basis on which theoretical research is carried out, it performs prognostic functions as regards the tendencies in the development of lexis; it is also significant for prescriptive activities.

A basic prerequisite for a detailed and systematic lexicographic description of neological lexis is the use of an empirical basis, which is as broad as possible. Updated Information Technology, various electronic resources (corpora, lexical databases, Internet archives, etc.) in particular, open up substantially new perspectives in this respect. Of course, conventional methods of collecting and processing lexicographical material have not lost their significance yet. The restricted use of these methods only, however, could not lead to satisfactory results. As the work of lexicographers from various Slavic countries shows (e.g. [14]), the combination of new and conventional methods of new-word lexicography is the only possible solution at this stage.

2 Empirical basis of the “Dictionary of New Words in Bulgarian (at the End of the 20th and the First Decade of the 21st c.)”

The latest academic neological Bulgarian dictionary – “Dictionary of New Words in Bulgarian (at the End of the 20th and the First Decade of the 21st c.)” [13], was compiled at the Department of Bulgarian Lexicology and Lexicography of the Institute of Bulgarian Language, as a sequel and a supplement to the one published in 2001, which was titled “New Words and Meanings Bulgarian Dictionary” [12]. The dictionary contains about 4300 newly-coined or newly-borrowed lexical units, 700 words, which have developed new meanings, more than 600 new compound names and terms and almost 150 new phraseological combinations.

When compiling the “Dictionary of New Words in Bulgarian”, a corpus-based lexicographic approach was applied. This approach involves the implementation of the basic part of lexicographic activities (in particular – the selection of units for the lexical frame, the study of the use of neologisms in contemporary written texts, the extraction of illustrative material, etc.) by means of using an adequately organized and balanced corpus of electronic texts. The corpus on the basis of which the “Dictionary of New Words in Bulgarian” was compiled is lexicographically specialized and was created at the Department of Bulgarian Lexicology and Lexicography of the Institute for Bulgarian Language (for more details see [4], [7])¹. This is a monolingual, structured, lemmatized, unannotated, open corpus of written texts, which contains more than 300 million words and includes more than 7600 electronic documents (digitalized versions of books or periodicals – newspapers, magazines, year-books, etc.).

The corpus consists of four sub-corpora, which have been set apart on the grounds of chronological characteristics:

1. A sub-corpus consisting of texts published in 19th c. This sub-corpus includes Bulgarian Revival Literature texts and periodicals.

¹ Since the beginning of 2009 the lexicographic corpus has been incorporated within the content of Bulgarian National Corpus (http://www.ibl.bas.bg/BGNC_bg.htm), which was constructed by means of combining electronic resources, developed at various departments of the Institute of Bulgarian Language at the Bulgarian Academy of Sciences [6].

2. A sub-corpus consisting of texts published at the beginning of 20th c. (1901) until 1944. This sub-corpus includes Contemporary Bulgarian Literature texts and periodicals from the period until the middle of 20th c.

3. A sub-corpus consisting of texts published between 1945 and 1989. The year of the last spelling reform in Bulgaria, which established the new rules of Contemporary Bulgarian Spelling, was arbitrarily adopted as the starting point of this period, and the end of the period is the year when significant socio-political and economic changes started in Bulgarian society at the end of 20th c. and the beginning of 21st c., a time when exceptional dynamics in Bulgarian language development processes were also noticed, more specifically in its lexical system.

4. A sub-corpus consisting of texts published from 1999 onwards. This corpus served as the main empirical basis for lexicographic description of the changes in the lexis of Bulgarian language at the end of 20th and the beginning of 21st c.

The sub-corpus consisting of the latest texts (published after 1990), includes more than 6500 electronic documents, which consist of more than 240 million words. Its content is presented in the following table:

Table 1. Contents of the sub-corpus of texts published after 1990

Sub-Corpus Material Types	Number of Electronic Documents	Number of Words (in millions)
Books (total)	1374	110
Bulgarian Literature	674	31
Translated Literature	700	79
Periodicals (total)	5289	132
Newspapers	4271	110
Magazines and Year-Books	1018	22

A basic requirement to the empirical material of lexicographic description is for this material to be exemplary, so that the objectivity and reliability of lexicographic solutions can be ensured [1]. The purpose, when selecting texts to be included within the neological sub-corpus, is to achieve a wide range of the styles and genres included. Texts have been selected of both the informative and fictional type, the proportion being approximately 80% of informative texts (scientific, popular science, documentary, memoir literature, political journalism, etc.) as against 20% of fictional texts (various genres of fiction, poetry, drama, etc.). The fact that informative (publicistic in particular) texts prevail is justified by the circumstance that neologisms acquire public acceptance and strengthen their position as regards use mostly by functioning on the level of the language of the Media. Additionally, the fact that the full text versions of periodicals are included allows for a wide range of genres to be covered in the publicistic style (reports, interviews, articles, etc.). The study of scientific literature of a wide range of topics makes it possible for new terminological vocabulary to be recorded. On the other hand, the inclusion of fictional texts allows for the range of neological phenomena studied to be expanded by the discovery of some original neologisms created by authors.

When selecting texts, the thematic criterion is applied as well, the aim being to encompass a maximum number of subjects. For instance, the scientific litera-

ture included covers the following branches of science²: Theory of Literature (65), History and Archaeology (50), Philosophy and Ethics (35), Political Science (14), Psychology (11), Linguistics (8), Economics (5), Sociology (5), as well as Computers and Informatics, Culture Studies, Law, Art History, Physics, Biology, Geology, etc. Some thematically specific (specialized and non-specialized) periodicals have been included, in the fields of: Art and Culture (23), Humanities (16), Economics (16), Computers (11), Leisure and Lifestyle (7), Health and Medicine (7), Law (6), as well as Sport, Music, Education, Humour, Hunting and Fishing, Apiculture, etc.

The Lexicographic Electronic Corpus of Bulgarian Language grows constantly, as new materials are added in order for a balance to be reached between its separate subdivisions and in order for efficiency to be ensured in tracing the innovative phenomena in lexis.

3 Procedure to semi-automatically extract neologisms from an electronic corpus

For the purposes of the lexicographic description of the latest Bulgarian lexis, a procedure was applied to semi-automatically extract neologisms from an electronic corpus [5]. This procedure includes the following steps:

1. Generating an alphabetical frequency list of word forms in the sub-corpus of texts published after 1990.
2. Generating a reference list of word forms on the basis of the following sources:
 - an alphabetical frequency list of word forms in a reference electronic corpus (which includes the first three of the sub-corpora described above, part of the Lexicographic Corpus of Bulgarian Language)
 - a specially created morphological database, which includes nearly 70 000 lemmas that cover the core part of Bulgarian lexis, and more than 700 000 forms, which were automatically generated out of these lemmas
 - a list of about 50 000 proper names (anthroponyms and toponyms)
3. Comparison, by means of a special computer programme (see Figure 1, p. 147), between the alphabetical frequency list of word forms in the sub-corpus with the latest texts and the reference list created, and generation of a list of units, which are not included in the reference list.
4. Manual processing of the generated list and eliminating the units, which are not neological in their nature: word forms, which were not recognized by the programme as forms of well-known words, proper names, words containing spelling mistakes and misprints, words containing foreign graphics, etc. The elimination of these units is carried out by the lexicographer on the basis of linguistic intuition, by juxtaposing them with the available Bulgarian language dictionaries and by checking their use in the Lexicographic Electronic Corpus and other sources.

² The digit in the brackets here and below indicates the number of titles.

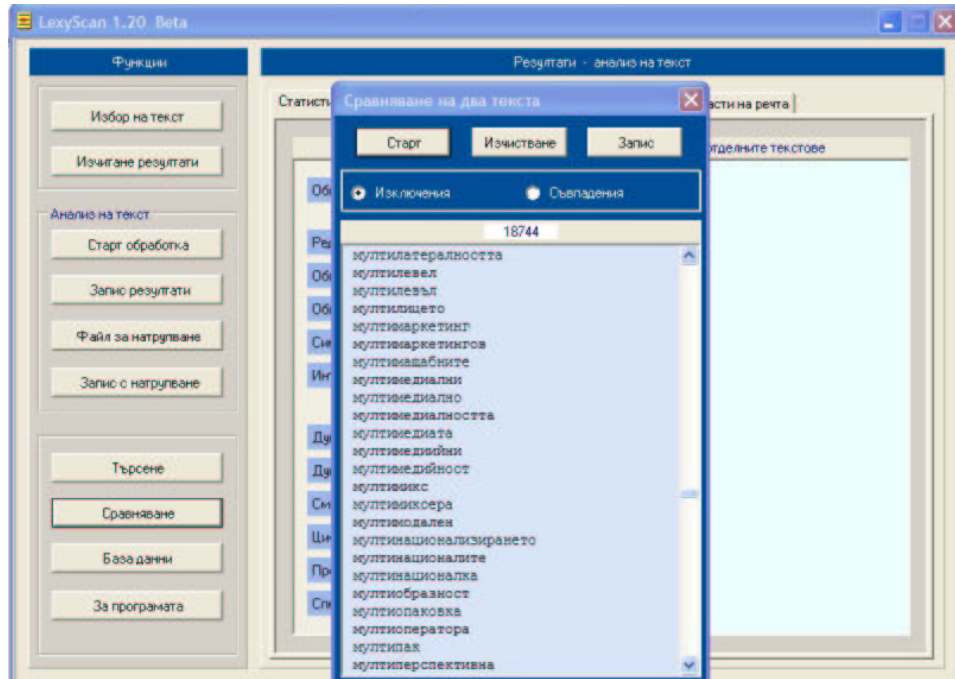


Fig. 1. Generating a list of word forms, which are not included in the reference list

An expanded list of about 30 000 neological units was created as a result of the procedure carried out and this list includes true neologisms, various types of occasionalisms, units newly created by authors, etc., which served as the basis on which the lexical frame of the “Dictionary of New Words in Bulgarian (at the End of the 20th and the First Decade of the 21st c.)” was formed.

Along with its indisputable advantages (fast processing of a huge array of texts, which is unattainable when excerpting manually), the method described, of semi-automatically extracting neologisms, has its obvious imperfections as well. Only lexical neologisms can be discovered by means of the procedure specified (word formation neologisms, newly-borrowed words), but not neosemantisms, new set phrases, terminological or other phrases. This is why it is necessary for the lexical frame of the neological dictionary to be expanded by manually excerpted units.

4 Optimizing the lexicographic description of neologisms by applying a corpus-based approach

The use of modern corpus-based approaches is a necessary condition in order for quality and objectivity to be achieved in lexicographic work. By studying large volumes of text arrays one can overcome any possible display of subjectivity and lack of organization, which is not impossible when collecting lexicographic material

by means of manual excerption. What is achieved is efficiency in examining the innovations in the lexical system, as well as greater thoroughness and reliability of their lexicographic description. The creation of a lexical frame for dictionaries of neologisms is facilitated. The greater quality of illustration of the neological phenomena described is ensured by means of quotations from authentic written texts.

The use of an exemplary empirical basis and the opportunity to apply tools for statistical analysis enable us to make well-grounded research decisions about various issues: when distinguishing between the different types of newly created lexical units (neologisms, occasionalisms, potential words), when hesitating with respect to spelling rules, when assessing the productivity of word-formation formants, when studying the words' semantics and their ability to combine, etc. The constant expansion of the empirical basis enables us to opportunely detect any possible changes in the status of newly created units (e.g. the process of usualization and activation, and as a result – the unit's shift from the periphery of the lexical system towards its centre).

Software tools, which are used to work with the corpus, enable us to analyze and mould (by using statistical methods) the contextual environment of lexical units and (when a reference corpus of a sufficient diachronic depth is used) to obtain systematized information regarding any possible changes in the contextual behaviour of the respective lexemes and their ability to combine. Changes of this kind are quite often evidence that the word has undergone some semantic development (which can be found in the noun *пространство* (space), for example, in newly coined phrases such as *публично пространство, медийно пространство, образователно пространство, религиозно пространство* and *интернет пространство, учеб пространство, информационно пространство*) or that new set phrases have been formed (for instance, *изборен туризъм, medien komfort, nakazatelen vot*, etc.).

The corpus-based approach (when the relevant diachronic depth is available, of the electronic corpus used) enables us to objectify and justify the neological status of individual units and to more or less accurately determine the chronology of their entering the language. For instance, the data in the Lexicographic Electronic Corpus indicate that a considerable part of the words and the phrases included in one of the latest neological Bulgarian dictionaries — [8], have in fact long been known in our language³, for example: *агентирам* (1979), *аеромобил* (1966), *бомбоубежище* (1981), *буржа* (1942), *голист*, *голистки* (1978), *воайор* (1982), *дефлектор* (1979), *земеподобен* (1971), *инцест* (1981), *килт* (1957), *колумбарий* (1971), *ледопад* (1971), *моторика* (1982), *полиграф* (1971), *прогнозист* (1971), *просвет* (1971), *психograma* (1979), *публична тайна* (1959), *ракетчик* (1973), *смъртник* (1956), *снегоход* (1972), *спиричуъл* (1979), *стрелжия* (1971), *унчестер* (1973), *футуристичен* (1979), *футуролог* (1972), etc. This shows that the psycholinguistic criterion of newness, when applied in the new-word lexicography, is not sufficiently reliable without the relevant methods used to verify the lexicographer's intuition.

³ The digit in the brackets indicates the year when the respective unit was initially recorded within the Lexicographic Corpus.

5 Conclusion

The complete automation of the extraction of various types of neological units from a text corpus is a task, which can find its solution in the more distant future. A necessary condition for this to happen is the presence of sufficiently reliable methods of automatic linguistic analysis, in particular, the automatic recognition of multiword lexical units, the automatic word sense disambiguation, etc. Bulgarian Computer Linguistics experts also work along these lines (for instance, see [2], [3], [10], [11], etc.); all results achieved (along with their other applications) will make the linguistic shaping of neology possible, as well as the development of algorithms for the automatic recognition and extraction of various types of neologisms.

Bibliography

- [1] Filipec, J. (1995). Teorie a praxe ednojazyčného slovníku vykladového. In: Manuál lexicografie. Praha, “H&H”, pages 14–49.
- [2] Koeva, S. (2007). Multi-Word Term Extraction for Bulgarian. In ACL 2007, Proceedings of the Workshop on Balto-Slavic NLP. Prague, pages 59–66.
- [3] Silberztein, M., Koeva, S. (2005). Bulgarian and English Semantic Dictionaries for the Purposes of Information Retrieval. In Computer Treatment of Slavic and East European Languages. Bratislava, “Veda”, pages 193–202.
- [4] Благоева, Д. (2008) Проблеми на изграждането на електронен корпус за лексикографски цели. In Lexikografie v kontextu informační společnosti. Praha, UJČ, 2008, pages 169–179.
- [5] Благоева, Д. (2008). Съвременни подходи в българската неография (проблеми и перспективи). – Български език, 2008, кн. 1, pages 5–14.
- [6] Благоева, Д., Коева, Св., Колковска, С. (2009). Български национален корпус. In Лексикографски преглед, 2009:10, pages 72–75.
- [7] Благоева, Д., Колковска, С. (2007). Електронен корпус за целите на “Речник на българския език” – състояние и перспективи. In Лексикографията и лексикологията в съвременния свят. Велико Търново, 2007, Знак’94, pages 277–286.
- [8] Бонджолова, В. (2003). *Неологичен речник (за периода 1998–2003)*. Велико Търново, “GABEROFF”, 2003.
- [9] Карева, О. М., Кочнев, В. В. (2006). Интерактивна база данни “нео-LEX”: опыт алгоритмизации лексикографической обработки неологизмов. In Русская академическая неография (к 40-летию научного направления). Санкт Петербург, “Наука”.
- [10] Коева, Св. (2004). Съвременни езикови технологии (приложения и перспективи). In Закони на/за езика. София, “Хейзъл”, pages 111–157.
- [11] Кукова, Хр. (2006). Подходи за автоматично отстраняване на семантична многозначност. In Български език, LIII:2, pages 75–84.
- [12] Пернишка, Е., Благоева, Д., Колковска, С. (2001). *Речник на новите думи и значения в българския език*. София, “Наука и изкуство”, 309.
- [13] Пернишка, Е., Благоева, Д., Колковска, С. (2009). *Речник на новите думи в българския език (от края на XX в. и първото десетилетие на XXII в.)*. София, “Наука и изкуство” (in print).

- [14] Рангелова, А. (1996) Изграждане на база за описание на посттоталитарната лексика в чешкия език. In *Езикът на тоталитарното и посттоталитарното общество*. София, "Проходка и Качармазов", pages 155–159.

DOROTA KOPCIŃSKA¹
JADWIGA LINDE-USIEKNIEWICZ¹

¹ University of Warsaw, Faculty of Polish, Warsaw, Poland

MATCHING FRAMENET FRAMES WITH POLISH SENSE DIVISIONS: THE CASE OF *JECHAĆ*

Abstract. The paper presents a proposal of a novel sense division for the Polish lexeme *jechać*, arising from an attempt to match the dictionary sense with an appropriate FrameNet frame. Polish lexicography proposes a single sense for all cases of humans going on land by some means of transportation, in order to contrast the verb *jechać* with the verb *iść* ‘to go on foot’. However, all factual examples of use of the verb *jechać* can easily be matched with one of three relevant frames: “Operate_vehicle”, “Ride_vehicle” and “Travel”, at least from the wider context. The paper shows how a parallel sense division can be established using purely linguistic, i.e. syntactic and semantic, criteria.

Keywords: Frame, frame semantics, sense division, Polish motion verbs.

The present paper discusses some of the issues that have arisen during the realization of the project RAMKI: **R**ygorystyczna **a**plikacja **m**etody **k**ognitywno-**i**nterpretacyjnej (ram interpretacyjnych) do opisu polszczyzny¹ launched in 2007. The project was founded by the Polish Ministry of Science and Higher Education, grant nr MNiSW N104 024 32/1840 (<http://www.ramki.uw.edu.pl/>). The project, referred to as RAMKI throughout the present text.

The project was designed as a pilot study concerning about 200 Polish verbal lexical units (LU), where each unit roughly corresponded to a single dictionary sense. The lexical units chosen for the project simultaneously met two criteria: the first one was the frequency of appearance in the chosen corpus, the second was their preliminary semantic correspondence to lexical units already described in other FrameNets [7].

The corpus chosen for the project was the IPI PAN Corpus [13]. The pilot study differed from the FrameNets established for other languages, i.e. the original Berkeley FrameNet Project (<http://framenet.icsi.berkeley.edu/>) for English; SALSA (the Saarbrücken Lexical Semantics Acquisition Project, <http://www.coli.>

¹ Lit. Little Frames: A rigorous application of the cognitive interpretative method (interpretative frames) to Polish.

uni-ssarland.de/projects/salsa/) for German; Spanish FrameNet (<http://gemini.uab.es:9080/SFNsite>) and Japanese FrameNet (<http://jfn.st.hc.keio.ac.jp/>). Since Polish is not a widely studied language and the resources are scarce, instead of starting from raw corpus data, we decided to evaluate the degree to which existing syntactic and lexicographic descriptions of Polish lexical units could be incorporated into the RAMKI project. Available syntactic descriptions comprised [12] and [14]. For preliminary lexicographic presentation two dictionaries were used. The lexicographic source of choice was [1], complemented by [4]. The rationale for using lexicographic description of the units studied and not their semantic analyses were given in [7]:

“We believe that the lexicographical practice, the more pre-theoretical or counter-theoretical the better, offers some valid information about the usage which can be easily translated into appropriate frames and/or frame elements. Lexicographers aim at giving information about the actual use and are rarely concerned about fine-honed distinction between internal-linguistic and extralinguistic phenomena. Thus the actual lexicographic descriptions appear to reflect fairly adequately the real-life language use and the speakers’ awareness of their own language”.

Because of the differences in design between the RAMKI project and other FrameNets, a specific, in-house software was devised. It allowed linking established senses to existing (previously implemented) frames, and providing examples of appropriate usage. The examples were chosen to show as many frame elements as possible, and were annotated both in terms of frame elements, and in terms of syntactic elements. For the latter syntactic schemata established by Świdziński [15, 16] were used. In particular, special attention was paid to the relation between superficially non-reflexive and superficially reflexive verbs, often featuring in a single lexicographic entry. An example of such annotation will be given below, in Figure 4.

The procedure adopted comprised of two basic stages. In the first stage, lexical units widely documented in the Corpus were preliminarily matched with appropriate frames. Lexical units represented by a given word were established accordingly to sense division in [1]. Preliminary frame matching was carried out using bilingual dictionaries (Polish-English). The lexicographers were provided with rough division into senses, appropriate frame descriptions taken from the Berkeley FrameNet site and with examples featuring the appropriate word. It was obvious from the start that the initial matching might be highly erroneous. First of all, there was a risk that it may be impossible to match corpus examples with appropriate lexical units. In many cases there were no examples corresponding to some units. The absence of examples did not invalidate the existence of a given lexical unit, since the examples for both [1] and [4] had been drawn from a different corpus (i.e. the full PWN Corpus, much greater than the one freely available at the PWN Corpus site, <http://korpus.pwn.pl/>). Secondly, dictionary equivalence could have been (and was) misleading. First of all, most dictionaries provide, sometimes indiscriminately, semantic equivalents and translation equivalents [8, 11]. Moreover, some translation

equivalents in the target text invoke different frames than the original words in the source text [10].

The question was further complicated by the fact that our original understanding of the Frame Semantics [5, 6] focused on the conceptual, i.e. fairly language independent notions of frames, as shown by some earlier work involving frame semantics, e.g. [18]. Working with FrameNet showed that frames are at least as much lexically geared as conceptually geared. Thus, while so-called “lexical frames” correspond to lexical units of a given natural language, the “non-lexicalized” frames are conceptual in nature and may correspond to notions general enough as to lack appropriate sets of lexemes (More on this in [3]). All the anticipated difficulties have appeared during actual lexicographic work on RAMKI. The case of the verb *jechać* is particularly illustrative of that.

First of all, one has to bear in mind that the system of motion verbs in Polish greatly differs from corresponding systems in Germanic and Romance languages. In particular, in Polish there is a clear distinction between *jechać* and *iść*. The former in its basic senses refers to motion on land, and involves the use of a vehicle or other device that helps with the movement (such as sledge, skis, skates, rollerskates, etc.). The verb may also refer to movement involving the use of an animal (horse, donkey, mule etc.). By contrast *iść* is not as general as English *to go*, since in its basic senses it means moving on foot. The importance of the vehicle, device or animal is borne out by the way this basic sense of *jechać* is defined in both [1] and [4]:

- | |
|--|
| <p>1. Jeśli jedziemy czymś, np. samochodem lub windą, albo na czymś, np. na nartach lub na koniu, to znajdujemy się w tym lub na tym i razem z tym poruszamy się w jakimś kierunku. <i>Kierowca tira jechał zbyt szybko i wpadł w poślizg. . . Mogliśmy jechać pociągiem albo autostopem. . . Ktoś jechał środkiem jezdnii na rowerze. . . Anglik jechał konno przy powozie. . . Do pracy jadę prawie godzinę. . . Dokąd jedziesz w tym roku na wakacje?</i></p> <p>2. Jeśli jakiś pojazd jedzie, to porusza się w określonym kierunku. <i>Pociąg kolebał się jadąc. . . Przepraszam, czy ten autobus jedzie do śródmieścia?</i></p> |
|--|

Fig. 1. The relevant part of the entry *jechać* in [1]

- | |
|--|
| <p>1. «przenieść się z miejsca na miejsce za pomocą środków lokomocji, odpowiedniego sprzętu sportowego lub na grzbiecie zwierzęcia; odbywać podróż»: Pół dnia jechałem w zatłoczonym pociągu. ○ Jechać samochodem, saniami, tramwajem, windą, wozem. ○ Jechać na rowerze, na sankach, na nartach, na koniu. ○ Jechać na urlop, na wycieczkę. ○ Jechać do miasta, na wieś. ○ Jechać po towar do magazynu.</p> <p>2. «o środkach lokomocji, o wszelkiego rodzaju pojazdach: być w ruchu, posuwać się naprzód»: Pociąg jedzie po szynach. Auto jedzie drogą.</p> |
|--|

Fig. 2. The relevant part of the entry *jechać* in [4]

It should be noted that in both dictionaries quoted here the main distinction between sense 1 and sense 2 involves the semantic characteristics of the subject:

human in 1 and a vehicle in 2. Interestingly, no distinction is drawn for the sense 1 between passengers and drivers, as shown by some examples from [1]:

*Kierowca tira **jechał** zbyt szybko i wpadł w poślizg...*
 ‘The TIR lorry driver was going too fast and went into a skid.’

*Mogliśmy **jechać** pociągiem albo autostopem...*
 ‘We could have gone by train or we could have hitch-hiked.’

*Do pracy **jadę** prawie godzinę...*
 ‘It takes me an hour to get to work.’

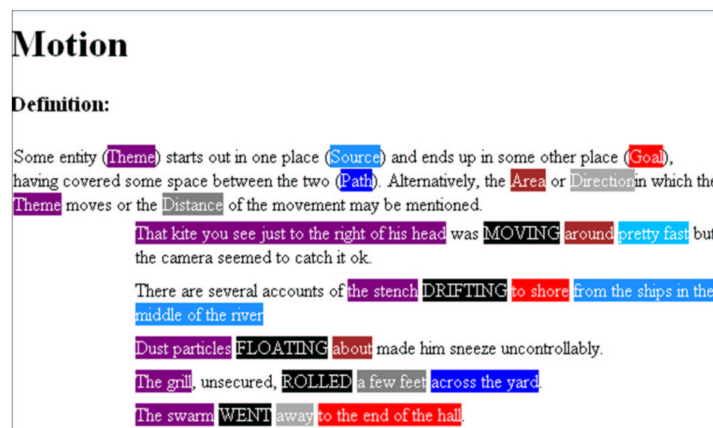


Fig. 3. The relevant part of the “Motion” frame, taken from <http://framenet.icsi.berkeley.edu>

It is not surprising that the initial frame matched to sense 1 was a very general, non-lexicalized frame of Motion, presented below in the Figure 3.

The examples found in Figure 3, compared with the examples from Figures 1 and 2, suggest that the Motion frame is but a poor match to the sense 1 of *jechać*. This can be also seen in the initial RAMKI description presented in the Figure 4².

However, it was quite easy to match some of the examples with another frame, the “Operate_vehicle” one, presented in the Figure 5.

All examples featuring bicycles and scooters were assigned this frame, as shown in Figure 6. Since the “Operate_vehicle” frame was not implemented in the software, the frame itself is not identified and the examples are annotated only roughly.

Obviously, some of the examples presented in the Figure 4 are easily understood as referring to an “Operate_vehicle” frame. This happens whenever the grammatical subject of the sentence refers to the Driver. Moreover, other examples may be easily identified with another frame relating to Motion, i.e. the “Ride_vehicle” frame,

² At present, the software does not allow for printing the results in a user-friendly way.

- , i to w dodatku nie swoich. Z Rosji_{Source} jechali tam_{Goal} tylko agenci, prowokatorzy i doradcy wojskowi_{Theme}. Natomiast [Usuń]
- , to angina, i na dodatek wyjątkowo podła. Jadę z mężem_{Lu0} do lekarza. Nasz niemiecki i francuski są [Usuń]
- , że co śnieg, to śnieg. - Szybciej_{Speed} jedzie się do progu_{Direction}, no i zimą atmosfera jest zupełnie [Usuń]
- ś chciała z nami, to nie ma sprawy. Jedziemy stopem, więc nie wiem, czy cię ten śmieciarz [Usuń]
- , zbudził ludzi i ruszył z nimi przed siebie. Jechał ku Warszawie_{Direction}, ale sam nie wiedział, po co [Usuń]
- - Toć to kulig, nie pogrzeb. Wacpan jedziesz na swym koniu_{Manner} za naszymi saniami_{Path} jak za trumną [Usuń]
- , biznes odpada, a może umarła krewna i jedziesz odebrać spadek_{Purpose?} Już wiem! – Paul Człowiek patrzy [Usuń]
- do jedzeni a, apetyt po całodziennym jeździe dopisywał, nieraz jechali nocą, bo naloty stawały się coraz częstsze, a [Usuń]
- konie , dążyli ku ostatecznemu celowi podróży. Jakiś czas jechali brzegiem_{Path_shape} rzeki płynącej w dość wąskiej dolinie, którą zasypał [Usuń]
- , gdzie kuguar zabił narciarkę. 30-letnia kobieta jechała na nartach biegowych_{Manner} wzdłuż jeziora Minnewanka, odległego o około [Usuń]
- , uciekającą w styczniu z panem Janvier na golgotę. Jadą ruską pobiedą_{Carrier...} Wskazówka szybkościomierza sięgała dziewięćdziesiątki [Usuń]

Fig. 4. The initial examples of *jechać* associated with the Motion frame

presented in Figure 7, where the subject of the sentence refers to the passenger, cf. the *Jadą ruską pobiedą* ‘They are going in a Russian Pobieda’ example (Pobieda being a Soviet car manufacturer). In addition, in some of the examples it was impossible to decide if the subject was actually denoting the passenger or the driver. These examples could not be matched with either vehicle-related frame, but they fitted another frame related to Motion, i.e. the Travel frame, presented in Figure 7 below. This frame is instantiated by such examples as:

Z Rosji jechali tam tylko agenci, prowokatorzy i doradcy wojskowi.

‘Only spooks, agents provocateurs and military advisors went there from Russia’

Jadę z mężem do lekarza

‘I am taking my husband to the doctor’

... a może umarła daleka krewna i jedziesz odebrać spadek

‘and maybe a distant relative has died and you are going to collect your legacy’

From what has been said so far, matching both corpus examples and dictionary examples with “Operate_vehicle”, “Ride_vehicle”, and “Travel” respectively seems quite easy. Nevertheless, the match itself does not constitute an adequate proof that the sense of *jechać* given as 1 both in [1] and in [4] actually covers three distinct senses of the verb in question. The fact that a “word” can refer to different real-life situations, or that it may have different translation equivalents in another

Operate_vehicle

Definition:

The words in this frame describe motion involving a **Vehicle** and someone who controls it, the **Driver**. Some words normally allow the **Vehicle** to be expressed as a separate constituent.

Tim **DROVE** **his car** all the way across North America.

Tom **SADDLED** **my canoe** across the Canadian border.

Other words in this domain are based on the names of vehicles, and do not normally allow the **Vehicle** to be expressed as a separate constituent.

The group **BIKED** all the way across the country.

However, a separate **Vehicle** constituent can occur if it adds information not included in the verb.

Tim **BIKED** across the country **on an old 10-speed**.

Fig. 5. The relevant part of the “Operate_vehicle” frame, taken from <http://framenet.icsi.berkeley.edu>

language does not necessarily prove its polysemy (cf. [2, 8] among others). However, the syntactic and semantic features of sentences with *jechać* instantiating different frames invoked so far do possess certain characteristics that provide evidence for the polysemy.

The verb *jechać* in the “Operate_vehicle” frame and the “Ride_vehicle” frame has different ways of instantiating the frame element “Speed”. Within the “Operate_vehicle” sentences this element is introduced by the expression *z szybkością* ‘at a speed’ acting as an adjunct to the verb itself, e.g.

Kierowca TIRA jechał z szybkością 100 km/h // nadmierną szybkością.
 ‘The TIR lorry driver was driving 100 km per hour // at excessive speed’

However, in “Ride_vehicle” sentences, the “Speed” element cannot be instantiated as a direct adjunct to the verb *jechać*,

**Jechał pociągiem z szybkością 200 km na godzinę.*
 ‘He was going by train at a speed of 200 km per hour.’

**Jechał nowiutkim porsche z szybkością 200 km na godzinę, wieziony przez rajdowego kierowcę.*
 ‘He was going in a new Porsche at a speed of 200 km per hour, driven by a professional rally-driver.’

The “Speed” element has to be introduced through a different structure, e.g.

Jechał pociągiem mknącym z szybkością 200 km na godzinę.
 ‘He was going by train moving at 200 km per hour.’

Jechał nowiutkim porsche wieziony z szybkością 200 km na godzinę przez rajdowego kierowcę.
 ‘He was going in a New Porsche, driven by a professional rally driver at a speed of 200 km per hour’

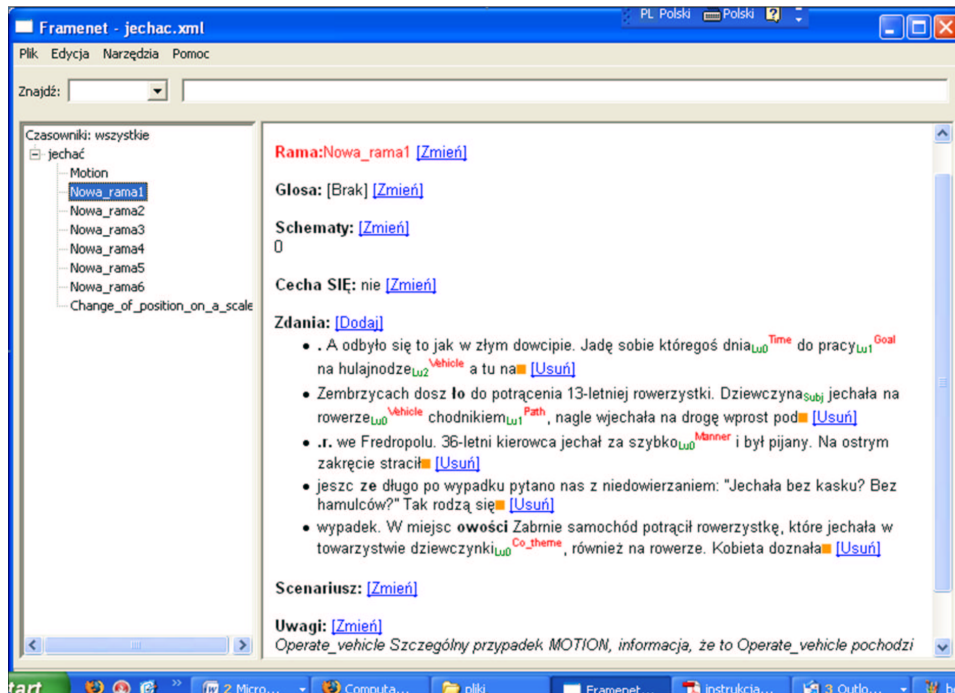


Fig. 6. Examples matching the “Operate_vehicle” frame

By contrast, the distinction between “Travel” frame and the two vehicle-involving frames is semantic in nature. Contrary to what had been said about *jechać* referring exclusively to motion on land, in this sense the verb may be used even if the journey is by sea or by air. Given appropriate pragmatic circumstances involving the air fares and the rates of exchange somebody in Warsaw may say:

Jutro jadę do Stanów po towar. Wylatuję o 8 rano.

‘I am going to the States to get the merchandise tomorrow. My flight is at 8 a.m.

without the sequence being in any way deviant. Moreover, for the “Travel” frame the distinction between actually driving or being driven or taken as a passenger (“Riding_vehicle” and “Operating_vehicle”) is irrelevant.

The attempt to match the senses of *jechać* with appropriate frames shows that the contrastive approach adopted for our project may bring forward some important, but hitherto overlooked, sense distinctions in Polish. However, the fact that the actual examples of the use of some verbs fall easily into distinct frames does not by itself constitute proof of polysemy. Polysemy can be neither proved through contrastive semantics nor through differences in reference and/or truth conditions. The latter can be accounted for by semantic vagueness of the item in question, thus further tests, including syntactic properties, areas of semantic discontinuity and other procedures have to be brought in evidence of proposed detailed senses.



Fig. 7. The relevant part of the “Travel” frame, taken from <http://framenet.icsi.berkeley.edu>

Bibliography

- [1] Bańko, M. (ed.) (2000). *Inny słownik języka polskiego*, Warsaw, Wydawnictwo Naukowe PWN.
- [2] Bogusławski, A. (1988). “Dwujęzyczny słownik ogólny. Projekt instrukcji z komentarzami”. In: Saloni, Z. (red.) *Studia z polskiej leksykografii współczesnej*, Wrocław: Ossolineum, 19–65.
- [3] Derwojedowa, M., Linde-Usiekniewicz, J. (forthcoming). „Od przypadków głębokich do elementów ram we FrameNecie”
- [4] Dubisz, S. (ed.) (2003). *Uniwersalny słownik języka polskiego*, Warszawa, Wydawnictwo Naukowe PWN.
- [5] Fillmore, C. (1982), “Frame semantics”. In: Yang I. (ed.) *Linguistics in the Morning Calm*, Seoul, Hanshin Publishing Co., 111–137.
- [6] Fillmore, C. (1985). “Frames and the semantics of understanding.” *Quaderni di Semantica*.
- [7] Linde-Usiekniewicz, J., Derwojedowa, M., Zawisławska, M. (2008). „Aspect and the Frame Elements in the FrameNet for Polish”, In: Bernal, E., DeCesaris, J. (Eds), *Proceedings of the XIII Eurolex International Congress*, 1511–1518.
- [8] Linde-Usiekniewicz, J., Olko, M. (2006). “Multilingual dictionaries on-line: reality and perspectives”. In: Koseska-Toszewa, V., Roszko, R. (Eds), *Semantyka a konfrontacja językowa*, t. 3. Warszawa: SOW, 43–59.
- [9] Obrębski, T. (2002). *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej*, (Ph.D. thesis, ms), Politechnika Poznańska, Poznań.
- [10] Padó, S. (2007). “Translational equivalence and cross-lingual parallelism: the case of FrameNet frames”. In: *Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*, Tartu, Estonia, 29–46.

- [11] Piotrowski, T. (1989)., “The bilingual dictionary – a manual of translation or a description of lexical semantics?” In: Saloni, Z. (ed.) *Studia z polskiej leksykografii współczesnej III*, Białystok: Dział Wydawnictw Filii UW, 41–52.
- [12] Polański, K., (ed.) (1980). *Słownik syntaktyczno-generatywny czasowników polskich*, t. I–V, , Wrocław: Zakład Narodowy im. Ossolińskich.
- [13] Przepiórkowski, A., Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version, Warszawa: IPI PAN.
- [14] Saloni, Z., Świdziński, M. (1998). *Składnia współczesnego języka polskiego*, Warszawa: PWN.
- [15] Świdziński, M. (1992). *Gramatyka formalna współczesnego języka polskiego*, Warszawa: Wydawnictwa UW.
- [16] Świdziński, M., 1996, *Własności składniowe wypowiedników polskich*, Warszawa: Dom Wydawniczy Elipsa.
- [17] Woliński, M. (2004). Komputerowa weryfikacja gramatyki Świdzińskiego, (Ph.D. thesis, ms), Instytut Podstaw Informatyki PAN, Warszawa.
- [18] Zawisławska, M. (2004). *Czasowniki percepcji wzrokowej. Ujęcie kognitywne*, Warszawa: Nakładem Wydziału Polonistyki.

JANUSZ S. BIENŃ

Department of Formal Linguistics, University of Warsaw, Poland

FACILITATING ACCESS TO DIGITALIZED DICTIONARIES IN DJVU FORMAT

Abstract. 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 the DjVu format has been used also for dictionaries, in particular there are 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 in particular 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.

We mention also another approach to the problem, exemplified by *djvu-xfgrep* program.

Keywords: digitalization, DjVu, dictionaries, *Poliqarp*, *djvu-xfgrep*

1 Digitalization

In the proper sense digitalization consist in representing an object by means of bits or numbers, but it can be done in many different ways. A page of text can be treated as a picture and divided into small points classified as black or white, so the page can be represented by an array of binary digits. Such points and their representations are called *pixels*, i.e. picture elements (*pix* was used as an abbreviation for *picture*, cf. <http://en.wikipedia.org/wiki/Pixel> or *Webster's New World Dictionary of the American Language*). Of course classifying all point as just black or white is often not sufficient, so there are also grayscale and color pixels in use.

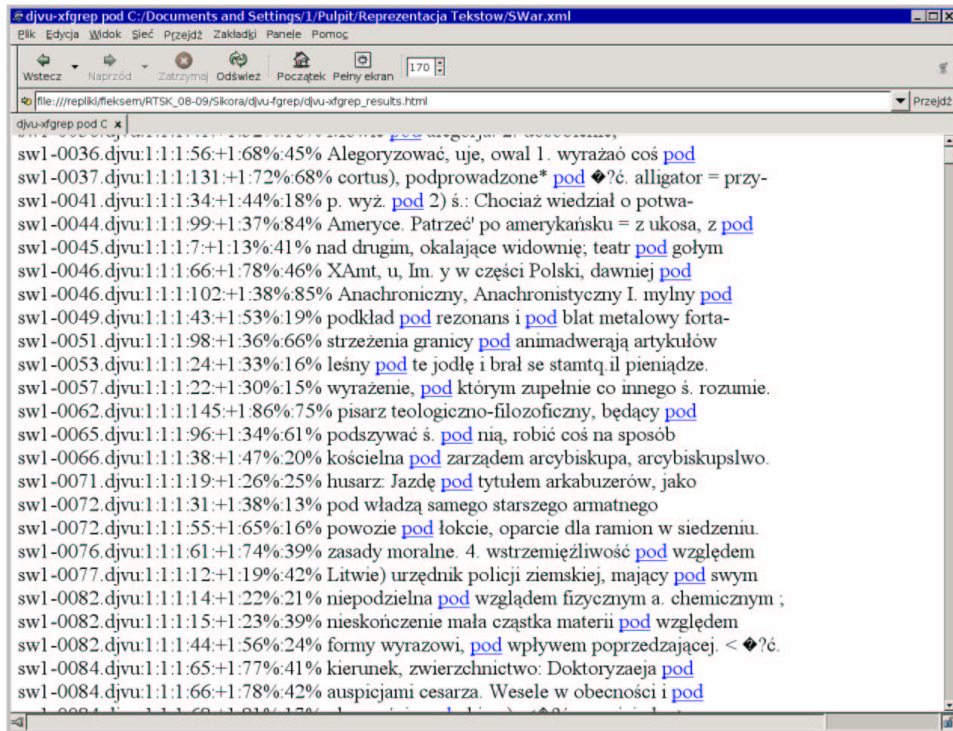


Fig. 1. Sample djvu-xfgrep output

Texts represented as pixels are usually the results of scanning printed texts, so they are called simply scanned texts or just scans (although the traditional scanners are more and more often replaced by various devices using digital photography). As such texts can be also digitally born, we advocate here a more precise term *pixel texts*. Such texts are often subject to the process of Optical Character Recognition (OCR). If the process is unattended and the results are neither verified nor corrected by humans, it is called „dirty OCR”. There is no convenient term for the textual results of OCR, so we propose here to call them *textel texts*, as they consist of characters, which can be considered the primary ‘textels’, i.e. text elements (the term *textel* is sometimes used as meaning texture element, but for this notion the term *texel* seems more popular). A typical digitally born electronic text is also a textel text.

Both pixel and textel texts can be supplemented by various additional data, extracted automatically from the text content or added by human intervention. The simplest and most useful forms of such information are outlines describing the document structure and facilitating navigation in the electronic text, annotations providing additional information about text fragments and hyperlinks e.g. to footnotes etc. These features can be provided using a commonly used standard or with

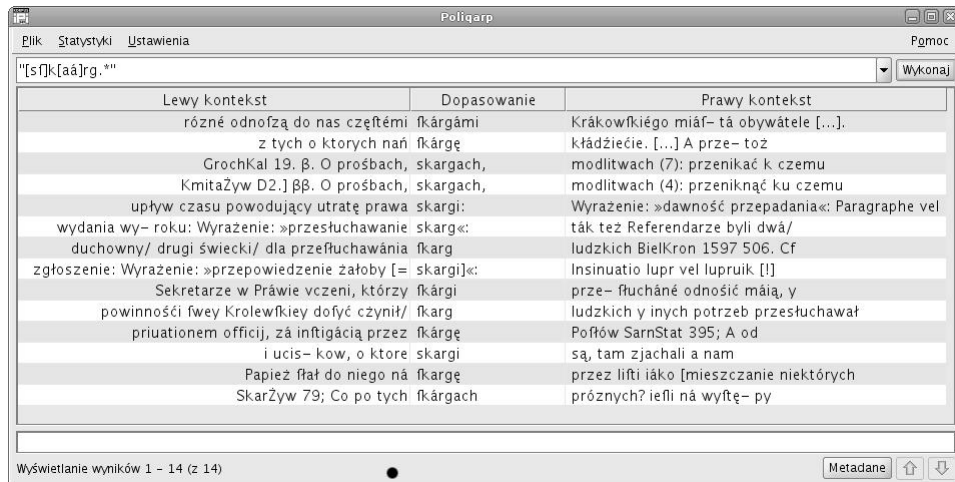


Fig. 2. Poliqarp — a query

a special purpose program. Although the term *digitalization* can be still used, the latter case is better described as computerization of a text.

2 DjVu technology

To quote [4], DjVu technology is

an image compression technique, a document format, and a software platform for delivering documents images over the Internet

It was developed by Yann Le Cun, Léon Bottou, Patrick Haffner, and Paul G. Howard at AT&T Laboratories in 1996. A document in DjVu format may consist of several layers: the pixel layer (actually split into foreground, background and stencil) representing the page images, and the textel one (called hidden text layer) which can be used for searching, cut and paste operations etc. For several years this feature was unique. In 2001 similar possibility was introduced by Adobe to the Portable Document Format (PDF) specification, but DjVu still has several advantages over PDF. The feature of DjVu, absent in PDF and most relevant for the present paper, is the possibility to store individual pages in separate files (so called *unbundled* or *indirect* documents) and to serve them over the Internet in any order. This is an important reason for preferring DjVu over PDF also for digitally born documents.

Several dictionaries available in the DjVu format have been mentioned already in [1] and [3]. One should note also *Jamieson's Etymological Dictionary of the Scottish Language Online* (<http://www.scotsdictionary.com/>) which is available both in DjVu and PDF format and is provided with sophisticated search facilities. There also interesting new acquisitions of Polish digital libraries, such as 19th century

Plik		Statystyki	Ustawienia
"[sɨ]k[aa]rg.*"			
Lewy kontekst		Dopas	
różné odnofzã do nas częftémi fkãrgãmi			
z tych o ktorych nań fkãrgę			
GrochKal 19. β. O prośbach, skargach,			
KmitaŻyw D2.] ββ. O prośbach, skargach,			
upływ czasu powodujący utratę prawa skargi:			
wydania wy- roku: Wyrażenie: »przesłuchawanie skarg«:			
duchowny/ drugi świecki/ dla przesłuchawãnia fkarg			

Fig. 3. Poliarp — query details

editions of Latin-Polish glosses by Bartłomiej from Bydgoszcz, first published in 1488. To locate all its volumes, visit the site of Federacja Bibliotek Cyfrowych (<http://fbc.pionier.net.pl>) and search for Bartłomiej z Bydgoszczy.

As of 10th September 2009, in Polish digital libraries (cf. e.g. [5]) there are 199 701 publications in DjVu format, which is 72% of the total (<http://fbc.pionier.net.pl/owoc/attr-stats>).

3 Dictionaries as texts

Every DjVu viewer allows for searching the hidden text layer, but for unbundled documents it is inefficient as it defeats the purpose of splitting the document into separate pages: to access the hidden text, all the pages have to be loaded, and if the search is repeated, they are reloaded multiple times.

Several possible solutions has been outlined in my note [2] and tentatively implemented in Java by Piotr Sikora. Two programs, `djvu-fgrep` and `djvu-xfgrep`, are available at <http://code.assembla.com/djvu-fgrep/subversion/nodes> on GNU GPL license; for the present paper only the latter program is relevant (x in the name stands for *eXtended*). The programs are named after a very popular Unix utility `fgrep`, which is a simplified version of `grep`. The `grep` program performs the function of searching globally for regular expressions and printing (i.e. displaying) the matches; the first letters of words *global*, *regular expression*, *print* make up the name of the program. The letter f in `fgrep` means that this version of the program allows only to search for *fixed* strings.

At present downloading the whole document to be searched is unavoidable, but it has to be done only once. Then the hidden text is extracted in XML format (with `djvutoxml`, a program from the free DjVuLibre library) and provided as the input to the `djvu-xfgrep` program together with the word to be searched. The results are in the form of a HTML file. Figure 1 (p. 162) shows the result of searching the word *pod* in the hidden text layer (obtained by „dirty OCR”) of the first volume of the so called Warsaw Dictionary (cf. <http://ebuw.uw.edu.pl/publication/255>).

Several columns on the left describe the localisation of the found word. First there is the name of the file containing the page in question, next the numbers of

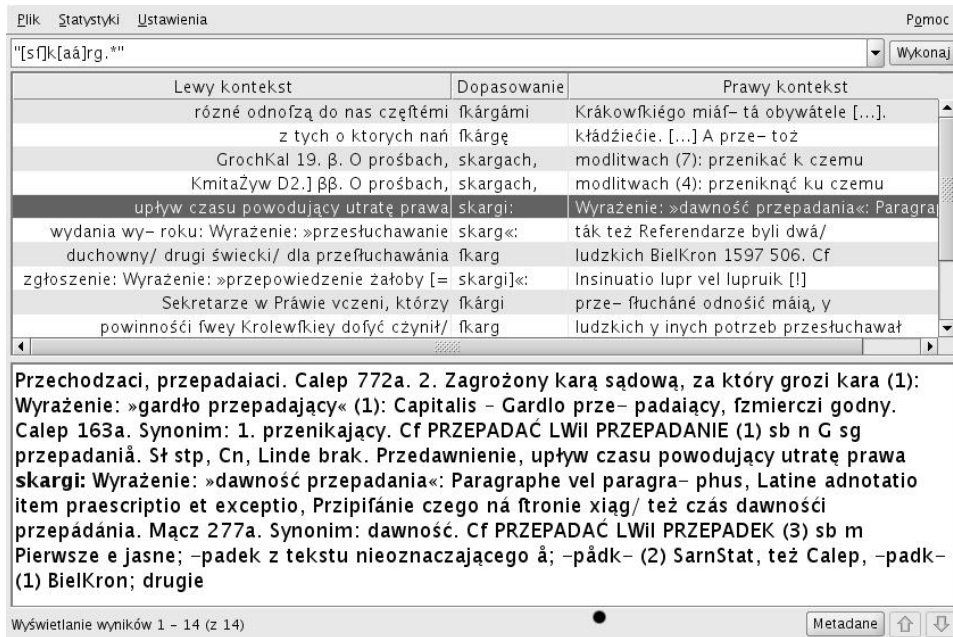


Fig. 4. Poliqarp — larger context

the column, the region, the paragraph and the line containing the word (when so detailed division of the page is inappropriate or not available, then all lines are assumed to belong to the same column, region and paragraph). After the plus sign there is the number of the word in the line and the approximate location on the pixel image of the page, expressed as percents of the horizontal and vertical size of the page. Last but not least, the context of the word is specified, and the word itself is a hyperlink to the image of the page with the word in question highlighted in a similar way as demonstrated on Figure 4 (p. 165).

Although the program is usable, implementing it was not a goal in itself, but just a milestone on the way to the converter described in the next section.

4 Dictionaries as corpora

Treating dictionaries as corpora has been suggested first, to the best of my knowledge, in [7]. Following this suggestion we propose to use corpus tools to investigate and search the text of dictionaries. For several reasons our preferred tool is Poliqarp (*Polyinterpretation Indexing Query and Retrieval Procesor*), a GPLed corpus query tool developed in the Institute of Computer Science of Polish Academy of Sciences (cf. <http://korpuz.pl/index.php?page=publications>).

Poliqarp is a client-server system. If a server is available on a local or global network, then there is no need to download the whole document (in particular a multivolume dictionary) for the purpose of searching. Another important feature of

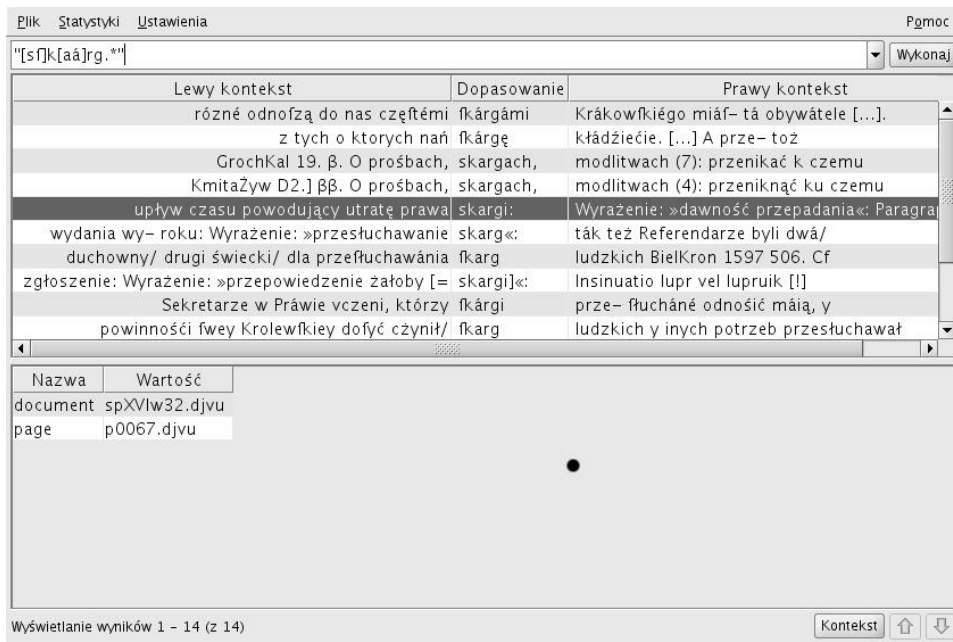


Fig. 5. Poliqarp — selected match localisation

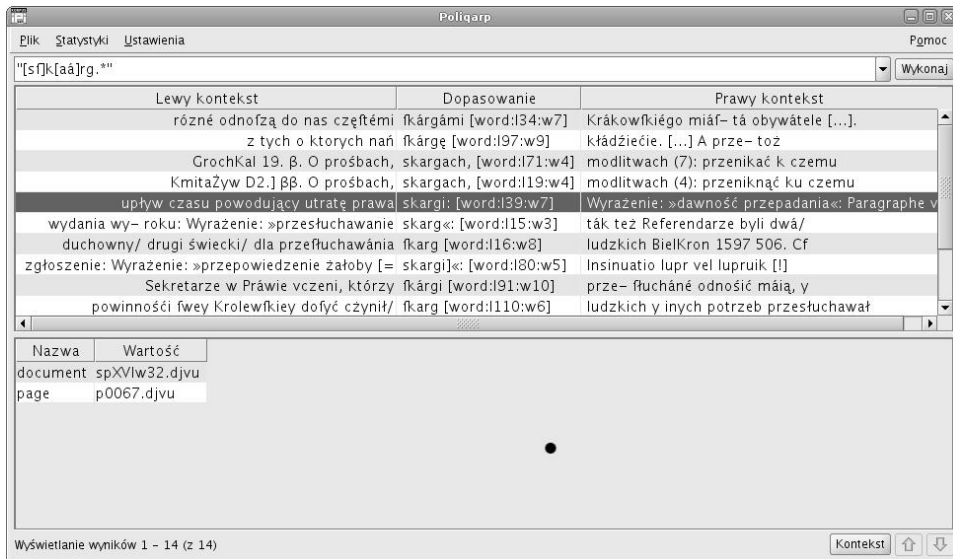
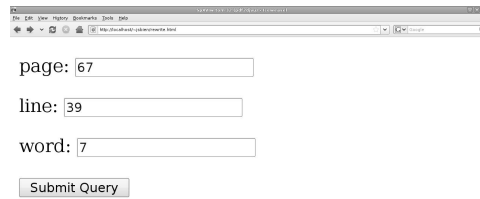


Fig. 6. Poliqarp — selected match in-page localisation



page:

line:

word:

Fig. 7. Locating the match

Poliqarp is sophisticated support of regular expressions, which can be used not only to circumvent the errors of dirty OCR, but also to accounts for different spellings in historical dictionaries.

We will illustrate the latter aspects using the 32nd volume of the Dictionary of the 16th century Polish. Thanks to the late Prof. Franciszek Pełowski, the former head of the dictionary team, there is a digitally-born version of the volume — to make a long story short, the PDF files used for printing the paper version have been converted to DjVu using Jakub Wilk’s excellent `pdf2djvu` program (<http://jwilk.net/software/pdf2djvu.html>, cf. also [6]). The converter from DjVu to **Poliqarp** has been also specified first in the note [2], then implemented in Java by Piotr Sikora and made available under the terms of the GPL license at <http://code.assembla.com/djvu-fgrep/>. An additional utility has been written by Jakub Wilk, who is also the current developer of **Poliqarp**.

Our intention is to search for the word *skarga* (meaning *complaint*) in all inflexional forms, so we search for words starting with the appropriate string — `.*` (the final part of the regular expression) means any number of any characters. On the other hand we cannot use the string *skarga* directly, because it can be spelled also with long *s* and/or with accented *a*, so we have to use list of alternative characters such as `[aá]`. Figures 2 (p. 163) and 3 (p. 164) show the result of our query (**Poliqarp** interface can be switched to English, but for this example this doesn’t seem appropriate).

Figure 4 (p. 165) demonstrates a standard feature of **Poliqarp**, namely the sub-window with the larger context of the match.

Figure 5 (p. 166) demonstrates another standard feature of **Poliqarp**, namely displaying the metadata of the document, but we use this in a non-standard way. Because we treat the dictionary as a corpus, the equivalent of a document is in our case just a page of the dictionary. In consequence the metadata consist of two fields: the reference to the dictionary as a whole (the index file of the unbundled DjVu document) and the reference to the component file containing the specific page.

Similarly Figure 6 (p. 166) demonstrates yet another standard feature of **Poliqarp**, namely displaying the tags of the matched word, and again we use this in a non-standard way. Although intended for tags such as part of speech etc., we use them simply to provide localisation of the word on the page, specifying for this purpose the line number and the running number of the word in the line.

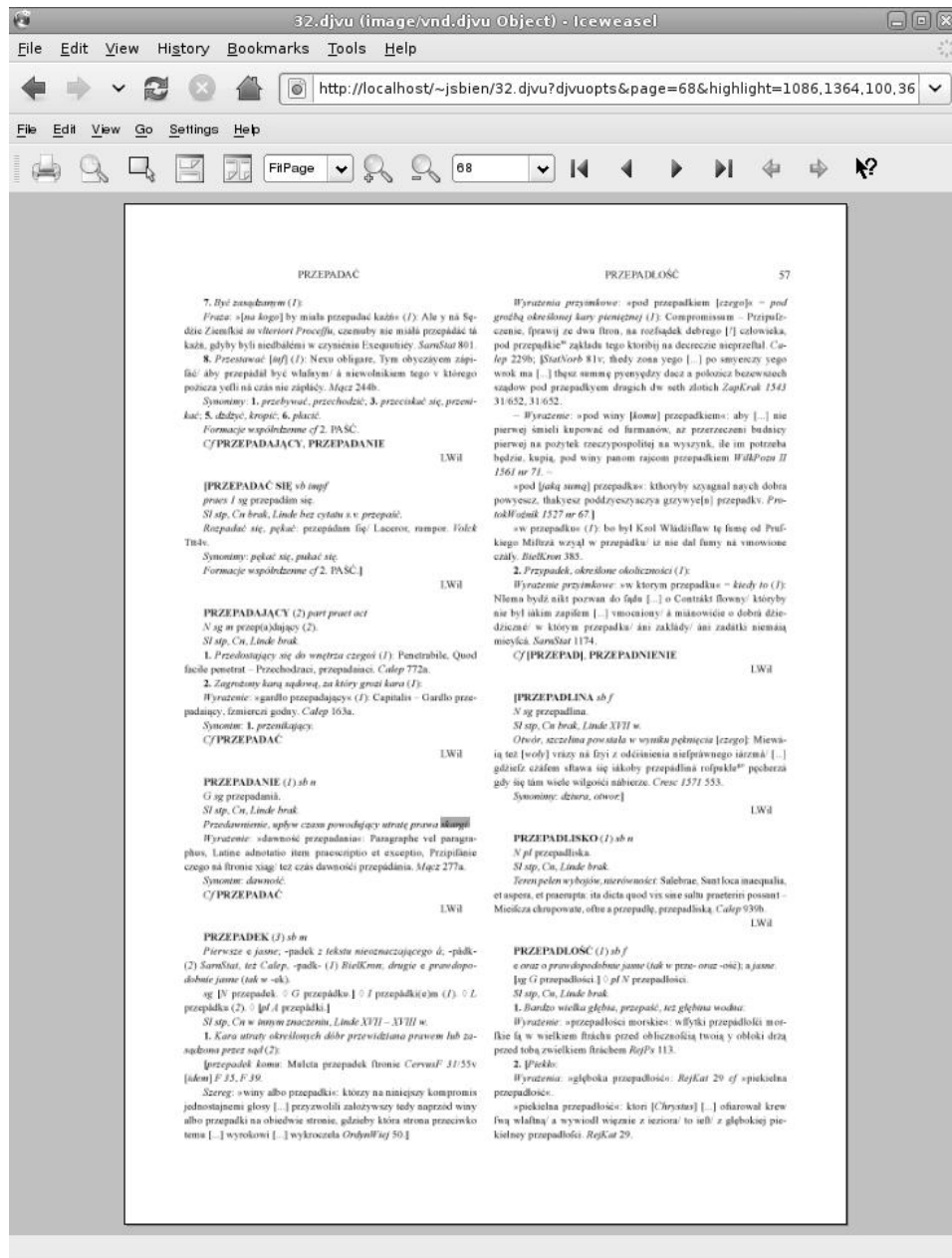


Fig. 8. The located match

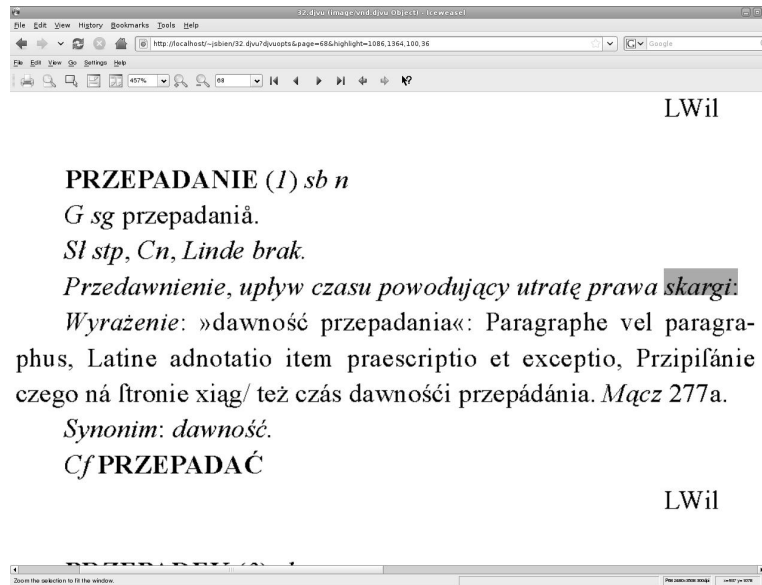


Fig. 9. Details of the match

Hence we have shown that PoliQarp can store all the information needed to locate the word in question in the pixel version of the dictionary. To make this process less cumbersome, Jakub Wilk prepared a simple tool presented on Figure 7 (p. 167).

After filling the form the user is redirected to the appropriate page with the relevant fragment highlighted, as demonstrated on Figure 8 (p. 168). Usually it will be convenient to zoom into the interesting area, as illustrated on Figure 9 (p. 169).

5 Concluding remarks

We have demonstrated that PoliQarp, supplemented by programs by Piotr Sikora and Jakub Wilk, can be used „as is” to facilitate access to large documents in DjVu format. Our plans are to make the process much more user friendly. It will be one of the goals of the *Digitalization tools for philological research* project supported by Grant N N519 384036 of the Ministry of Science and Higher Education. The project lasts from 13 May 2009 to 12 November 2011, more information will be in due time available at <http://wbs.klf.uw.edu.pl>.

Bibliography

- [1] Bień, J. S., 2006. Kilka przykładów dygitalizacji słowników *Poradnik Językowy* z. 8 (637), s. 55-63. <http://ebuw.uw.edu.pl/publication/250>.
- [2] Bień, J. S., 2008. Narzędzia do analizy tekstowej warstwy dokumentów DjVu. Unpublished note. <http://bc.klf.uw.edu.pl/105/>.

- [3] Bień, J. S., 2009 Digitalizing dictionaries of Polish. In: *Methods of Lexical Analysis: Theoretical assumption and practical applications*. Białystok, s. 37-45. <http://bc.klf.uw.edu.pl/71/>
- [4] Yann Le Cun, Léon Bottou, Andrei Erofeev, Patrick Haffner, and Bill W. Riemers. “DjVu document browsing with on-demand loading and rendering of image components” in *Internet Imaging*, San Jose, January 2001. <http://leon.bottou.org/papers/lecun-2001>.
- [5] Wałek, A., 2009. Biblioteki cyfrowe na platformie dLibra. Wydawnictwo SBM, Warszawa.
- [6] Wilk, J., 2008. Rozbudowa pakietu oprogramowania DjVuLibre. <http://jw209508.hopto.org/papers/thesis/>.
- [7] Żmigrodzki, P., 2005. Słownik jako korpus tekstów — korpus tekstów jako słownik. Perspektywy polskiej leksykografii naukowej. *Poradnik Językowy* nr 6, s. 3-14.

JELENA PARIZOSKA

Faculty of Humanities and Social Sciences, University of Zagreb, Croatian

IDIOM VARIABILITY IN CROATIAN: THE CASE OF THE CONTAINER SCHEMA

Abstract. In cognitive linguistics most idioms are considered to be motivated by various cognitive mechanisms which link the meaning of the idiom with the meanings of its constituents. One of these mechanisms is the CONTAINER image schema. In Croatian it is reflected in idioms containing the preposition *u* ('in'). The CONTAINER schema serves to structure abstract conceptual domains like SITUATIONS, EVENTS and STATES. For example, being in a difficult situation is conceptualized as being in a container. In addition to motivating the idioms with the constituent *u*, the CONTAINER schema also constrains their variability. This means that variations are not unpredictable, but are dependent on the underlying cognitive mechanism. The aim of the paper is to show that the Croatian idioms describing difficult situations vary their lexico-syntactic structure systematically to reflect the CONTAINER image schema. Based on the data from the Croatian National Corpus, we will show that the idioms share a common element, the construction *u* + NP, which constitutes the fixed core of each of the expressions and serves as the basis for variant realizations. The lexical and syntactic variations reflect the different ways in which the relation between the trajector (a person) and the landmark (a container-like object) is conceptualized. The variability of the expressions raises the issue of the criteria used in establishing the citation form in monolingual dictionaries of idioms.

Keywords: idiom variability, conceptual motivation, the CONTAINER image schema, Croatian

1 Introduction

Stability of form has been established in phraseology as one of the defining properties of idioms: one cannot normally replace the components of an idiom with other words, add new components, change the word order or the syntactic structure of an idiom, etc. Corpus-based studies have relativized the notion of stability, showing that many idioms possess one or more institutionalized variants and may also be creatively exploited in the discourse [19, 4; 18].

It has also been shown that a number of idioms have a fairly flexible structure and contain open slots which can be filled by a range of lexical items [5]. Psycholin-

guistic studies on the variability of idioms have proved that the type of lexical and syntactic variations an idiom may occur in are not unpredictable, but are dependent on and constrained by the cognitive mechanism or mechanisms motivating the expression [8, 9]. One such mechanism is the CONTAINER image schema, which serves to structure abstract concepts such as SITUATIONS, EVENTS and STATES [13, 16]. In Croatian this image schema is reflected in idioms which contain the preposition *u* ('in'). For example, a difficult or unpleasant situation is conceptualized as a location (a three-dimensional container-like object or a substance), and acting in difficult circumstances is conceptualized as being contained in a location or surrounded by a substance (e.g. *biti u klopki* (lit. be in a trap.LOC, 'caught in a trap'; *biti u sosu* (lit. be in a sauce.LOC, 'be in trouble')).

The data from the Croatian National Corpus show that idioms describing difficult situations appear in various lexical and syntactic realizations, in which the relation between a trajector (usually a person) and a landmark (an unpleasant situation, expressed by a prepositional phrase) may be construed in different ways. Let us examine the following sentences:

- (1) *Grad je u velikim dugovima.* (lit. in huge debts.LOC)
'The city is in huge debt.'
- (2) *Uletjeli smo u klopku, zar ne?* (lit. fly into a trap.ACC)
'We are caught in a trap, aren't we?'
- (3) *Iračko vodstvo u Bagdadu dovedeno u škripac* (lit. be brought into a clamp.ACC)
'Iraqi leaders in Bagdad forced into a corner.'

These examples show that different conceptualizations of the trajector-landmark relation are reflected in the choice of specific lexical items and syntactic constructions. In (1) the trajector *grad* ('city') is located in the interior of the landmark (*dugovi* 'debts'), and this static spatial relation is expressed by the use of the construction *u* + locative and the copular verb *be*. In (2) and (3) the trajector-landmark relation is construed dynamically, which is signalled by the accusative construction, implying movement of the trajector towards the landmark or its interior (*škripac* 'clamp' and *klopka* 'trap' respectively). The accusative constructions both involve a trajector moving from a source along a path to a goal (a landmark), which is signalled by the use of motion verbs. They do, however, differ in one important respect: in (2) the intransitive verb *uletjeti* ('fly into') designates self-propelled motion of the trajector, whereas in (3) the use of the transitive verb *dovesti* ('bring to') and the passive construction reflect caused motion, i.e. energy transfer from one object to another, with the trajector *iračko vodstvo* ('Iraqi leaders') being the entity receiving the energy, which causes it to move towards the landmark.

The aim of this paper is to show that the Croatian idioms describing difficult situations vary their lexical components and syntactic structure systematically to reflect the cognitive mechanism underlying idiomatic meanings – the CONTAINER image schema. More specifically, on the basis of the data from the Croatian National Corpus we will show that each of the given idioms has a core element, the construction *u* + noun phrase, which serves as the basis for variant realizations that express how the relation between the trajector and the landmark is construed. Relative to

this fixed core, different conceptualizations of the trajector-landmark relation are encoded through various syntactic constructions (locative and accusative) and lexical items (verbs). The variability of the idioms raises the issue of the criteria used in choosing a construction that is listed in dictionaries as the standard form associated with idiomatic meaning. We will show that in order to establish the conventional form of an idiom, we need to perform the grammatical and conceptual analyses using a large corpus.

The paper is organized as follows. The second section gives a brief account of image schemas in cognitive linguistics, with special emphasis on the CONTAINER schema as a mechanism motivating idiomatic meanings. The third section presents the results of the study of eight idioms describing difficult situations in the Croatian National Corpus. The fourth section is a discussion, interpreting the results of the study. The final section is the conclusion.

2 Image schemas and idioms

In cognitive linguistics most idioms are considered to be motivated by various cognitive mechanisms which link the meaning of the idiom with the meanings of its constituents [7, 14, 16]. This means that idioms are products of our conceptual system and that their meanings arise from our knowledge of the world [14, p. 201]. Psycholinguistic studies have shown that the cognitive mechanisms which motivate a large number of idioms are conceptual metaphor, conceptual metonymy, conventional knowledge and image schemas [6, 10, 15].

Johnson [13, p. xiv] defines an image schema as a “recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience.” One of the most common features of our everyday experience is containment. As Johnson [13, p. 21] points out, we experience a great number of objects in our immediate surroundings as three-dimensional containers into which we put certain things: cups, boxes, bags, rooms, vehicles and our own bodies. Thus, for an object to be experienced as a container three structural elements are required: interior, boundary and exterior [16]. It has been shown that the CONTAINER image schema underlies the concepts lexicalized by the preposition *in* [16, 17, 21]. According to [21, p. 48], the central meaning of *in* is entering or being inside a container. The container may be any three-dimensional object in which an entity can be located or which it can enter.

In Croatian, as in many other Slavic languages, spatial relations are expressed by means of prepositions, case marking on noun phrases and prefixes (for Croatian see [22]; for Czech see [12]). A situation in which one entity (the trajector) is contained within another (the landmark) is typically expressed by the construction *u* (‘in’) + locative (e.g. *u kutiji* ‘in the box’; *u sobi* ‘in the room’). On the other hand, the use of *u* + accusative designates that the trajector is moving towards the landmark or its interior (e.g. *ići u sobu* ‘go to the room’; *staviti u kutiju* ‘put in the box’). Thus, the construction *u* + accusative invokes both the CONTAINER and the SOURCE-PATH-GOAL image schemas, the latter implying that the landmark or its interior is the endpoint of the trajector’s movement. The use of the preposition *u* relatively frequently coincides with the use of verbs formed with the prefix *u-* [23,

p. 100], signalling that the trajector changes its location and moves into the interior of the landmark, which is a container-like object (e.g. *ubaciti* ‘throw into’; *ući* ‘go into’; *uletjeti* ‘fly into’).

It has been shown that image schemas give rise to metaphorical extensions into non-spatial domains [13, 16]. For example, we conceptualize emotions and states in terms of containment, which is the reason why the preposition *u* is used to express non-spatial meanings. In Croatian a number of idioms containing the preposition *u* describe situations in which the trajector is a person and the landmark is an unpleasant or dangerous situation. Here are some examples:¹

- (4) *Slušaj, stari, u govnama si do grla ...* (lit. be in excrement.LOC)
 ‘Look, mate, you’re in deep trouble.’
- (5) *... jer ne treba od svakog kiksa padati u totalnu komu* (lit. fall into a coma.ACC)
 ‘You mustn’t get depressed every time you make a mistake.’

In Croatian we talk about being in trouble (*biti u govnama*) (4) or falling into a depression (*padati u komu*) (5) because abstract concepts like SITUATIONS, STATES and EMOTIONS are structured by the CONTAINER image schema. This results in the conceptual metaphor STATES ARE CONTAINERS. The idiomatic meanings ‘be in trouble’ and ‘get depressed’ are related to spatial meanings via conceptual metaphor.

Studies of the lexical and syntactic behaviour of idioms [3, 9, 9, 11, 20] have established that variations are not unpredictable, but are closely dependent on the cognitive mechanism or mechanisms motivating the idioms. Therefore, it would be reasonable to assume that the variability of the Croatian idioms describing difficult situations is constrained by the CONTAINER image schema. In the remainder of the paper we will look into the ways in which the relation between a trajector (a person) and a landmark (an unpleasant situation) is construed and check for any systematic variability of lexical components and syntactic structure which would reflect the underlying CONTAINER image schema.

3 Corpus research and results

For the present study we chose eight idioms which have similar meanings and structure.² Each expression describes a situation in which a person faces an unpleasant, difficult, or dangerous situation or a big problem. All the expressions share a common element, the construction *u* (‘in’) + a noun phrase. The noun phrases are as follows: *dug/dugovi* (lit. debt/debts); *gabula* (lit. a difficult situation in a card game)³; *govna/drek* (lit. excrement); *kaša* (lit. porridge); *klopka/stupica/zamka* (lit. trap); *sos* (lit. sauce); *škripac/procijep* (lit. clamp); *žrvanj* (lit. millstone).

¹ The examples have been taken from the Croatian National Corpus.

² The idioms have been taken from two monolingual dictionaries of Croatian idioms — *Frazeološki rječnik hrvatskoga ili srpskog jezika* [29] and *Hrvatski frazeološki rječnik* [30].

³ The word *gabula* denotes a situation in a card game in which a player holding a jack and a king is bound to lose both cards (cf. [24] headword *gabula*).

We performed a study of the eight expressions in the Croatian National Corpus (www.hnk.ffzg.hr). We looked for occurrences of the common element, the pattern *u* ('in') + noun phrase, within the span of 5 words. Croatian is a fleective language, so queries were built in such a way to identify all forms of the nouns making up the idioms. We eliminated instances of nouns used in the literal sense. In this way we obtained a sample of 681 variant realizations of the eight idioms describing difficult situations from the Croatian National Corpus.

Our results show that the combination of the preposition *u* and the noun phrase remains unaltered, which implies that it is the core element of each of the idioms. This construction designates a difficult situation that a person faces. The core is subject to different conceptualizations of the relation between the trajector (a person) and the landmark (a difficult situation), which is reflected in the choice of the syntactic construction (*u* + locative or *u* + accusative) and peripheral lexical constituents, namely verbs. The results are presented in Table 1.

Construction	Number of examples	%
<i>u</i> + LOC	331	49%
<i>u</i> + ACC	350	51%
Total	681	100%

Table 1. Locative and accusative constructions in the Croatian National Corpus

The construction *u* + locative is used to express a static spatial relation: the landmark is seen as a container and the trajector is an entity located in its interior. In 54% of the cases, the static variant of the spatial relation is expressed by the verb *be*, as in the following examples:

- (6) *Liberali su u stupici!* (lit. be in a trap) 'The Liberals are in a trap.'
- (7) *Ako nam to propadne, onda smo tek u gabuli.* (lit. be in a difficult situation) 'If that fails, then we're really in trouble'
- (8) *Svi smo u istom sosu...* (lit. be in the same sauce) 'We're all in the same boat.'

These examples show that being in a difficult situation is conceptualized as being located in an enclosed place which resembles a container. The nouns which make up the idioms and the verbs which occur in the locative constructions highlight various aspects of containment. For example, *klopka/stupica/zamka* 'trap', *škripac* 'clamp' and *žrvanj* 'millstone' imply that the trajector is affected by the physical force of the container (a solid material), which holds it in a particular position and prevents it from moving. The landmark may also be conceptualized as an amorphous substance which covers the trajector to a certain degree: *dug/dugovi* 'debt/debts', *govna/drek* 'excrement', *sos* 'sauce', *kaša* 'porridge'. In such cases the landmark lacks clear borders, but it is conceptualized as a container by virtue of the fact that the deeper the trajector is confined within a given substance, the

more difficult it is for them to move and go outside. This is signalled by the use of verbs such as *grcati* ‘choke’, *gušiti se* ‘suffocate’ and *utopiti se* ‘drown’.

The use of the construction *u* + accusative indicates that the relation between the trajector and the landmark is construed dynamically. In each case getting into a difficult situation is conceptualized as movement of the trajector along a path towards the interior of the landmark, which is a container-like object. Thus, the construction *u* + accusative invokes the CONTAINER schema as well as the SOURCE-PATH-GOAL schema.

The results show that there are two types of dynamic construals. 77% of the accusative constructions relate to self-propelled motion, which is expressed by the intransitive construction with the trajector as the subject. Here are some examples:

- (9) *Kako je i zašto Dinamo uopće upao u dugove?* (lit. fall into debts)
 ‘How and why did Dinamo run a debt in the first place?’
- (10) *Vojnici, kada dođu u škripac, neće reagirati ukočeno...* (lit. come to a clamp)
 ‘When soldiers are forced into a corner, they don’t freeze...’
- (11) *Ona je, međutim, tek ušla u žrvanj političkih igara.* (lit. go into a millstone)
 ‘However, she is new to the political arena.’

In sentences (9)-(11) the interior of the landmark expressed by the prepositional phrase is the endpoint of the trajector’s motion. These examples also show that several elements contribute to the directional interpretation of the trajector-landmark relation: the preposition *u*, case marking on noun phrases and motion verbs. The fact that the verbs used in these expressions belong to a restricted set shows that changing state is conceptualized as changing position relative to a container, with the accusative phrase as the destination of the trajector’s movement.

Some verbs which occur in the accusative constructions clearly indicate that getting into a difficult situation is involuntary, as in the following examples:

- (12) *Siguran sam da će se uloviti u zamku koju ću mu lijepo nastaviti ...* (lit. get caught in a trap.)
 ‘I am sure that he will get caught in the trap that I will set for him’
- (13) *... darovitoga i sposobnog odvjetnika koji se upleo u kartaške dugove.* (lit. become entangled in debts)
 ‘...a talented and competent lawyer who got into gambling debts.’

The use of the verbs *uloviti se* ‘get caught’ in (12) and *uplesti se* ‘become entangled’ in (13) signals that the trajector has little or no control of its movement towards the interior of the container.

The other type of dynamic construals reflect caused motion: the trajector is brought into a difficult situation by another entity. There are 23% of such examples. This is expressed by the transitive construction with the moving trajector as the direct object:

- (14) *Sabina uvidi da ju je taj upit bacio u škripac...* (lit. throw into a clamp)
 ‘Sabina realized that her question put her in a difficult position’

- (15) *Oni koji su uništili klub, koji su ga gurnuli u dugove veće od 20 milijuna maraka.* (lit. push into debts)
 ‘The people who ruined the club, who ran up a debt of over 20 million marks’
- (16) *Želim vidjeti može li taj momak i dalje plesati ako ga uvalimo u govna.* (lit. push into excrement)
 ‘I want to see whether that kid will still be doing it if he’s boxed into a corner’

These examples show that the dynamic construals in this group are realized as energy transfer from an agent to the trajector, which brings about the motion of the trajector toward the interior of the landmark. The caused-motion schema is reflected in the use of force-dynamic verbs *baciti* ‘throw’ (14), *gurnuti* ‘push’ (15) and *uvaliti* ‘push into’ (16).

4 Discussion

The results show that the eight Croatian idioms we studied share a common element, the construction *u* + noun phrase, which profiles a difficult or unpleasant situation that a person is in. This element constitutes the core of each of the expressions and serves as the basis for variant lexical and syntactic realizations. The relation between the trajector (a person) and the landmark (an unpleasant situation) may be conceptualized in a number of ways. Therefore the core is subject to different construals: static or dynamic, which is expressed by the constructions *u* + locative and *u* + accusative respectively; the self-propelled motion of the trajector, signalled by the use of the intransitive construction; and the motion of the trajector brought about by energy transfer from another entity, which is reflected in the use of the transitive construction with the trajector as clausal object. In other words, there are three variants: one in which the construction *u* + locative is used with the verb *be* (as in (6)–(8)), one in which the construction *u* + accusative is used with an intransitive verb (typically a motion verb, as in (9)–(11)), and one in which the accusative phrase is used with a transitive verb (usually a force-dynamic verb, as in (14)–(16)).

Our findings confirm a correlation between the variant realizations and conceptual motivation. The use of the construction *u* + locative and the copular verb *be* reflects the CONTAINER image schema: the trajector is an entity contained in a location, which implies that being in a difficult situation is conceptualized as being located in a container-like object. The accusative construction invokes both the CONTAINER schema and the SOURCE-PATH-GOAL schema, implying that the trajector changes location and moves along a path towards the interior of the landmark, which is the goal of the trajector’s movement. In other words, the accusative phrase indicates a change of state. A person may get into an unpleasant situation by virtue of their own movement or may be brought into difficulties by another entity. As a result, the verbs in the accusative constructions vary systematically to reflect self-motion or caused motion: getting into a difficult situation by virtue of one’s own movement is expressed by motion verbs (e.g. *upasti* ‘fall into’, *uletjeti* ‘fly into’, *ući* ‘enter’, *doći* ‘come’), and causative relations are signalled by force-dynamic verbs (e.g. *uhvatiti* ‘catch’, *wući* ‘drag into’, *ubaciti* ‘throw into’, *staviti* ‘put’). In 34%

of accusative constructions the verb is prefixed with *u-*. This exhibits double motivation by the CONTAINER schema: on the one hand, entering or being inside a container is reflected by the use of the preposition *u*; on the other, it is reflected by the prefix *u-* which is attached to the verbs, signalling movement of the trajector towards the interior of a container. Thus, the variant lexical and syntactic realizations of the Croatian idioms describing difficult situations systematically reflect the CONTAINER image schema, which makes their meanings motivated.

The variability of the idioms describing difficult situations raises the issue of establishing the dictionary citation-form. In monolingual dictionaries of Croatian idioms, which are based on hand-collected data, the locative and accusative constructions as well as the constructions reflecting self-motion and caused motion are listed alphabetically by key word, under separate entries. This would suggest that they are treated as separate expressions. Corpus evidence shows that they are not individual expressions but variant realizations of a more schematic configuration, *u* + noun phrase, which is motivated by the CONTAINER image schema. Moreover, the choice of specific lexical items and syntactic constructions depends on how the relation between the trajector and the landmark is construed, which implies that the variant realizations are constrained by the underlying cognitive mechanism motivating the expressions.

The results of experiments conducted with learners of English as a foreign language have proved that knowledge of cognitive mechanisms which link idiomatic meanings to literal ones greatly facilitates idiom learning [1, 2, 15]. Cognitive linguistic tools to language learning combined with corpus data have been applied in monolingual English dictionaries, so that idioms are arranged both by key word (with information on the variability of a given expression and the range of possible forms) and according to theme, i.e. meaning (cf. e.g. *Cambridge International Dictionary of Idioms* [25], [31] *Longman Idioms Dictionary* [28], *Collins Cobuild Dictionary of Idioms* [26]). We therefore suggest that in establishing the conventional form of idioms and organizing entries in Croatian dictionaries three factors should be taken into consideration: conceptual motivation, lexico-syntactic structure and corpus data. It has been proved that idiom learning can be significantly aided by knowledge of conceptual motivation, and large corpora provide sufficient evidence to describe the forms that idioms occur in.

5 Conclusion

Based on the data from the Croatian National Corpus we have established that the Croatian idioms describing difficult situations have a fixed core, the construction *u* + noun phrase, which contains an open slot that may be filled by verbs belonging to restricted sets, namely motion verbs and force-dynamic verbs. Even though the choice of specific lexical items cannot be predicted, they must be compatible with the underlying conceptual motivation. Thus, the variant realizations reflect the cognitive mechanism motivating idiomatic meanings – the CONTAINER image schema – in a systematic way.

We have also shown that in establishing the standard form of an idiom we need to analyze the syntactic and semantic features of its components using corpus

evidence. Large computer corpora do not only provide accurate information on the frequencies and distribution of idioms, but they also show that the forms of a number of idioms are relatively variable. Thus, dictionary entries should provide a more schematic account of standard forms to allow for variability.

Bibliography

- [1] Boers, Frank (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics* 21.4: 553–571.
- [2] Boers, Frank, Murielle Demecheleer (2001). Measuring the impact of cross-cultural differences on learners' comprehension of imageable idioms. *ELT Journal* 55.3: 255–262.
- [3] Cacciari, Cristina, Sam Glucksberg (1991). Understanding Idiomatic Expressions: The Contribution of Word Meanings. Simpson, Greg B., ed. *Understanding Word and Sentence*. North-Holland: Elsevier Science Publishers B.V., 217–240.
- [4] Cignoni, Laura, Stephen Coffey, Rosamund Moon (2002). Idiom variation in English and Italian: two corpus-based studies. *Languages in Contrast* 2.2: 279–300.
- [5] Fillmore, Charles J., Paul Kay, Mary C. O'Connor (1988). Regularity and idiomatcity in grammatical constructions: The case of *let alone*. *Language* 64.3: 501–538.
- [6] Gibbs, Raymond W., Jr. (1990) Psycholinguistic studies on the conceptual basis of idiomatcity. *Cognitive linguistics* 1.4: 417–451.
- [7] Gibbs, Raymond W., Jr. (1994). *The Poetics of Mind: Figurative Thought, Language and Understanding*. Cambridge: Cambridge University Press.
- [8] Gibbs, Raymond W., Jr., Nandini Nayak (1989). Psycholinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology* 21: 100–138.
- [9] Gibbs, Raymond W., Jr., Nandini Nayak, John Bolton, Melissa Keppel (1989). Speakers' assumptions about the lexical flexibility of idioms. *Memory and Cognition* 17: 58–68.
- [10] Gibbs, Raymond W., Jr., Jennifer O'Brien (1990). Idioms and mental imagery: the metaphorical motivation for idiomatic meaning. *Cognition* 36: 35–68.
- [11] Glucksberg, Sam (1993) Idiom meanings and allusional content. Cacciari, Cristina, Patrizia Tabossi, eds. *Idioms: Processing, Structure, and Interpretation*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3–26.
- [12] Janda, Laura A., Steven Clancy (2006). *The Case Book for Czech*. Bloomington, Indiana: Slavica Publishers.
- [13] Johnson, Mark (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago - London: The University of Chicago Press.
- [14] Kövecses, Zoltán (2002). *Metaphor. A Practical Introduction*. Oxford University Press.
- [15] Kövecses, Zoltán, Péter Szabó (1996). Idioms: a view from cognitive semantics. *Applied Linguistics* 17.3: 326–355.
- [16] Lakoff, George (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago – London: The University of Chicago Press.

- [17] Lakoff, George, Mark Johnson (1980). *Metaphors We Live By*. Chicago: The University of Chicago Press.
- [18] Langlotz, Andreas (2006). *Idiomatic Creativity*. Amsterdam – Philadelphia: John Benjamins.
- [19] Moon, Rosamund (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.
- [20] Nunberg, Geoffrey, Ivan Sag, Thomas Wasow (1994). Idioms. *Language* 70.3: 491–539.
- [21] Rudzka-Ostyn, Brygida (2003). *Word Power: Phrasal Verbs and Compounds: A Cognitive Approach*. Berlin: Mouton de Gruyter.
- [22] Silić, Josip, Ivo Pranjković (2005). *Gramatika hrvatskoga jezika*. Zagreb: Školska knjiga.
- [23] Šarić, Ljiljana (2008). *Spatial Concepts in Slavic. A Cognitive Linguistic Study of Prepositions and Cases*. Wiesbaden: Harrassowitz Verlag.

Sources

- [24] Anić, Vladimir (2004). *Veliki rječnik hrvatskoga jezika*. Zagreb: Novi Liber.
- [25] *Cambridge International Dictionary of Idioms* (1998). Cambridge: Cambridge University Press.
- [26] *Collins Cobuild Dictionary of Idioms*. 2nd Edition. (2002). Glasgow: HarperCollins Publishers.
- [27] Croatian National Corpus. Available at: www.hnk.ffzg.hr
- [28] *Longman Idioms Dictionary* (2001). Harlow: Longman.
- [29] Matešić, Josip (1982). *Frazeološki rječnik hrvatskoga ili srpskog jezika*. Zagreb: Školska knjiga.
- [30] Menac, Antica, Željka Fink-Arsovski, Radomir Venturin (2003). *Hrvatski frazeološki rječnik*. Zagreb: Naklada Ljevak.
- [31] Spears, Richard A. (1998) *NTC's Thematic Dictionary of American Idioms*. Lincolnwood: NTC Publishing Group.

MATEUSZ-MILAN STANOJEVIĆ¹
BARBARA KRYŻAN-STANOJEVIĆ¹

¹Faculty of Humanities and Social Sciences, University of Zagreb, Croatian

LEVELS OF CONSTRUCTIONAL MEANING:
THE CONFLUENCE OF THE DATIVE AND MIDDLE VOICE
IN POLISH AND CROATIAN

Abstract. 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 Tudman Vuković forthcoming). 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.

Keywords: dative, middle voice, grammaticalization, Croatian, Polish.

1 Introduction

In various types of cognitive constructional approaches, such as Langackerian cognitive grammar [18] and Goldberg’s construction grammar [6], the meaning of the construction is not the sum of its parts. Rather, there is emergent meaning which may be motivated by the parts of the construction, but cannot be entirely predicted by it. Well known cases include the *let alone* construction [5] or the *What’s X doing Y?* construction [13].

Studies so far have mostly focused on two issues: the development of the construction and the type of emergent meaning. In studying the development of the construction, the focus is on how the final construction is grammatically and semantically motivated by its parts — this is the case of Langacker’s exploration of the present perfect, the passive, etc. [20]. In the “emergent meaning” study, the characteristics of the emergent constructions are attended to, explaining the rich grammatical and pragmatic detail, without necessarily looking at its motivation (as is the case in the *let alone* construction).

This study is done along the lines of “developmental” studies, with one important difference — we will be looking at the way in which two constructional units form a single larger constructional unit. The first constructional unit is the *se/się* construction and the second unit is the dative construction. The *se/się* + dative construction includes examples such as:

- (1) *Ako se [Hrvatska] pridruži Partnerstvu za mir*
If refl [Croatia] joins Partnership-DAT for peace
‘If Croatia joins the Partnership for peace...’
- (2) *Zależy ... komu się ją opowiada*
depends who-DAT refl. it-ACC told
‘it depends on who it is told to’
- (3) *zlo koje se dogodilo Hrvatskoj*
evil which refl. happened Croatia-DAT
‘... the evil that happened to Croatia’

In Croatian example the *se* construction refers to self motion of the agent (Hrvatska ‘Croatia’) which moves towards the dative. The dative is not aware of the agent’s movement, and the verb form with *se* is the only one possible. As opposed to that, the Polish example shows a typical usage in which the dative construction of energy transfer (signaled by the verb *opowiadać* ‘talk’) is combined with the *się* construction. The *się* construction in this case is a clear way of defocusing the agent who is doing the talking. Significantly, this sentence has a straightforward non-middle counterpart, where the verb simply appears with the agent doing the action of talking. Finally, is also composed of agent defocusing and the dative as being influenced by the action. However, there is no clear non-middle counterpart. The question is then — what is the difference, if any, between (1), (2) and (3)?

In this paper we will be looking into the ways in which the *se/się* construction interacts with the dative construction, trying to prove that there are two basic levels of interaction — the level where there is no confluence between the two constructions and no emergent meaning (as in (1)), and the level where there is confluence between the two constructions and new emergent meaning occurs (as in (2) and (3)). Based on a detailed look at various contexts where the two constructions co-occur in Croatian and Polish we will prove that the confluence — emergent meaning — is possible only in the grammaticalized senses of each of the constructions, which are more schematic in respect to the “basic”, diachronically older senses.

The paper is organized as follows. The following section gives a short description of the dative and *se/się* construction in Croatian and Polish using the theoretical framework of Langacker’s cognitive grammar. Section 3 examines the interaction between the two constructions, defining levels of confluence, and discussing why there is confluence in some cases and no confluence in others. Section 4 is the conclusion.

2 The dative and the *se/się* construction in Croatian and Polish

2.1 The dative construction

The prototypical Croatian and Polish dative are dominion constructions, which involve the transfer of a thematic entity into the dative’s dominion. For instance, in the Croatian example (4) the trajector *haljinu* ‘dress-ACC’ is transferred to the dative *mi* ‘me’:

- (4) ...[ona] *mi* *je dala svoju haljinu.*
 ...[she] me-DAT is gave her dress-ACC.
 ‘She gave me her dress.’

This is schematically depicted in Figure 1, where the agent (*ona* ‘she’) transfers some energy onto the theme (*haljinu* ‘dress-ACC’) moving it from the source domain to the target domain. In the target domain the dative (*mi* ‘me’) establishes mental contact with the theme. The two ovals signal the dominion — a conceptual region or a set of entities to which a particular reference point affords mental access [20, p. 170], [21, p. 6].

This type of analysis of the dative has been used for a number of Slavic languages, including Czech [10], Polish [3, 25, 31]), and Croatian [26, 27]. Mental access of the dative and the entity within its dominion spells out the dative’s awareness of the landmark or the dative’s affectedness by it. This is reflected in several dative patterns, a case in point being the transfer pattern as in (4), where a thematic entity is transferred to the dative. Other dominion patterns include the assessment pattern (where the dative assesses a thematic entity or its surroundings), and the “dative of possession” (where the dative affords mental access to its “possession”). Let us have a look at two examples, one in Polish (5) and one in Croatian (6).

- (5) *Jest mi gorąco.*
 Is me-DAT hot
 ‘I am hot.’

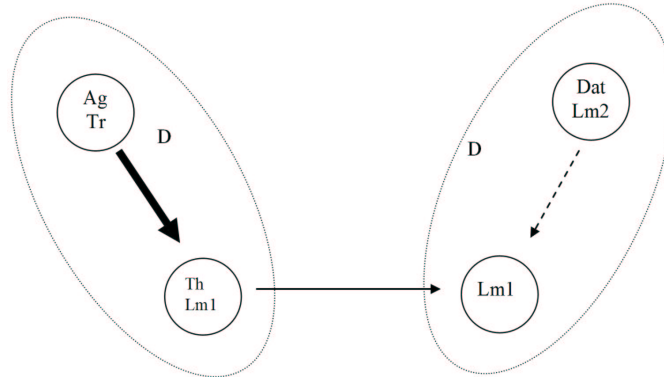


Fig. 1. The dative transfer sense

- (6) *Roditelji su mu stanovali u Selešu.*
 parents are him-DAT lived in Seleš-LOC
 ‘His parents lived in Seleš.’

Example (5) signals the dative’s assessment of its setting, which is simultaneously its dominion. The dative (*mi* ‘me’) finds his/her setting hot. The dominion here signals affectedness of the dative — the heat of the setting affects the dative. Example (6) is traditionally dubbed the “dative of possession” (in the broad sense of the word, including parts of the body, relatives and friends). Here the dative pronoun *mu* ‘him’ is used to locate the entity *roditelji* ‘parents’. Affectedness is also clear — anything that happens to the dative’s “possessions” affects the dative. Polish also includes “dative of possession” examples (e.g. *Ogolili Pawłowi głowę* ‘They shaved Paweł’s-DAT head’; example from [3, p. 121]).¹

In addition to dominion examples of the dative, both the Croatian and the Polish dative contain examples where no dominion is constructed around the dative. These uses are diachronically older, and contain examples of the trajector’s physical or abstract movement towards the dative (which we term the allative) and matching of the dative’s and the trajector’s force (which we term the competitor sense (after [11])). Thus, Croatian example (7) illustrates the trajector’s (*igrač* ‘player’) running towards the dative (*njoj* ‘her’).

¹ Polish is somewhat more restricted than Croatian in this sense. The restriction is due to the empathy hierarchy, which expresses “different degrees of the speaker’s empathy with the referent of the dative” [15, p. 330] — or, in other words, the centrality of the elements within the dative’s dominion. The empathy starts with inalienable possessions (e.g. body parts, kin), and extends to various objects that a person may be interested in, etc. It seems that Croatian allows dative marking for a wider range of elements on the scale regardless of possible affectedness, whereas in Polish dative may be used with central elements of the hierarchy, and may be used with peripheral members only if affectedness is clearly contextually marked [28, p. 14–15].

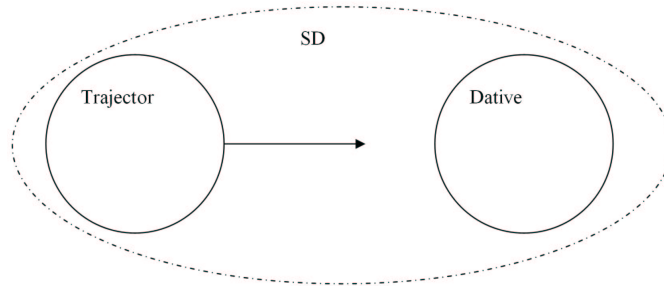


Fig. 2. The allative pattern and the search dominion

- (7) *Igrač je ... potrčao prema njoj*
 player is ... started running toward her-DAT
 ‘... the player ... started running toward her.’

Polish does not have many allative examples: they have been superseded by the *do* + genitive construction, although Kempf does give a Polish example (cited in [3, p. 49]):

- (8) *Wszyscy wybiegli mu na spotkanie*
 everyone ran out him-DAT on meeting
 ‘Everyone ran out to meet him.’

Still, there are plenty of other non-physical examples which we would include in the allative category, with experiential verbs, where the dative denotes the object of experience. The subject, so to speak, mentally moves towards the object, or, as Dąbrowska puts it, directs his “mental gaze” at the dative [3, p. 49]. In Croatian these are, e.g., *čuditi se* ‘wonder at’, *radovati se* ‘look forward to’, and in Polish, e.g. *przyglądać się* ‘look at’, *dziwić się* ‘be surprised at’, etc. The allative category is characterized by a search dominion [27] — although the dative is used as a reference point to locate the trajector which moves towards it, unlike in other examples, there is no affectedness. This is illustrated in Figure 2.

The allative is related to the competitor pattern, where the forces of the trajector and the dative are matched. Let us elaborate on two examples (example (9) is in Croatian and example (10) is in Polish):

- (9) *Hrvatski šumari oštro se protive proglašenju novih zaštićenih područja.*
 Croatian foresters adamantly refl. oppose declaration-DAT new protected areas
 ‘Croatian foresters are adamantly opposed to the declaration of new protected areas.’
- (10) *Stany Zjednoczone przeciwstawią się jakiegokolwiek porozumieniu...*
 States United will oppose refl. any agreement-DAT
 ‘The United States will oppose any agreement...’

In both examples there is a trajector (*hrvatski šumari* ‘Croatian foresters’ in (9), and *Stany Zjednoczone* ‘the United States’ in (10)), whose force clashes (signaled by the verb *protiviti se* and *przeciwstawić się* ‘oppose’ and similar verbs) with the inherent force of the dative entity — *proglašenje* ‘declaration’ in (9) and *porozumienie* ‘agreement’ in (10). The dative in the competitor pattern may be any type of entity with an inherent force, for example illocutionary force (as is the case with declaration in (9)), legal force (agreement in (10)), etc.

In this short overview of the dative construction we showed that the Croatian and Polish dative construction consists of two main subtypes — a dominion and a search dominion construction. In the dominion construction the dative signals an entity that establishes mental contact with the thematic item, and may be affected by it. Examples of this construction prototypically involve the transfer of a thematic entity towards the dative, and the dative’s affectedness by the entity being transferred into its dominion. In the diachronically older search dominion construction, the dative may be used as a reference point to locate an entity, but does not establish mental contact with it, and is, hence, not affected by it. This pattern includes examples of self-movement of the trajector towards the dative or energy clash between the trajector and the dative.

2.2 The *se/się* construction

Actually, what we dub the *se/się* construction is a family of constructions in both Croatian and Polish, including reflexive, reciprocal, and various middle meanings (for a cognitive review of Polish cf. e.g. [3, p. 73–76], [12, 30]; some Croatian *se* constructions are discussed in [1, p. 37–53]). In the reflexive and reciprocal meaning there is energy given off by the agent, and the *se/się* construction marks its direction — reflexively back onto the same agent (as in (11)) or reciprocally between two agents (in (12)).

- (11) *Marek się goli.*
Marek refl. shaves
‘Marek is shaving.’

- (12) *Momci se tuku.*
Boys refl. fight
‘The boys are fighting.’

In (11) the action of shaving is reflexive — a single subject is simultaneously the agent and the patient. As opposed to that, in (12) the action is reciprocal — there are two boys, and they are fighting with each other — both are giving and receiving blows. Note that the reflexive pronoun is a clitic form, and a non-clitic form *sebe/siebie* is also possible, as in:

- (13) *Marek myje siebie.*
Marek washes refl.
‘Marek is washing himself.’

Although (13) seems similar to (11), there is a significant difference. When the non-clitic form of the pronoun is used, the two participants (agent and patient)

remain distinct (and *siebie* may be conjoined with patient noun phrases; cf. [3, p. 74–75]). As opposed to this, the agent and the patient in the *se/się* construction are not distinguishable to the same degree as in a prototypical transitive construction, where the agent and the patient are two different participants [7, p. 638], [14, p. 211], [3, p. 74]. This is also evident in “reflexive” verbs that relate to emotional experiences, such as for instance, *bojati se/bać się* ‘be afraid of’, *brinuti se/martwić się* ‘worry’, etc., which cannot take the non-clitic pronoun *sebe/siebie*.

In middle constructions the *se/się* construction marks an unknown, unidentifiable or simply a non-salient agent ([3, p. 74–75], [30, p. 9–10]), as exemplified in (14)–(16):

- (14) *Drzwi się otworzyły.*
 door refl. opened
 ‘The door opened.’
- (15) *Smračilo se.*
 got-dark refl.
 ‘It got dark.’
- (16) *Ovdje se mnogo radi.*
 here refl. much works
 ‘Here people work a lot.’

All three constructions are similar because the (more or less prototypical) agent is outside the immediate scope of view. In Polish example (14) there is some energy which causes the door to open, but it is not coded in the sentence itself (i.e. it is outside the immediate scope of view). In Croatian example (15) there simply cannot be an agent — only the end point of the energy process is seen (in part also because of the *l*-participle). In other words, the change of state is “internal”. Finally, examples such as (16) (from Croatian) are traditionally called impersonal constructions — although the agent is identifiable, it is simply not salient (for various reasons).

2.3 Grammaticalization, the dative and the *se/się* construction

Essentially, both the development of the dative in Croatian and Polish as well as the development of the *se/się* construction follow a familiar path of grammaticalization. We consider grammaticalization to be the development from lexical to grammatical forms and from grammatical to more grammatical forms ([9, p. 1], [8, p. 2]).

As far as the Croatian and Polish dative are concerned, its benefactive (malefactive) and indirect object functions clearly developed from its earlier allative meanings ([16, p. 190–196], [17, p. 142], [29, p. 238–269]), which is accordance with the grammaticalization cline from allative to the dative [8, p. 37–38]. This is a process whereby a human or animate goal of motion, because of its animacy develops a dominion around it. This enables the animate dative to be construed as affected by any entity coming into its dominion, as seen above. The development from the dative to a possessive-like construction (as in (6) above; cf. [2], [28]) is the next step in this grammaticalization chain, and is also attested in a number of languages

[8, p. 103–104]. Finally, let us note that synchronic data for Croatian also indicate some formal changes in the dative grammaticalization chain (allative > dative > assessment/possessive). This is the change from a predominance of nouns in the dative (in the allative pattern) to an equal number of nouns and both non-clitic and clitic pronouns (in the central dative pattern) to predominantly cliticized forms in the assessment/possession pattern [27], which is also a sign of grammaticalization.

The examples of the Croatian and Polish *se/się* construction also exhibit a grammaticalization cline from less grammaticalized reflexive and reciprocal uses to more grammaticalized middle uses, which is a typical development [14]. Again, it is a part of a larger grammaticalization chain: reflexive > reciprocal > middle ([7, p. 628–639], [8, p. 252–254]). We have already noted that this chain is due to a change in the semantic distinguishability of the participants [7, p. 638]. Let us just add that this is also evident structurally in the distribution of the non-clitic vs. clitic forms of the reflexive pronoun (*sebe/siebie* vs. *se/się*). The non-clitic form is possible only when the two participants are notionally distinct (cf. example (13) vs. (11)), whereas the more grammaticalized reciprocal and middle patterns appear only with the clitic pronoun *se/się*.

3 The interaction between the *se/się* and the dative

In order to explore the interaction between the two constructions, we looked for examples of a noun or pronoun in the dative immediately followed by *se* or *się*. The Croatian data were taken from the Croatian National Corpus, and the Polish data were taken from the IPI PAN corpus. In this paper we will explore only the semantic characteristics of the data, and we will not go into any quantitative analyses. We will present the data based on the four groups of dative meanings: allative (and competitor), transfer, assessment and reference point/affectedness.²

3.1 Allative/competitor

In the allative + *se/się* group the dative construction is the abstract goal of motion and the *se/się* construction refers to the self-movement of the agent. Here is a Croatian example (example (1) is repeated as (17) for convenience):

- (17) *Ako se [Hrvatska] pridruži Partnerstvu za mir.*
 If refl [Croatia] joins Partnership-DAT for peace
 ‘If Croatia joins the Partnership for peace’

In this example abstract movement of the agent (Croatia) towards the dative goal (*Partnerstvu za mir* ‘Partnership for peace’) is signaled by using a reflexive verb *pridružiti se* ‘join’. As mentioned before, the reflexive verb here marks the fact that the agent and patient role are not very clearly distinguishable. In case of (abstract) movement this is taken to mean that the energy is agent-internal — i.e. that this is a case of self-propelled motion. Polish has no allative examples (except

² These four patterns were found in Croatian [27], and they largely coincide with patterns found by Dąbrowska for Polish [3].

from the one cited under (8)), but does include various competitor examples, which have a similar dynamic (example (10) is repeated for convenience):

- (18) *Stany Zjednoczone przeciwstawią się jakimukolwiek porozumieniu...*
 States United will oppose refl. any agreement-DAT
 ‘The United States will oppose any agreement...’

In (18) the *się* construction is again taken to mean that the energy is agent-internal, and the dative is the abstract goal, unaware of the trajector’s movement [27]. The allative/competitor + *se/się* construction is identical to the one shown in Figure 2, with the difference that the *se/się* construction clearly marks the energy as being internal to the trajector. Significantly, the trajector may only be an entity with internal energy, i.e. an entity that can move by itself. The energy marked by *se/się* agrees with the dative as the abstract goal of externalized energy. Whenever the *se/się* construction does not code internal energy (such as, for instance, with verbs requiring an experiencer subject, e.g. Croatian *bojati se* ‘be afraid of’, *brinuti se* ‘worry’ or Polish *bać się* ‘be afraid of’, *martwić się* ‘worry’) it cannot be combined with the dative as the abstract goal.³ This may also be one of the reasons why verbs of motion without the *se/się* marker (such as Croatian *ići* ‘go’, *hodati* ‘walk’ and Polish *iść* ‘go’, *chodzić* ‘walk’) are restricted in taking the dative — they are simply not energetic enough to combine with the allative sense.⁴ In other words, the allative+*se/się* is an energetic intransitive construction.

3.2 Transfer

In the transfer group (see example (4)) an entity is transferred into the dative’s dominion, and the *se/się* construction defocuses the agent which provides energy for the transfer:

- (19) *szczegółowy nadzór nad tym, komu daje się zezwolenie*
 ...detailed monitoring over that who-DAT give refl. permit
 ‘[it is necessary to]... closely monitor who is given the permit’
- (20) *stalno im se govori o tom problem*
 constantly they-DAT refl. tells about this problem
 ‘They are being told about this problem all the time.’

In (19) a permit is transferred to the dative entity (signaled by the pronoun *komu* ‘who-DAT’), and in (20) a message (incorporated in the meaning of the verb

³ For a treatment of *se* as a marker of an energetic construal cf. [24].

⁴ In Croatian this restriction refers to the fact that prepositionless allatives (such as *dodi mami* ‘come to mommy-DAT’) appear predominantly with animate and human datives, and are virtually impossible with inanimate datives (except in a few idiomatic examples). However, Croatian does allow allatives with prepositions with a full range of datives (*ići prema kući, mami* ‘go towards home-DAT, mommy-DAT’). Polish is much more restrictive in this sense — it simply has no allatives (and not many dative prepositions have survived). Finally, let us note that this energy explanation does not clash with the pragmatic interaction explanation offered in [27].

govoriti ‘speak’) is transferred to a person (signaled by the pronoun *im* ‘they-DAT’). Significantly, the verbs appearing in this pattern clearly signal the transfer of an entity, which is why various verbs of acquisition or loss are included here (such as *dati/dać* ‘give’, *posuditi/požyczyć* ‘lend/borrow’), verbs of communication (e.g. *govoriti/mówić* ‘talk’, *pričati/opowiadać* ‘tell’), verbs of energy transfer (*pomóc/pomóc* ‘help’, *oteżati* ‘make difficult’), etc. The dative is typically a person.

Several things need to be mentioned. Firstly, the trajector which moves towards the dative entity has no internal energy, i.e. the energy source causing the trajector’s movement is outside the trajector. In other words, the pattern exemplifies the billiard ball model (as opposed to self-propelled motion). Secondly, the *se/się* construction marks the fact that the energy source does not appear in the immediate scope of view. It detransitivizes an originally transitive construction. Thirdly, the dative entity is the recipient or the beneficiary of the transferred entity — in other words, the transferred entity is within the dative’s dominion, and the dative entity establishes mental contact with it. This is schematically shown in Figure 3.

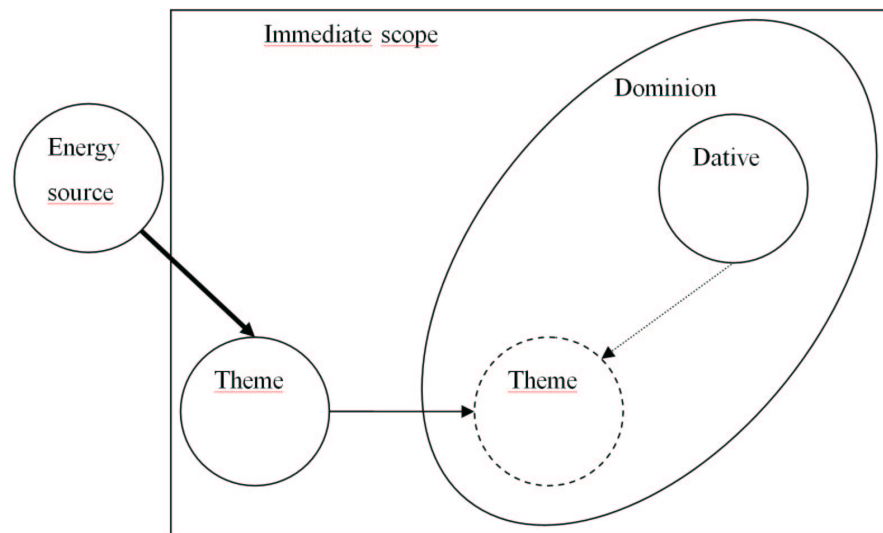


Fig. 3. Transfer + the *se/się* construction

The transfer pattern is naturally connected to the Polish construction *udać się* ‘succeed in doing something’, as in:

- (21) ...*udało jej się zadzwonić*
 succeeded her-DAT refl. to telephone
 ‘...she managed to call.’

Here the *się* construction marks that there is in fact no agent, and the dative signals an entity which is affected by the stroke of good luck which enabled a particular action. In other words, what happens here is that some “good energy”

is transferred towards the dative, removing any possible obstacles and allowing the dative to achieve something. Note that Croatian expresses this without a *se* construction (*uspjeti* ‘succeed’, *poći za rukom* ‘manage’).

3.3 Assessment

In the assessment sense, an entity within the dative’s dominion is assessed by the dative. The group contains reflexive verbs requiring an experiencer subject, such as Polish *ukazać się* ‘appear’, *podobać się* ‘like’, *wydawać się* ‘appear, seem’, *znudzić się* ‘be bored (with something)’, *śnić się* ‘dream’ or Croatian *činiti se* ‘appear, seem’, *pričiniti se* ‘seem to someone’, *sviđati se* ‘like’. Let us have a look at two examples:

- (22) ... *żołnierzom się znudził nasz program.*
 ...soldiers-DAT refl. bored our program.
 ‘... the soldiers were bored by our program.’
- (23) *meni se činiło kao da slušam istoga gospodina*
 me-DAT refl seemed like compl. listen same gentleman
 ‘... it appeared to me as if I were listening to the same gentleman.’

In both examples the dative assesses an entity within its dominion — in (22) the soldiers feel bored by a program, and in (23) one situation appears to the speaker as similar to another one. In this group the datives are naturally human (or at least animate), because they are experiencers. This also means that there is a dominion around the dative. The dative establishes mental contact with an entity (thing or relation) within its dominion. The mental contact spells out affectedness (see e.g. [26]). The reflexive verb again facilitates a construal in which there is no energy source within the immediate scope of view. However, this construal is one step further from the previous senses. Whereas in the previous senses there was an objective energy source which caused the movement or event, in the assessment sense the energetic construal is subjectified ([18, p. 128–132], [19], [22], [23]). The theme does not objectively move towards the dative. Instead, the dative (being the affected experiencer) **construes** the theme as entering its dominion without its own involvement — as if the abstract entry into the dative’s dominion was caused by an external energy source. This is illustrated in Figure 4.

In Figure 4 the dative is the conceptualizer of a theme which is within its dominion. The dative establishes mental contact with the theme and the theme’s subjective movement into the dative’s dominion, which is signaled by the *se/się* construction. Dative’s affectedness by the theme results in a “controlled reaction” to it [24, p. 29].

In Polish this pattern also includes examples where a relation signaled by the *się* construction is assessed, and the assessment is expressed by means of an adverb:

- (24) *dobrze mi się siedziało na kolanach Azara*
 well I-DAT refl. sat on knees Azar’s
 ‘It felt good sitting on Azar’s knees.’

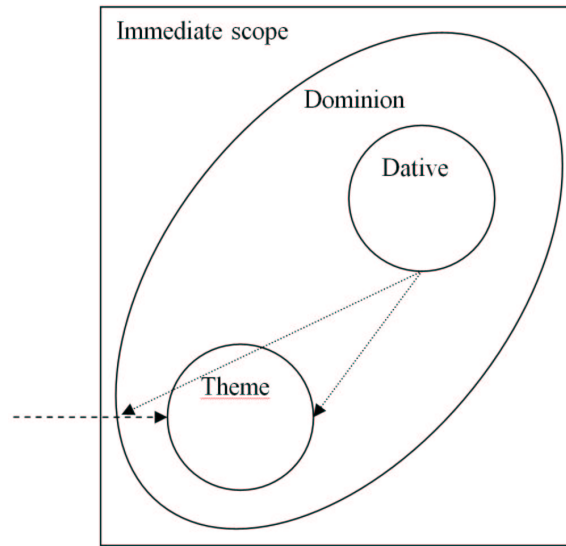


Fig. 4. Assessment dative + the *se/się* construction

In (24) the dative assesses the relation of sitting on Azar’s knees as something pleasant. The dative experiencer establishes mental contact with the relation of sitting on Azar’s knees, which just “happens to be” within its dominion. It is as if the action of sitting down onto Azar’s knees had already been performed by the coreferential agent, which now becomes an experiencer who cannot help but notice that the sitting is pleasant (see [30, p. 13] for a similar explanation). Again, note the involuntariness of the process.

Also note that the entire setting may be construed as the entity that the dative assesses, as in:

- (25) *wszystko po to, aby dzieciom się nie nudziło.*
 all because so that children-DAT refl. not bored
 ‘... and all this so that the children do not feel bored.’

In the Polish example (25) children (the dative) assess their own state of mind as being bored within a setting that they are in. The dative along with the *się* construction emphasize that the dative is affected by the setting where it got without its own control. The *się* construction is possible here because there is no agent (only an experiencer) — and this role is taken by the dative.

Croatian is somewhat different in this respect — the *se* construction may be used with various imperfective verbs to signify that the dative is experiencing an uncontrollable craving, but there seem to be no verbs such as the Polish *nudzić się* ‘be bored’ with a reflexive element.

- (26) *Meni kad se plače plačem...*
 me-DAT when refl. cries cry...
 ‘When I feel like crying, I cry...’

In the Croatian example (26) the dative has an incontrollable craving (the person wants to cry) within a particular temporal setting. Again, the *se* construction completely eliminates the agent, and a dative experiencer takes its place. The construal of craving is brought about by the confluence of grammatical factors (dative experiencer, the *se* construction, an imperfective verb) and by the fact that there is no competing construction such as the Polish *nudzić się* ‘be bored’. By the same token, Polish does not have constructions like (26), because it has a competing construction *chcieć się* ‘fell like (doing something)’.⁵

3.4 Reference point/affectedness

In the reference point/affectedness sense a process takes place, and it is construed as being within the dative’s dominion. Thus, the dative no longer receives a theme which was directed towards it, but witnesses an event happening within its dominion. The *se/się* construction is used to defocus the agent, and the dative is used to construe the event as within its dominion, and thus affecting it. Let us have a look at two examples:

(27) ... *wpatrywać się w telewizor (dopóki mu się nie zepsuł)*
 ... stare at refl. in TV (until him-DAT refl. not broke)
 ‘...staring at the TV (until it got broken on him).’

(28) *Razbila mu se šalica...*
 broken him-DAT refl. cup...
 ‘His cup got broken. / The cup got broken on him. /
 / He accidentally broke the cup’

The events happening in examples (27) and (28) are marked by the *se/się* construction, which means that the agent is outside the immediate scope of view, either because it is unknown or unidentifiable. The dative construes the relation marked by the *se/się* construction as happening within the dative’s dominion, i.e. the dative is construed as establishing mental contact with it and being affected by it. This is illustrated in Figure 5.

We call this the reference point/affectedness pattern because it is parallel to the “dative of possession” (see example (6) and its explanation) — in the “dative of possession” the dative is a reference point [21] to locate an entity, and here it is the reference point to locate an event. In both cases, dominion is construed around the dative, and because of mental contact with the entity/event, the dative is construed as affected. Finally, note that no objective relationship is required between the dative and the relation construed as within its dominion — something simply needs to be construed as happening within one’s dominion, and this is taken to mean that it affects the dative. This may have some pragmatic consequences. For instance, coding the actual agent as an experiencer in a construction with a defocused agent may be seen as a way of downplaying one’s own role in an

⁵ There may be other grammatical reasons behind this as well, notably the difference in the nature of Croatian and Polish imperfectives. For a cognitive analysis of Slavic aspect see [4]

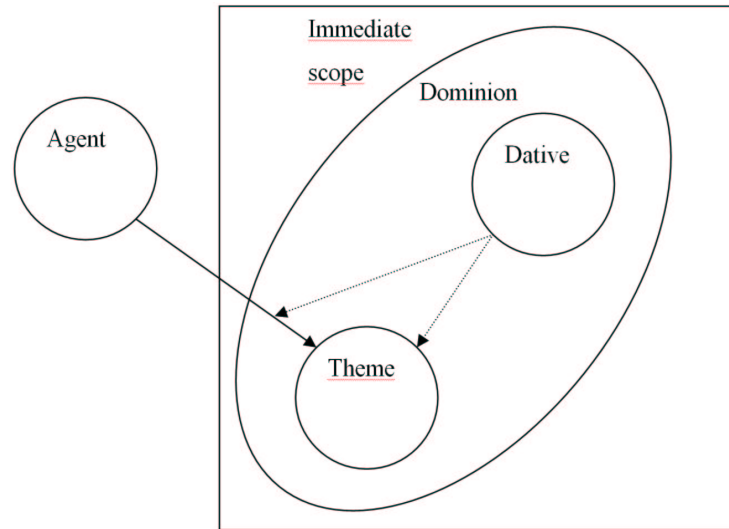


Fig. 5. Reference point/affectedness + the *se/sie* construction

undesirable event. Thus, (28) may imply that it was in fact the dative that broke the cup. Note that this is also connected to examples such as (21), where one's own role is also downplayed.

4 Discussion

In the allative + *se/sie* construction each of the elements simply does what it normally does. The allative signals an abstract goal of motion or energy. The variant of the *se/sie* construction which is used signals that the energy (resulting in motion or some other type of externalization) is internal to the trajector. Of course, the two constructions do limit each other's scope — you can use only the *se/sie* form of a verb signaling internal energy with the allative (that is why, e.g., *bać sie/bojati se* 'be afraid of' does not work) and vice versa. Still, the integration of the two constructions does not result in any additional constructional meaning that is not already symbolically signaled by each of the constructions — the search dominion and abstract goal are due to the allative, and the internal energy (and its externalization) are due to the *se/sie* construction.

In the transfer pattern, there is a combination of a different *se/sie* construction with the transfer sense of the dative. The *se/sie* construction marks the fact that the agent is outside the immediate scope of view, because it is unknown, unidentifiable or non-salient. The dative marks a recipient or a beneficiary which makes mental contact with a theme entering its dominion. Again the two constructions quite naturally limit each other's scope, inasmuch as only this particular *se/sie* construction (defocused agent of a verb marking transfer) may be used with this type of dative. The integration, however, still does not result in any additional construc-

tional meaning. The dative signals establishing mental contact with a theme put into motion by a defocused agent, which is defocused by the *se/się* construction.

In the assessment sense the dative establishes mental contact with an entity in its dominion, and is affected by it. The *se/się* construction allows an energetic construal of the theme — as if the theme were “moved” into the dative’s dominion by some unknown agent or energy source. In fact, this “movement” is subjective, and further emphasizes the dative’s affectedness. The type of affectedness is lexically expressed as assessment (using a reflexive verb as in (22), (23) and (25) or a verb and an adverb as in (24)). New constructional meaning not symbolically expressed by either component structure is seen in Croatian, where a combination of an imperfective reflexive verb and the dative signals someone’s uncontrollable craving (as in (26)).

The reference-point/affectedness construction also results in some additional meaning which is due to the construction, rather than to its components taken in isolation. The dative construction signals establishing mental contact with an event in its dominion, and the dative’s affectedness by it. The *se/się* construction that is used is a detransitivizing construction, signaling that the agent is defocused. The integration of the two may bring about a pragmatic confluence of the elements, where the defocusing of the agent and the affectedness of the dative causes an interpretation according to which the referent of the dative expression was the defocused agent. Thus, the agent’s role is downplayed, implying that the described event was an accident which affected the referent coded as the dative (as in (28)). The reference point/affectedness construction is similar to the last assessment construction inasmuch as it has constructional meaning. As opposed to it, however, the constructional meaning here is not symbolic, but rather of a pragmatic nature. In other words, only a particular context will determine the exact type of affectedness at play.

New constructional meaning in the last two patterns is made possible because of the grammaticalization of the dative and the *se/się* construction. Thus, the dative in the last two constructions has been stripped of all its detailed roles (abstract goal, recipient, beneficiary), and is construed as a schematic reference point which merely locates an entity within its dominion (as in the reference point/affectedness pattern) or assesses an entity located within its dominion (as in the assessment pattern). What is common to both is the schematic grammaticalized function of affectedness, which is present in both Croatian and Polish. Similarly, in the last two patterns, the *se/się* construction does not refer to a reflexive or reciprocal meaning, but rather to the more grammaticalized intransitive impersonal and inherent reflexive (as in the assessment pattern) or middle meaning (as in the reference point/affectedness pattern). What is common in both *se/się* patterns is the defocusing (or indeed complete removal) of the agent. The lack of agent intersects with the affectedness of the dative bringing about constructional meaning — uncontrollable craving in the assessment sense in Croatian or various pragmatic meanings in the reference-point/affectedness sense. New constructional meaning is possible only in those cases where the two constructions are very schematic — grammaticalized so far that neither contributes material detailed enough to “overpower” the other construction.

5 Conclusion

The aim of this paper was to look into the ways in which the *se/się* construction interacts with the dative construction. Based on a semantic analysis of examples from Croatian and Polish we showed that four patterns appear: the allative/competitor pattern, the transfer pattern, the assessment pattern and the reference point/affectedness pattern. With regard to emergent meaning, the patterns may be divided into those without emergent meaning (allative/competitor and transfer), and those with emergent meaning (assessment and reference point/affectedness). We showed that emergent meaning was symbolic and pragmatic in nature, and that it appeared only when the grammaticalized senses of each construction were combined. The reason behind this is that the grammaticalized senses are more schematic in relation to non-grammaticalized ones.

This conclusion begs the question whether the influence of grammaticalization onto the emergent meaning is just a fluke connected with the incorporation of the two constructions presented here, or perhaps a more systematic phenomenon. Based on the fact that in this paper similarities were found between two languages — Croatian and Polish, it would be reasonable to assume that this may be a systematic factor. Still, no final judgments can be made before a more general look is taken, exploring other complex grammatical constructions unconnected to this one.

Bibliography

- [1] Belaj, B. (2004). *Pasivna rečenica*. Osijek: Filozofski fakultet Sveučilišta Josipa Jurja Strossmayera.
- [2] Cienki, A. (1993). Experiencers, Possessors, and Overlap Between Russian Dative and u + Genitive. *Berkeley Linguistics Society* 19: 76–89.
- [3] Dąbrowska, E. (1997). *Cognitive Semantics and the Polish Dative*. Berlin, New York: Mouton de Gruyter.
- [4] Dickey, S. M. (2000). *Parameters of Slavic Aspect: A Cognitive Approach*. Stanford: Center for the Study of Language and Information.
- [5] Fillmore, Ch. J., Kay, P., O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of “let alone”. *Language* 64, no. 3: 501–538.
- [6] Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument*. Chicago: University of Chicago Press.
- [7] Givón, T. (1990). *Syntax: A Functional-Typological Introduction*. Vol. 2. Amsterdam, Philadelphia: John Benjamins Publishing Co.
- [8] Heine, B., Kuteva, T. (2002). *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.
- [9] Hopper, P. J., Traugott, E. C. (2003). *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.
- [10] Janda, L. A. (1993). *A Geography of Case Semantics: The Czech Dative and the Russian Instrumental*. Berlin – New York: Mouton de Gruyter.
- [11] Janda, L. A. (2002). Cases in collision, cases in collusion: the semantic space of case in Czech and Russian. In *Where One's Tongue Rules Well: A Festschrift for*

- Charles E. Townsend, ed. Laura A Janda, Steven Franks, and Ronald Feldstein, 13:43–61. Indiana Slavica Studies. Columbus, Ohio: Slavica.
- [12] Kardela, H. (2007). Event structure: a force dynamics/absolute construal account. In *Cognition in language. Volume in honour of Professor Elżbieta Tabakowska*, ed. Władysław Chłopicki, Andrzej Pawelec, and Agnieszka Pokojńska, 150–167. Kraków: Tertium.
- [13] Kay, P., Fillmore, Ch. J. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* 75, no. 1: 1–33.
- [14] Kemmer, S. (1993). *The middle voice*. Amsterdam, Philadelphia: John Benjamins Publishing Co.
- [15] Kučanda, D. (1996). What is the dative of possession? *Suvremena lingvistika* 22, no. 41: 319–332.
- [16] Kuryłowicz, J. (1964). *The inflectional categories of Indo-European*. Heidelberg: Carl Winter, Universitätsverlag.
- [17] Kuryłowicz, J. (1977). *Problèmes de linguistique indo-européenne*. Wrocław, Warszawa, Kraków, Gdańsk: Zakład Narodowy im. Ossolińskich.
- [18] Langacker, R. W. (1987). *Foundations of Cognitive Grammar*. Vol. 1. 2 vols. Stanford: Stanford University Press.
- [19] Langacker, R. W. (1990). Subjectification. *Cognitive linguistics* 1, no. 1: 5–38.
- [20] Langacker, R. W. (1991). *Foundations of Cognitive Grammar*. Vol. 2. 2 vols. Stanford: Stanford University Press.
- [21] Langacker, R. W. (1993). Reference-point constructions. *Cognitive linguistics* 4, no. 1: 1–38.
- [22] Langacker, R. W. (2000). *Grammar and conceptualization*. Berlin – New York: Mouton de Gruyter.
- [23] Langacker, R. W. (2005). *Obserwacje i rozważania na temat zjawiska subiektywifikacji*. Trans. Małgorzata Majewska. Kraków: Universitas.
- [24] Maldonado, R. (2002). Objective and subjective datives. *Cognitive linguistics* 13, no. 1: 1–65.
- [25] Rudzka-Ostyn, B. (1996). The Polish dative. In *The Dative. Volume 1: Descriptive studies*, ed. William van Belle and William van Langendonck, 341–394. Amsterdam, Philadelphia: John Benjamins Publishing Co.
- [26] Stanojević, M.-M., Geld, R. (2008). The dative in Croatian as a dominion phenomenon. *Études Cognitives* 8: 95–108.
- [27] Stanojević, M.-M., Tuđman Vuković, N. (forthcoming) Dominion, subjectification, and the Croatian dative.
- [28] Šarić, L. (2002). On the semantics of the “dative of possession” in the Slavic languages: an analysis on the basis of Russian, Polish, Croatian/Serbian and Slovenian examples. *Glossos* 3. <http://seelrc.org/glossos/>.
- [29] Šarić, L. (2008). *Spatial Concepts in Slavic. A Cognitive Linguistics Study of Prepositions and Cases*. Wiesbaden: Harrassowitz Verlag.
- [30] Tabakowska, E. (2003). Those notorious Polish reflexive pronouns: a plea for Middle Voice. *Glossos* 4. <http://seelrc.org/glossos/>.
- [31] Wierzbicka, A. (1986). The meaning of a case: a study of the Polish dative. In *Case in Slavic*, ed. R. D. Brecht and J. S. Levine, 386–426. Columbus, Ohio: Slavica.

MAKSIM DUŠKIN

Institute of Slavic Studies, Polish Academy of Sciences, Poland

CONCERNING EXPONENTS OF ADNUMERAL APPROXIMATION IN POLISH AND RUSSIAN

Abstract. This article attempts to present the most important results of an analysis of Polish and Russian exponents of adnumerative approximation (e.g. Pol. *około*, *przeszło*, Rus. *более*, *примерно*, etc.).

“Approximation” is defined here as a mechanism of numeric attribution, based on designating a segment in a series of numbers, instead of designating one particular point. This definition encompasses several different ways to designate a segment. All of those ways have separate exponents in Polish and Russian. Analysis indicates that exponents of the same sense can differ in terms of connectivity, semantics and stylistics. These differences should be taken into consideration while establishing Polish and Russian translational equivalents.

Keywords: Approximation, Approximate, Polish, Russian.

1 Introduction

This article will attempt at presenting the most important results of the analysis of Polish and Russian exponents of adnumerative 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. The last part of the study

was to analyse individual exponents. Observable syntactic, stylistic and semantic differences between them were of particular interest to the study.

The analysis of the exponents of approximation leads to the establishment of Polish-Russian and Russian-Polish translation equivalents. Such a task is not an easy one and requires further, substantial research, and this work will restrict itself to presenting general problems which arise during this process.

2 What is approximation?

Approximation is widely connected with numbers, but it can also be understood in a broader context ([8], [14], [12]). 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” [8, p. 29].

This article, 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’).

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 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 [13] 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)¹.

Secondly, expressions such as *dużo/m mało*, recognized by Grochowski [8, 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 [4: 329].²

3 Classifications of exponents of adnumeral approximation

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 powyżej*, 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. The list of known Polish and Russian exponents of these six types is presented below:

(1) Exponents, setting the exclusive lower boundary of the segment (exponents of the sense ‘<little> more than X’)

POL. *po, ponad, powyżej, przeszło, więcej niż*;

RUS. *более (чем), больше (чем), за, свыше*;

Examples:

Pol. *Na salę wszedł człowiek po czterdziestce/powyżej czterdziestki, Jan ma ponad 2 metry wzrostu; Piotr ma przeszło/więcej niż czterdzieści lat,*

Rus. *Ян ростом больше/более двух метров, Ян ростом более чем/больше чем два метра, Яну за сорок* (mainly denoting people’s age), *В заповеднике живет свыше 30 представителей этого редкого вида птиц*

¹ More on the phenomenon of roundedness, see: [6].

² Cf.: “[...] *A is tall = A’s size (possibly: which is conspicuous) is greater than his possible size which would not attract attention*” [4: 329].

(2) Exponents, setting the inclusive lower boundary of the segment (exponents of the sense ‘X or <little> more than X’):

POL. *co najtmniej, minimum, najtmniej, nie tmniej niż, przynajtmniej*;

RUS. *как минимум, минимум, не менее (чем), не меньше (чем), по меньшей мере, самое меньшее*;

Examples:

Pol. *Jan nie był w Krakowie co najtmniej/minimum od 5 lat, Pobyt Jana w Krakowie trwał najtmniej/nie tmniej niż/przynajtmniej 3 tygodnie*;

Rus. *Ян не был в Кракове как минимум/минимум/не менее чем/не меньше чем/по меньшей мере/самое меньшее пять лет; Ян не был в Кракове не меньше/не менее пяти лет*

(3) Exponents, setting the exclusive upper boundary of the segment (exponents of the sense ‘<little> less than X’):

POL. *bez mała, blisko, tmniej niż, niecały* (with units of measurement and nouns such as *tysiąc, milion* etc.), *niemal, niespełna, pod, prawie*

RUS. *без малого, едва (ли) не, менее (чем), меньше (чем), неполный;³ под, почти; чуть (ли) не.*

Examples:

Pol. *Szliśmy bez mała/blisko/tmniej niż/niemal/niespełna/prawie godzinę, Szliśmy niecałą godzinę, Jan ta pewnie pod dwa metry wzrostu*

Rus. *Мы шли без малого/едва ли не/меньше чем/менее чем/почти/чуть ли не час, Мы шли менее/меньше часа, Мы шли неполный час; Температура под 38 градусов*;

(4) Exponents, setting the inclusive upper boundary of the segment (exponents of the sense ‘X or <little> less than X’):

POL. *co najwyżej, do, maksimum, najwyżej, nie więcej niż*;

RUS. *до, максимум, не более (чем), не больше (чем), от силы, самое большее*;

Examples:

Pol. *Jan powiedział, że będzie czekał co najwyżej/maksimum/najwyżej/nie więcej niż 10 minut, Zabójcy grozi do 10 lat rozbawienia wolności*;

Rus. *Ян сказал, что подождет максимум/не более чем/не больше чем/самое большее десять минут, Ян сказал, что подождет не более/не больше десяти минут, До самолета оставалось от силы десять минут, Убийцу ждет до 10 лет лишения свободы.*

(5) Exponents, designating the centre of the segment (exponents of the sense ‘<little> less than X or X or <little> more than X’):

POL. *circa, gdzieś, jakieś, koło, tmniej więcej, około, plus minus, rzędu, w przybliżeniu, z/ze*;

RUS. *где-то, около, порядка, приблизительно, примерно, с*; inversion (e.g. *человек пять*).

Examples:

³ This exponent appears in conjunction with some quantitative nouns, e.g. *неполный литр, неполная тысяча рублей*.

Pol. *Lista zawiera około/koło dwustu książek, Lista zawiera circa/gdzieś/jakieś /mniej więcej/plus minus/w przybliżeniu/z dwieście książek; To jest kwota rzędu 100 milionów złotych*

Rus. *Список содержит где-то/приблизительно/примерно двести книг, Список содержит около/порядка двухсот книг; Дом напротив строили с год, Список содержит книг двести,*

Exponents of the type (6), setting both boundaries of the segment (exponents of the sense ‘X–Y’) are co-appearances of two numeric expressions in both languages, like 10–20, Pol. *metr-dwa*, Rus. *метр-два* (ellipsis of the numeral 1). Some types of numeric expressions, e.g. Polish phrases such as *dwadzieścia ileś* (‘21–29’), *dwadzieścia kilka* (‘22–29’) and others could also be considered as exponents of this type.

Individual exponents of numeric approximation differ not only in the way they define the segment they are representing, but also in several other features.

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 (Мельчук [11], Grochowski [7]). 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.

It is noticeable, that some researchers only consider expressions which, in connection with a numeral, set **small segments of numbers** (e.g. Wierzbicka [14]) to be exponents of approximation, while others treat expressions which, while setting segments, do not **denote their size** (e.g. Bogusławski and Karolak [5], Мельчук [11], Grochowski [8]). 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.

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*)⁴;
- 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 above-mentioned groups, see Table 1 on page 207⁵.

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 some exponents, treated as adnumeral by other researchers (e.g. the expression *bez mała* by Grochowski in [7]), have been transferred to the group of not strictly adnumeral exponents (cf. *bez mała familiarne stosunki*).

As aforementioned, the concept of numeric approximation includes a number of ways to designate a segment, all of which have Polish and Russian exponents. Six preliminary ways to designate a segment were discussed (6 approximate senses):

1. ‘<little> more than X’ (*przeszło sześćdziesiąt*)
2. ‘X or <little> more than X’ (*co najmniej sześćdziesiąt*)
3. ‘<little> less than X’ (*bez mała sześćdziesiąt*)
4. ‘X or <little> less than X’ (*nie więcej niż sześćdziesiąt*)
5. ‘<little> less than X or X or <little> more than X’ (*około sześćdziesięciu*)
6. ‘from X to Y’ (*20–30*)

Here the component ‘little’ is optional.

The analysis of individual exponents of these six senses suggests that:

- all exponents of the sense ‘<little> less than X or X or <little> more than X’ contain the component ‘little’ in their meaning;
- none of the exponents, expressing the alternative ‘<little> less than X or X’ or ‘X or <little> more than X’, contain the component ‘little’ in their meaning;

⁴ Wierzbicka [14] 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.

⁵ 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.

- most exponents of the sense ‘<little> less than X’ contain the component ‘little’ in their meaning;
- one component of the sense ‘<little> more than X’, the Polish *przeszło*, supposedly contains the component ‘little’ in its meaning, while other exponents, both Russian and Polish, do not.

These observations have given birth to the conclusion, that eight, and not six ways of designating a segment, which are represented by Polish and English exponents of approximation, should be mentioned. These are the following ways:

1. ‘little less than X or X or little more than X’ (e.g. *około*)
2. ‘less than X’ (e.g. *mniej niż*)
3. ‘little less than X’ (e.g. *niespełna*)
4. ‘more than X’ (e.g. *więcej niż, ponad*)
5. (?) ‘little more than X’ (*przeszło*)⁶;
6. ‘less than X or X’ (e.g. *do*)
7. ‘X or more than X’ (e.g. *minimum*)
8. ‘X–Y’ (*20–30*)

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 *W każdej z grup jest mniej niż 20 osób* (it could be 14, 18, 19 and even 10).

The lists of exponents of approximation, included in this article, bear some corrections in relation to suggestions of other authors.

First of all, the list of Russian exponents of the sense ‘less than X or X’ has been enlarged by the strictly adnumeral expression *от сущи*, which is not mentioned in other works. Moreover, the list of exponents of the sense ‘little less than X or X or little more than X’ includes the Russian *порядка* and Polish *rzędę*.

Secondly, some expressions were eliminated from the list of exponents of approximation (e.g. Rus. *эдак/этак*, which is treated as an exponent of the sense ‘little less than X or X or little more than X’ by some researchers).

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

⁶ It should be mentioned, that *przeszło* requires further research, as to the component ‘little’. The research could indicate, that the exponent does not contain the component in that meaning. Then, the sense ‘little more than X’ should be eliminated from the list, because the rest of Polish and Russian expressions, designating the segment ‘more than X’, does not contain the component.

of units of measurement. Authors of some works (e.g. Wierzbicka in [14]) 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.

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.

4 The issue of establishing Polish-Russian and Russian-Polish translational equivalents

While establishing Polish-Russian and Russian-Polish translational equivalents, each exponent should be individually analysed. Some (but only some) exponents are limited in terms of connectivity with numeric expressions of certain types (e.g. Russian *за сорок <лет>*; cf. **за сорок человек*), while some are stylistically marked (e.g. Russian *с/со* — colloquial, while *приблизительно* — rather literary), and individual exponents carry additional components in their meaning (such components, that are absent from the meaning of other exponents of approximation). Additionally, conjunctions of some exponents with numeric expressions appear only in specific syntactic roles.

Let us begin by discussing the latter characteristic. Some exponents of approximation impose a certain case upon a nominal construction (numeral + noun). These are prepositions (e.g. Rus. *около*, Pol. *do*) and some Russian comparative forms (*более, больше, менее, меньше, свыше* <but not *более чем, больше чем* etc.>).

Cf. e.g. *Пришло около* → (Genitive) *двадцати человек* (*Przyszło około dwudziestu osób*), *Мат* → (Genitive) *кош 20 złotych*. Such approximate descriptions only appear in positions, where a noun in Nominative or Accusative would be required [11, p. 367].

Cf. *Врач вылечил* (→Accusative) *Диму/свыше двадцати пациентов*,
vs. **Врач помог* (→Dative) *свыше двадцати пациентам*; **Врач помог* (→Dative) *свыше двадцати пациентов*.

Exponents that do not impose the usage of a specific case can appear in other positions (cf. *Врач помог более чем двадцати пациентам*). This quality should be taken into consideration while translating e.g. the Polish exponent *około*. The exponent is often used as a particle (and then it does not impose the case) [2: *około* <vol. 1, p. 1153>], while the Russian *около* is a preposition. So, the Polish exponent

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, jakiś, koło, 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, niespełna; pod</i> ; Rus. <i>под</i>	Pol. <i>bez mała, niecały, niemal, prawie</i> ; Rus. <i>без малого, едва (ли) не неполный, почти, 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. <i>–</i>	–	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>

Table 1. Division of exponents of approximation

około can appear in such a position in the sentence, where, for example, the Locative case is required (*Tekst się zmieścił na około pięćdziesięciu stronach*), while the Russian *около* cannot (**Текст уместился на около 50 страниц*, **Текст уместился на около 50 страниц*). So the exponent *około* in the sentence above should be translated rather into e.g. the Russian *примерно*, or alternatively *приблизительно* or *где-то*: *Текст уместился примерно на 50 страницах* (/ *приблизительно/где-то на 50 страницах*). *Примерно, приблизительно* and *где-то* are not limited in distribution in the abovementioned way, which is specific to prepositional exponents (including *около*).

It is also good to mention, that some exponents contain limitations of connectivity with numeric expressions of certain types. Let's consider some examples.

The Russian exponent *с* usually occurs together with names of units of measurement containing the ellipsis of the numeral 1 (*Посплю с часок*), while it rather does

not co-occur with other numeric expressions. Because of this, while translating the Polish exponent *z/ze* in the sentence *Przyjdzie z pięć osób*, inversion should rather be used instead of the Russian *с* (e.g. *Придет человек пять*, but not *?Придет с пять человек*).

Russian *за* and Polish *po* appear almost exclusively with descriptions of an object's age (mostly connected with people). Cf. *Ему уже за пятьдесят*; *Do pokoju wszedł facet po pięćdziesiątce*. The connectivity of these exponents is thus far more limited than the connectivity of other Polish and Russian exponents of the sense 'more than X', cf. e.g.: **Пришло за пятьдесят человек* (but: *Пришло более 50 человек*, *Пришло свыше 50 человек* is possible, etc.).

So, while establishing translational equivalents of an exponent, the types of numeric expressions this exponent can be conjoined with should be taken into consideration.

It is also worth mentioning, that some exponents are stylistically marked. For example, the Russian *с* is more colloquial, so the exponent should not be used while translating Polish exponents, appearing in official documentation, academic text, the news etc. Cf. e.g.: *Budowa trwała około roku* vs. **Стройка продолжалась с год* (improper e.g. in the context of television news). Therefore, stylistic characteristics must also be taken into consideration while searching for translational equivalents for a given Polish or Russian exponent.

And finally, it should be noted that some exponents contain not only components of designation of a numeric segment, but also some additional components in their meaning. For example, according to the authors of the work *Новый объяснительный словарь синонимов русского языка*, the Russian expression *без малого* communicates not only the approximate content, but also the quantitative assessment 'a lot' („больше нормы или больше того, что ожидалось”) [1: 278; see also 279–280]. E.g. the expression *без малого 50* approximately states the number, while informing, that the number is large. The structure of the Polish exponent *niepełna* also contains the quantitative assessment component. In the explication of *niepełna*, proposed by Grochowski (*S-ów jest niepełna P*), this component is stated as "... sądzę, że S-ów jest mało" [8: 72]. It seems that a quantitative assessment component also contains the Russian exponent *от силы* (*Пришло от силы 5 человек* — the speaker informs that 5 or a few less people came, at the same time stating that this is not many).

The presence of additional components within the meaning of exponents should be taken into consideration while establishing translational equivalents. For example, if the Russian sentence *Пришло от силы 5 человек* is translated into Polish as *Przyszło nie więcej niż 5 osób*, we lose the information, contained in the original sentence, that the number of people is considered to be small. Whereas if we translate *Przyszło nie więcej niż 5 osób* into *Пришло от силы 5 человек*, we distort the original message, adding components, absent from the original.

5 Summary

Approximation is a mechanism of numeric designation, based on defining a segment instead of one point of the arithmetic sequence. Thus defined, the concept

encompasses a number of ways to define a segment. Each of these has numerous exponents in both Polish and Russian.

The analysis in this article indicates that Polish and Russian exponents setting a segment in the same way should not be automatically treated as equivalents. Exponents of the same sense could differ in terms of connectivity, semantics and stylistics. Each exponent should be analysed individually, and only then its equivalents in other languages can be established. Establishing translational equivalents in the field of exponents of approximation is not an easy task, and requires further work, because both analysed languages contain a significant number of such exponents, while not all of them have been sufficiently described in the literature of the subject.

Bibliography

- [1] Апресян, Ю. Д. (ed.) (2003). *Новый объяснительный словарь синонимов русского языка. Третий выпуск*, Языки славянской культуры, Москва.
- [2] Bańko, M. (ed.) (2000). *Inny słownik języka polskiego PWN*, t. 1–2, PWN, Warszawa.
- [3] Bogusławski, A. (1966). *Semantyczne pojęcie liczebnika i jego morfologia w języku rosyjskim*, Zakład Narodowy im. Ossolińskich, Wrocław.
- [4] Bogusławski, A. (1994). *Measures are measures. In defence of the diversity of comparatives and positives*, In: Bogusławski, A., *Sprawy słowa* (Word matters), Veda, Warszawa, p. 323–329.
- [5] Bogusławski, A., Karolak, S. (1973). *Gramatyka rosyjska w ujęciu funkcjonalnym*, wydanie drugie, poprawione, Wiedza Powszechna, Warszawa.
- [6] Duszkin, M. (2003). *Dokładniej o dokładności: “zaokrągloność” dokładnych określeń ilościowych*, Prace Filologiczne XLVIII, Warszawa, p. 133–142.
- [7] 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.), Polska Akademia Nauk, Instytut Języka Polskiego, Kraków, p. 31–37.
- [8] Grochowski, M. (1997). *Wyrażenia funkcyjne. Studium leksykograficzne*, Polska Akademia Nauk, Instytut Języka Polskiego, Kraków.
- [9] Koseska-Toszewa, V. (1991). *O języku-pośredniku i badaniach konfrontatywnych*, In: *Problemy teoretyczno-metodologiczne badań konfrontatywnych języków słowiańskich*, Běličova, H., Nieszczimienko, H., Rudnik-Karwatowa, Z. (eds.), Instytut Słowianoznawstwa PAN, Warszawa, p. 7–19.
- [10] Koseska-Toszewa, V. (1993). *Gramatyka konfrontatywna rosyjsko-polska. Składnia*, SOW, Omnitech Press, Warszawa.
- [11] Мелчук, И. А. (1985). *Поверхностный синтаксис русских числовых выражений*, Wiener Slawistischer Almanach, Sonderband 16, Wien.
- [12] Сахно, С. Л. (1983). *Приблизительное именование в естественном языке*, In: *Вопросы языкознания*, 1983:6, п. 29–36, Академия наук СССР, Издательство “Наука”, Москва.

- [13] Санников, В.З. (1999). *Русский язык в зеркале языковой игры*, In: Языки русской культуры, Москва.
- [14] Wierzbicka, A. (1991). *Cross-Cultural Pragmatics. The Semantics of Human Interaction*, Mouton de Gruyter, Berlin, New York.

JOANNA SATOŁA-STĄSKOWIAK

Institute of Slavic Studies, Polish Academy of Sciences, Poland

**TRANSLATING INTO SOMETHING THAT DOES NOT EXIST...
LITERARY WAYS OF TRANSLATING POLISH SENTENCES
WITH UNINFLECTED PERFECT PARTICIPLES
INTO THE BULGARIAN LANGUAGE**

Abstract. The aim of this work is a syntactic analysis of Bulgarian constructions translated from sentences with uninflected perfect participles of the poem *Pan Tadeusz* (“Mr Thaddaeus”) by Adam Mickiewicz. The analysis is based on the observation known from many other dissertations that if syntactic constructions, whose centres are both uninflected and inflected participles with regard to the meaning and syntactic function, cannot correspond in the translation to syntactic constructions with a participle as the centre, they correspond to clauses in precisely defined types of compound sentences.

Keywords: semantic layer, concurrent character, interpretation of the sequence of events, indicator of connection, translating

1 Introduction

A starting point for this work is an attempt to show literary ways of translating Polish sentences with uninflected perfect participles into the Bulgarian language.

Because of the difference between the system of Polish and Bulgarian participles, caused by the disappearance of some of the Bulgarian participles (including the Polish equivalent of the perfect participle discussed here), Błaga Dymitrowa’s work as a translator of *Pan Tadeusz* became quite complicated. She had to express often original participial constructions with the help of other linguistic means. The poem by Mickiewicz contains a huge amount of adjectival and adverbial participles.

2 Anteriority in the original and its equivalents

2.1 Anteriority in the original and in the translation

2.1.1 Translation equivalent has the form of the Bulgarian participle with -л (минало свършено причастие) The Bulgarian past participle perfective describes a completed activity — perfective. It is because of the fact that

this participle is formed from the verbs of perfective aspect. A group of compound sentences presented below contains examples where one of the clauses has a participle with *-a*, thus emphasising the anteriority of an event with respect to another event or state.

Księga I

Wprzagliwszy w swój rydwan orły złote zamiast srebrnych,
Od puszczy Libijskich *latał* do Alpów podniebnych,
Ciskaając grom po gromie, w Piramidy, w Tabor,
W Marengo, w Ulm, w Austerlitz.

орлите златоглави **впрегнал** в колесница,
в Либийската пустиня *литва* като птица,
пронесъл гръм след гръм до Алпите поднебни
в Улм, в Аустерлиц, в Маренго.

Księga II

Wreszcie **podniosłszy** trzonek z powagą do góry
Jak łaskę marszałkowską, *nakazał* milczenie.

— sequence of tenses: *podniosłszy* – *nakazał*

и *викна* в гнева си,
маршалски жезъл сякаш **вдигнал** над главите:

— reversal of order: *викна* – *вдигнал*

The group under discussion also contains sentences which, despite the correspondence participle- participle, convey a modified content. The modification occurs on the level of the Bulgarian predicate different from the original.

Księga VI

I jeszczeż po tym wszystkim, com tobie powiedział,
Będziesz spokojnie, ręce **założywszy** *siedział*,
Gdy działać trzeba!

И тъкмо днес, когато всичко ще разкроя,
ръце си **скръстил**, Литва *чакаш* на тепсия,
а трябва сам...

Księga XI

pomiędzy ciemnemi
Puszczami chłop, którego dziady i rodzice
Pomarli **nie wyjrzawszy** za lasu granice,

дълбоко **врасли** в тая почва, **невидели**
дори насъне други някакви предели,
тук раждан и погребан,

2.1.2 Translation equivalent has the form of a personal form of the verb

Compound sentence in the translation

Sentences belonging to this group are temporal compound subordinate clauses. The degree of bond in this kind of closed compounds is expressed by the indicators of connection *цом* and *след туй*. Each of the indicators of connection presented in this group carries a different lexical meaning, which is discussed below.

Equivalent *след туй* as an indicator of the sequence of sentences
An equivalent *след туй* is a proof of the sequence of sentences. One of the described events finished earlier (was anterior), the other, as the indicator of connection shows, took place after it. The time separating both events did not have to be distant, provided that it was long enough to allow to realise — complete the first one.

Księga II

I dwakroć **uderzwszy** głowy obie mocne
Jedną o drugą jako jaja wielkanocne,
Rozkrzyżował ramiona na kształt drogoskazu
I we dwa kąty izby rzucił ich od razu;

ударя две глави плешиви и корави
като яйца червени, чукнати за здраве,
след туй ръце *простря* — същ пътепоказател —

Equivalent *-цом* as an indicator of anteriority

Щом according to *Граматика на съвременния български книжовен език – морфология* “свърза подчинени темпорални причинни, и условни изречения с главното изречение.” [24]. This conjunction “Wnoszą informację o uprzedniości, ale także o odstepie czasu, jaki upływa między obydwoma zdarzeniami, komunikowanymi w zdaniu — jest to bardzo mały odcinek czasu, sekwencja faktów ma charakter zbliżony do równoczesności. Zdania te są bardzo bliskie konstrukcjom z *kiedy* / *когато*, w których sygnalizowana jest koincydencja zdarzeń.” [14]. The sentences below are examples of linking a subordinate clause with a main clause.

Księga II

Zaś jastrząb, pod jasnymi wiszący błękity,
Trzepie skrzydłem jak motyl na szpilce przybity,
Aż **ujrzawszy** wśród łąki ptaka lub zająca,
Runie nań z góry jako gwiazda spadająca.

А там увиснал ястреб пърполи в зенита —
с карфица пеперуда сякаш е забита;
щом жертва си **съзре** в далечната ливада,
се стрелва изведнъж – звезда като че пада.

In the examples presented here one can also find sentences where the translated content has been modified by means of a change of meaning of participial equivalent.

Księga II

Tu **wszedłszy** starzec głowę zadumana *skłonił*
I twarz zakrył rękami, a gdy ją odsłonił,
Miała wyraz żałości wielkiej i rozpaczu.

Щом старецът **погледна** тия рамки слепи,
глава *обори* мълком в треперящи шепки.
Когато пак я вдигна, Графът трепна смаян.

2.2 Anteriority in the original and its disturbances in the translation

2.2.1 Translation equivalent has the form of the present participle The fact that the perfect participle is replaced by the present participle changes the sentence considerably. The activity which was expressed before by the perfect participle and now by the present participle is almost concurrent with the activity in the main clause in the translation.

Księga VII

Ja, z konia **zsiadłszy**, zaraz *padłem* na kolana
Dziękując Panu Bogu.

Аз, **слизайки** от коня, *паднах* на колени,
благодарих на бога.

— the sequence of facts has the character similar to simultaneousness

Księga XI

Potem **spuściwszy** oczęta
Dodała:

Тук **свеждайки** очи, едва — едва *зашушна*:

In these examples there are also sentences in which lexical modification of the text on the level of the predicate of the main clause took place, for example:

Księga XII

Podkomorzy rusza
I z lekka **zarzuciwszy** wyloty kontusza,
I węża podkręcając, *podał* rękę Zosi

Подкоможи *слуша*,
запретвайки встрани полите на контуша.

2.2.2 Compound sentence in the translation In this case compound sentences are linked by means of an indicator of connection *u* or non-conjunctively by means of a comma. Both indicators of connection suggest that the sentences have the concurrent coordinating character. However, the examples listed below show that the concurrency of two events is not so obvious.

Conjunction *u* as an indicator of concurrency

I have mentioned in the article *The realization of uninflected present participle in the Bulgarian literature — on the basis of “Pan Tadeusz” by Adam Mickiewicz* that an indicator of connection *u* characterises inclusive coordinate clauses. However, it is not an unequivocal statement. An indicator *u* has not got only one function. Sometimes a different interpretation (discussed below) is allowed. One of the two possible interpretations is determined by the type of activity indicated by the verbal forms.

In [2] W. Doroszewski and B. Wieczorkiewicz suggest that semantic relationship between inclusive coordinate clauses may be based on the sequence in time and on the tangency, i.e. getting closer in time. This tangency is represented by examples classified into group A in which verbal forms imply concurrency of the activities in time. Group B allows the possibility of the sequence of activities indicated by the verbs.

According to [24, 25] the sequence of activities indicated by the verbs can be expressed by means of conjunctions. Such cases can be observed in sentences linked by conjunction *u*.

A. Conjunction *u* indicates the concurrent character

I have included in this group typical examples of inclusive sentences in which it is possible to do simultaneously activities indicated by the verbs. There are no physical contraindications of concurrency which would allow the sequence of events and states. Here are the examples:

Księga I

Wlowszy kropelkę wina w szklanę panny Róży,
A młodszej **przysunąwszy** z talerzem ogórki,
Rzekł:

Бавно Подкоможи
пред малката си щерка краставици сложи,
доля на Ружа вино и така **настави:**

Księga II

Policmajster powinność służby swej rozumiał,
Bardzo się nad zuchwalstwem czynownika zdumiał
I **odwiódłszy** na stronę, po bratersku *radził*,
By przyznał się do winy i tym czyn swój zgładził.

Сърдит от дързостта на тоз чиновник дребен,
отведе го встрани и бърз *свет му даде:*
греха си да признае, тъй да го заглади.

B. Conjunction *u* allows the interpretation of the sequence of events

In this group the activities indicated by the verbal forms cannot be physically done at the same time. It leads to an assumption that the indicated activities did not happen simultaneously but one after the other (although the time separating them could have been just a moment).

Księga I

1.

Przeprosiwszy go grzecznie, na miejscu swym *siadła*
Pomiędzy nim i stryjem, ale nic nie jadła;

След туй **се извини** почитателно и *седна*
на оня празен стоп. Но блюдо не погледна.

— She apologised and [then] sat down.

2.

Dał mu poważnie rękę do pocałowania
I w skroń **ucałowawszy**, uprzejmie *pozdroził*:

Той за целувка – строг – подава му десница,
младежкото чело **целува** с думи прости
и *се обръща* пак към драгите си гости.

— She kisses and [then] gives her regards. (In this example the aspect of the verbs has been additionally changed.)

Comma as an indicator of concurrency

This indicator of connection, despite earlier interpretations, similarly as conjunction *u* has more than one function (cf. [20]). In the analysis of sentences linked by a comma one can consider a different possibility — recognition of the indicator of connection discussed here as an indicator of the sequence of sentences. The choice of interpretation is determined by the type of activities indicated by the verbal forms and sometimes even by physical impossibility.

A. Comma indicates the concurrent character

Księga VI

Jako raz zapozwany szlachcic z Telsz, Dzindolet,
Rozkazał mu, **oparłszy** o piersi pistolet,

Така веднъж под съд бе даден Дзиндолета
от Телши. Той о гръд **опря** му пистолета,
застави го да върши разни щуротии:

B. Comma allows the interpretation of the sequence of events

Księga I

1.
Właśnie dwukonną bryką wjechał młody panek
I **obiegłszy** dziedziniec *zawrócił* przed ganek,
Wysiadł z powozu;

И его, шляхтич млад веднъж *пристигна* тука,
дома **обиколи**, на входа не почука.

— He arrived, [then] walked around (lexical modification)

2.
Sędzia, z boku rzuciwszy wzrok na Tadeusza
I **poprawiwszy** nieco wylotów kontusza,
Nalał węgryzna i rzekł:

А чичото за миг Тадеуш с поглед смери,
поправи си с ръка широките ревери,
унгарско си *наля* и каза:

— He sat more comfortably, [then] poured

2.2.3 Singular sentence in the translation The group presenting singular sentences in the translation is not too numerous. It includes sentences in which the perfect participle of the original became a predicate and a predicate of the main clause was expressed by e.g. a prepositional phrase in the form of an adverbial of manner or an adjectival phrase. In a singular sentence the anteriority of events in relation to other, happening after them, was disturbed. In singular sentences one can observe only the concurrency. The anteriority, after the change of the perfect participle into a predicate and a predicate into a prepositional phrase, disappeared.

Księga I

1.
Lecz na wzmiankę Warszawy *rzekł* **podniósłszy** głowę:
но **вдига** в миг глава при името В аршава: endverse

— Lack of a predicate of the main clause in the translation.

2.
Inaczej bawiono się w drugim końcu stoła,
Во там **wzmógłszy się** nagle stronnicy Sokoła
Na partyję Kusego bez litości *wsiedli*:

По-иначе се развличат гостите отсреща.
Ловджийската им разпра **става по-гореща**.
Охулват Куси те, Сокола хвалят само.

Adverbial function of manner expressed by a prepositional phrase with abstractum

Księga I

Wojski na ostrym końcu wśród myśliwych siedział,
Słuchał zmrużywszy oczy, słowa nie powiedział,
 Choć młodzież nieraz jego zasięgała zdania,
 Bo nikt lepiej nad niego nie znał polowania.

На най- последно място Войски омърлушен,
с премрежени очи във споровете *вслушан*,
 при все че е най-вещ по всеки лов във Литва
 и младежта при спор до него се допитва –
 ни дума не пророня.

Księga II

To **wyrzekłszy** Sędziego *ścisnął* za kolana.

При тия думи той му *тупна* колената.

Księga XII

1.

I **skłoniwszy się** grzecznie, w pierwszą parę *prosi*.

И сучейки мустак, пред Зося *свс покана*
 за първа двойка той учтиво **се покланя**.

2.

Pan Sędzia **skłoniwszy się** *opuszczył* biesiadę;

Нануца със поклон пан Съдията пира

Translation equivalent has the form of the passive participle

One has to also distinguish cases in which translation equivalent had the form of the passive participle. Sentences belonging to this group are on the borderline between the concurrent and anterior sentences. Such state is caused by the participles themselves as they have a built-in perfective aspect of an activity.

Księga VI

Załamal ręce Książdz zdziwiony.

Wlepiwszy oczy w Sędzie, **ruszywszy** ramiony,
Rzekł: „To gdy Napoleon wolność Litwie niesie,
 Gdy świat drży cały, to myślisz o procesie? (...)”

Робак, **втренчен** в Съдията,
 ръце закърши смаян, **сви** си рамената
 и *рече*; — Знъчи, тъй: Родината ще бъде
 свободна, а пък той си мисли да се съди!

Structure of a compound sentence conveyed by two separate clauses

In this group one should focus on the relation between predicates. The construction of a compound sentence was divided. The content of the participle was conveyed by a separate sentence which does not grammatically indicate the anteriority of the event described by this participle. The predicates of these sentences are not formally linked.

Księga II

Gładził ją ręką, podszedł i jeszcze raz nisko
Skłoniwszy się, rzekł smutnie: „Mopanku. Panisko, —
 Daruj mi, że tak mówię, Jaśnie Grafie Panie,
 To jest mój zwyczaj, nie zaś nieuszanowanie: (...)

Повторно **се покланя**, приближава с почит.
 И със безкрайна скръб словата му *се точат*:
 — Мопанко! Че така наричам те, прости ми!
 Ти знаеш, туй са мои думици любими.

Księga VII

Ciągnął mowca **spojrzawszy** bystro, dość dwie słowie,
 “Nieprawdaż?” — “Prawda!” rzekli.

Сам Робак неведнъж *загатна* с думи смътни.
 — Да, знаем! — Е, добре! — Ораторът под вежда
със поглед прозорлив събраните **изглежда**.

— reversal of order events in relation to the original.

3 Summary

The problem of Polish sentences with the perfect participles and their translation into the Bulgarian language differs from sentences with uninflected present participles. This difference is caused by the lack of the perfect participle in Bulgarian. Sentences containing forms *part. praet. act. with -l* are the closest to the perfect participles. I have found few examples of sentences with this participle.

Deliberations, contained in this article, on the sentences with perfect participles in *Pan Tadeusz* and their equivalents in the Bulgarian translation by Błaga Dymitrowa lead to the conclusion that the translator has not used the whole set of temporal forms to express anteriority and resultativeness. However, this fact does not determine whether her rendering could be considered as a mediocre work. Thanks to many other qualities of the translation, it is now recognised as one of the best renditions of *Pan Tadeusz* in Bulgaria.

Despite the asymmetry of systems in the semantic layer on the level of content the rendering is close to the original. The translator has used diverse linguistic means. She has displayed a great sense and command of the Polish language.

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, 404 pp.
- [3] Георгиев, С. (1993). *Българска морфология*, Абагар, Велико Търново.
- [4] Grybosiova, A. (1978). *Rozwój funkcji imiesłowów nieodmiennych w języku polskim. Związki z nomen*. Wrocław.
- [5] Grzegorzczkova, R. (1999). *Wykłady z polskiej składni*. Wydawnictwo Naukowe PWN, Warszawa.
- [6] Jadacka, H. (1991). Imiesłowowy równoważnik zdania — norma a praktyka językowa. In *Prace Filologiczne*, 1991, R. XXXVI, pages 183–193.
- [7] 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., editors, pages 97–113, Warszawa.
- [8] Jodłowski, S. (1971) Stanowisko imiesłowów w systemie gramatycznym. In *Studia nad częściami mowy*, Warszawa.
- [9] Klemensiewicz, Z. (1937). *Składnia opisowa współczesnej polszczyzny*. Kraków.
- [10] Klemensiewicz, Z. (1963). *Zarys składni polskiej, wyd. IV*. Wydawnictwo Naukowe PWN, Warszawa.
- [11] Klemensiewicz, Z. (1980). *Historia języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- [12] Klemensiewicz, Z. (1982). *Składnia, stylistyka, pedagogika językowa*. Wydawnictwo Naukowe PWN, Warszawa.
- [13] Klemensiewicz, Z. (1983). *Podstawowe wiadomości z gramatyki języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- [14] Korytkowska, M. (1998). Uprzedniość, rezultatywność a zdania czasowe w języku bułgarskim i polskim. In *Studia z Filologii Polskiej i Słowiańskiej*, 34 SOW, pages 205–224, Warszawa.
- [15] Мицкевич, А. (1979). *Пан Тадеуш. Шляхтишка история от годините 1811–1812 в дванадесет стихотворни книги*. Народна Култура, София.
- [16] 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.
- [17] Москов, М. (1974). *Български език и стил*. Наука и изкуство, София.
- [18] Saloni, Z. (1971). *Błędy językowe w pracach pisemnych uczniów liceum ogólnokształcącego. Próba analizy językoznawczej*. Warszawa.
- [19] Saloni, Z., Świdziński, M. (1998). *Składnia współczesna języka polskiego, wyd. IV*. Wydawnictwo Naukowe PWN, Warszawa.
- [20] Satoła-Staškowiak, J. (2009). Translation of Polish Uninflected Present Participle In Bulgarian Literature — on the Basis of “Pan Tadeusz” (Mr Thaddaeus) by Adam Mickiewicz. In Koseska-Toszewa, V., Dimitrova, L., Roszko, R., editors,

- Representing Semantics in Digital Lexicography, Mondilex Fourth Open Workshop, Warsaw, Poland, 29 June – 1 July, 2009, Proceedings*, pages 180–188, Institute of Slavic Studies Polish Academy of Sciences, Warszawa.
- [21] Szober, S. (1968). *Gramatyka języka polskiego*, wyd. IX, Wydawnictwo Naukowe PWN, Warszawa.
- [22] Śmiech, W. (1971). *Funkcje aspektów czasownikowych we współczesnym języku ogólnopolskim*, Łódź.
- [23] Wróbel, H. (1975). *Składnia imiesłowów czynnych we współczesnej Polszczyźnie*. Prace Naukowe Uniwersytetu Śląskiego w Katowicach, Uniwersytet Śląski, Katowice.
- [24] Тилков, Д., Стоянов, С, Попов, К. (editors) (1993), *Граматика на съвременния български книжовен език. Морфология*, Издателство на Българската Академия на Науките, София, p. 452, p. 446–475.
- [25] Тилков, Д., Стоянов, С, Попов, К. (editors) (1993), *Граматика на съвременния български книжовен език. Синтаксис*, Издателство на Българската Академия на Науките, София, p. 293–294.

JULIA MAZURKIEWICZ-SUŁKOWSKA¹
AGATA MOKRZYCKA²

¹The faculty of Philology, University of Lodz, Lodz, Poland

²The Humanistic faculty, The Maria Curie-Skłodowska University, Lublin, Poland

FROM THE WORKS ON THE BULGARIAN-POLISH DICTIONARY OF VERBO-NOMINAL ANALYTICAL CONSTRUCTIONS

Abstract. The following article discusses the work on the *Bulgarian-Polish dictionary of verbo-nominal analytical constructions*. The goal of the authors was to create a dictionary that would address contemporary lexicographical issues and thus be of use to both interpreters and linguists.

Keywords: lexicography, verbo-nominal analytical constructions, normative problems

1 Introduction

The dictionary should be helpful for students - Bulgarists and Slavists alike, as the existing Slavic lexicography doesn't approach the phenomena of verbo-nominal analytical constructions with an adequate precision.

The dictionary may also be of great help in learning or perfecting Polish and Bulgarian. It will, without a doubt, help develop language competence and make it easier to identify "what connects with what", or to be more exact, which verbs create which constructions of the VNC type (verbo-nominal constructions).

2 The criteria for defining the verbo-nominal construction (VNCs)

For the purpose of our work on the *Bulgarian-Polish dictionary of verbo-nominal constructions* we consider a phrase to be a VNC only if it is semantically indivisible, and consists of a verbal part and a nominal exponent of the predicative notion. Therefore a VNC is a unit traditionally called periphrastic predicate of the type: Pl. *brać udział* (*Jan bierze udział w konferencji.*), Bul. *вземам участие* (*Иван взема участие в конференция.*), or a construction in which the nominal part is in the nominative case and formally is the subject of the sentence e.g. Pl.: *radość*

kogoś ogarnia (*Radość ogarnęła Marię*), Bul. *обхва на някого радост* (*Мария я обхва на радост*) [2, p. 17–30], [4, p. 51–52]; [8, p. 13–19].

The criteria for defining VNCs were taken from P. Zmigrodzki's *Właściwości składniowe analitycznych konstrukcji werbo-nominalnych w języku polskim* [8].

The primary condition of defining a construction as a VNC is the presence of an abstract noun. The term “abstract noun” here refers to non-objective notions, that is nouns that are the exponents of a predicative expression. The consequence of using this criterion is that it excludes from the analysis those expressions which contain in their structure a primary concrete noun like: *rzucić okiem*, *wydać owoce*.

The next criterion is the lack of the possibility of dividing a VNC into a compound sentence. According to Lewicki's test [7], a lexical unit is a VNC if transformed into a compound sentence with *że* or *to* it creates an unacceptable sentence e.g.: *wpaść w zachwyty* — **Jan wpadł w to, co jest zachwytem; pałać nienawiścią* — **Jan pała tym, że nienawidzi Piotra*. Jędrzejko emphasizes [5] that the test is not always accurate. For example some causative constructions of the type: *doprowadzić do szaleństwa* — *Jan doprowadził do tego, że Anna oszalała* do not fulfill this criterion.

In the literature concerning the subject one of the conditions of classifying a construction as a VNC is often the existence of a synonymous synthetical unit in the form of a verb (cognate or not). There are, however, several constructions failing to fulfill also this condition e.g.: *udzielać audiencji*, *zaspokoić ambicje*.

Apart from the criteria defining the set of VNCs we must also delimit them from idiomatical constructions. As is known, the feature that differentiates an analytical construction from an idiom is mainly the clear dual structure of a VNC: verbaliser – predicator. An idiom on the other hand is an indivisible (both syntactically and semantically) expression. For example the meaning of the idiom *puszczać farbę* is not the sum of the meanings of its parts. The next feature distinguishing a VNC from an idiom is the possibility of nominalizing an analytical construction reducing it to its nominal part, whereas idioms lack this feature. See [8]. Eg.:

- (1a) *Student przeprowadza analizę dzieła literackiego.*
- (1b) *Pochwalono go za poprawne przeprowadzenie analizy.*
- (1c) *Pochwalono go za jego analizę.*
- (2a) *Szef wystął mnie na zieloną trawkę.*
- (2b) *Wystanie mnie na zieloną trawkę było dla mnie ciosem.*
- (2c) * *Zielona trawka była dla mnie ciosem.*

The expected content of the dictionary is approximately 2000 entries consisting of about 6000 Bulgarian VNC's with their Polish equivalents.

Every analytical construction introduced in the dictionary consists of a nominal part expressed by an abstract noun (a nominalized exponent of the predicate) and a verbal part (the verbalizer of the predicative content). The second role may be fulfilled by a verb with a structural function only (maximally close to an auxiliary verb) or a verbalizer introducing additional predicative content: causativity, passivity, intensity, or additional aspectual content (inchoativity, durativity, iterativity, momentality, terminativity).

The units enclosed in the forthcoming studium represent:

- basic vocabulary, which is common for different language variants of Bulgarian and Polish and represents the communicative basis;
- official vocabulary, that is bookish, formal, specialistic (legal and administrative language), publicistic, elevated (it is said that VNCs characterize the official register and are one of stylistical markers of such language variants);
- colloquial, unofficial vocabulary typical for daily language communication.

3 Macrostructure of the dictionary

The main part of the *Bulgarian-Polish dictionary of verbo-nominal analytical constructions* is a bilingual Bulgarian-Polish study of these specific verbo-nominal constructions. The entries are systematized according to the Cyrillic alphabet, and Bulgarian abstract nouns are the entry words. Apart from the Bulgarian-Polish part concerning Bulgarian VNCs and their Polish functional equivalents there are also two appendixes: an index of verbalizers and an index of synthetical equivalents. Enclosing those two lists is supposed to positively affect the functionality of our study, facilitate its usage and, last but not the least, provide information about existing equivalents of analytical constructions.

Below is an excerpt of a main page from the dictionary. The structure of the entries will be characterized.

ПОДБОР — *dobór*

◆извършвам//извърша ~; ◆правя//направя ~ | на някого/на нещо | dokonywać//dokonać doboru | *kogoś/czegoś* | , robić//zrobić *dobór* | *kogoś/czegoś* | ; dobierać//dobrać | *kogoś/coś* |

Извършихме/направихме подбор на най-стойностните детайли от историческа гледна точка. — Dokonałiśmy doboru najbardziej wartościowych z punktu widzenia historii szczegółów.

ПОДВИГ — *bohaterski czyn*

◆извършвам//извърша ~; ◆творя//сътворявам//сътворя ~ dokonywać//dokonać *bohaterskiego czynu*

Той извърши подвиг в името на отечеството. — On dokonał czynu *bohaterskiego* ku chwale Ojczyzny.

ПОДГОТОВКА — *przygotowanie*

◆извършвам//извърша ~; ◆правя//направя ~ | за нещо | czynić//poczynić *przygotowania* | *do czegoś* | ; przygotowywać//przygotować *coś*

Извършихме сериозна подготовка за това тържествено събитие. — Poczyniliśmy poważne *przygotowania* do tego uroczystego wydarzenia.

ПОДДРЪЖКА — wsparcie

◆оказвам//окажа ~ | на някого/на нещо | okazywać//okazać wsparcie | komuś/czemuś | , udzielać//udzielić wsparcia | komuś/czemuś | ; wspierać//wesprzeć | kogoś/coś |

Родителите оказват поддръжка на децата си. — Rodzice udzielają wsparcia swoim dzieciom.

ПОДИГРАВКА — kpiny, drwiny

◆отправлям//отправля ~ | към някого — ; правя//направля ~ | с някого/с нещо | robić (sobie) | z kogoś/z czegoś | żarty, stroić (sobie) | z kogoś/z czegoś | żarty, wystawiać//wystawić | kogoś/coś | на kpiny; kpić//zakpić | z kogoś/z czegoś | , drwić//zadrwić | z kogoś/z czegoś |

Това, че мразил чалга, не ти дава основание да си правиш подигравки с хората. — To, że gardzisz muzyką disco polo, nie daje ci podstaw by stroić z innych kpiny.

There will be an index of the verbalizers in the dictionary consisting of a list of all the verbs constituting VNCs with appropriate abstract nouns with which they pair.

This part systematizes the material from the verbal side. The user of the dictionary will easily gain knowledge about the connectivity of a certain verb with nouns, and in addition to this quickly check if the unit he seeks is analyzed in the dictionary.

An excerpt from the Index of verbalizers:

- | | |
|------------------------------|------------------------------|
| ● ВДИГАМ//ВДИГНА | ● ВЗЕМАМ//ВЗЕМА |
| ● вдигам//вдигна буря | ● вземам//взема връх |
| ● вдигам//вдигна врява | ● вземам//взема думи |
| ● вдигам//вдигна гюрултия | ● вземам//взема души |
| ● вдигам//вдигна наздравница | ● вземам//взема за образец |
| ● вдигам//вдигна обсада | ● вземам//взема мерки |
| ● вдигам//вдигна поглед | ● вземам//взема на подбив |
| ● вдигам//вдигна скандал | ● вземам//взема надмощие |
| ● вдигам//вдигна тост | ● вземам//взема наем |
| ● вдигам//вдигна тревога | ● вземам//взема направление |
| ● вдигам//вдигна шум | ● вземам//взема отпуска |
| | ● вземам//взема пауза |
| | ● вземам//взема под аренда |
| | ● вземам//взема под внимание |

An index of the synthetical equivalents of the Bulgarian VNCs is also expected to be built in the dictionary. It will contain both synthetical equivalents that are cognate to the abstract noun, and equivalents that have a different root but function as synonymous to the VNC in question.

An excerpt from the Index of synthetical equivalents of VNCs:

авансирам	— давам//дам аванс
агитирам	— върша//извърша//извършвам, правя//направя, вода// проведа//провеждам агитация
агонизирам	— в агония съм
акламирам	— правя//направя акламации
акцентирам	— поставям//поставя акцент
алармирам	— вдигам//вдигна аларм
амнистирам	— давам//дам, правя//направя амнистия
анализирам	— извършвам//извърша, правя//направя, подлагам//по- дложа на анализ
анатемосвам	— налагам//наложа анатема
ангажирам се	— поемам//поема ангажимент
анексирам	— извършвам//извърша анексия; подлагам//подложа на анексия
апелирам	— отправям//отправя апел
апелирам	— правя//направя апелация

4 Microstructure of the dictionary (the shape of an entry)

The main unit of a lexicographical study is an entry, marked with a headword written in bold capital letters, in our case the abstract noun in Bulgarian. Then after a dash comes its Polish equivalent (or equivalents). If the headword is not neutral an appropriate label is introduced.

The main part of each entry consists of: the verbalizers organized according to Cyrillic alphabet with the appropriate abstract nouns and their polish equivalents (VNC other language forms). The arguments required for generating a correct sentence are given next to the predicator. Every VNC is accompanied with a sentence exemplifying its use and its polish translation. The inspiration for those examples was taken from excerpted from the internet “living” utterances of these units in publicistic texts, notes from daily press, discussions at internet forums. The examples are generated in such a manner so that they comprehensively illustrate the functioning of the unit in a text.

АКЦЕНТ — akcent^a

◆поставям//поставя ~ | *върху нещо* | *przen. kłaść//położyć akcent* | *na co* | *na czymś* | *akcentować//zaakcentować* | *coś* |

Председателят на Комисията за защита от дискриминация постави акцент върху превенцията. — Przewodniczący Komisji Ochrony przed Dyskryminacją położył akcent na prewencję.

^a **The headword** with capital letters, the polish equivalent after a dash with non-capital letters.

The entries are organized according to the Cyrillic alphabet.

АГРЕСИЯ₁^a — agresja, wroga postawa

◆ **проявявам**// **проявя** ~ przejawiać agresję, być agresywnym
Гладните кучета проявяват агресия. — Głodne psy są agresywne.

АГРЕСИЯ₂^a — napadź zbrojna, atak

◆ **извършвам**// **извърша** ~ | *срещу някого* / *срещу нещо* | dokonywać//dokonać, dopuszczać się//dopuszczyć się agresji | *przeciwko komuś* / *przeciwko czemu* |
Германия през 1939 г. е извършила агресия срещу Полша. — Niemcy w 1939 r. dopuściły się agresji przeciwko Polsce

^a **Equiform words** are defined each in its own entry with a down numerical index by the headword.

АЛИБИ — алиби, **przen.**^a **usprawiedliwienie**

◆ **давам**// **дам** ~ | **на някого/на нещо** | dawać//dać | komuś /czemuś | alibi; *przen.* usprawiedliwiać//usprawiedliwić | kogoś /coś |
Търговец дава alibi на подкупни прокурори. — Kupiec daje alibi przekupnym prokuratorom.

^a **Labels** are written in italic type.

БЕЗДЕЛИЕ — nieróbstwo

◆ **отдавам се**// **отдам се на** **~**^a oddawać się//oddać się nieróbstwu; próżnować
Да ползва богатата на живота, като се отдава на безделие – това е идеалът на френския аристократ. — *Korzystać z uroków życia, oddając się nieróbstwu – oto cel francuskiego arystokraty.*

^a **A tilde (~)** after verbalizers is written instead of the abstract noun.

АВАНС — czołowa pozycja, prowadzenie, przewaga

◆ **вземам**// **взема** ~ wychodzić//wyjść, wysuwać się//wysunąć się na czołową pozycję, na prowadzenie

Петкратният шампион Михаел Шумахер взе аванс преди днешната квалифика-

ция за Гран при на Австрия.^a — Pięciokrotny mistrz Michael Schumacher objął

prowadzenie przed dzisiejszą kwalifikacją do Grand Prix Australii.^b

◆ **губя**// **загубя**// **загубвам** ~ tracić//stracić przewagę

Мидълзбро в първата среща загуби аванс от два гола. — Middlesbrough w pierwszym spotkaniu staciło przewagę dwóch goli.

◆ **задържам**// **задържа** ~ zachowywać//zachować czołową pozycję, prowadzenie

Еврото не задържа своя аванс на световната валутна сцена. — Euro nie zachowało swojej czołowej pozycji na światowej scenie walutowej.

◆ **имам** ~ mieć przewagę

Френската състезателка имаше аванс от две точки след първата серия. — Francuska zawodniczka miała przewagę dwóch punktów po pierwszej serii.

^a After the polish equivalents, written in a new line goes **the example sentence** in Bulgarian.

^b **Polish translations** of the examples follow after a dash.

АТАКА — atak

◆^aвпускам се//впусна се в ~; ◆^aхвърлям се//хвърля се в ~ | срещу някого/срещу нещо | ruszać//ruszyć, rzucić się//rzucić się do ataku | przeciwko komuś/przeciwko czemuś | | na kogoś/na coś | przypuszczać//przypuścić atak | przeciwko komuś/przeciwko czemuś | | na kogoś/na coś |

Стиснах зъби и без колебание се хвърлих в атака срещу спама. — Zaciśnięłem zęby i bez wahania rzuciłem się do ataku przeciwko spamowi.

◆ минавам//мина в /към^b ~; ◆ преминавам//премина в /към^b ~ | срещу някого /срещу нещо | przechodzić//przejsć do ataku | przeciwko komuś/przeciwko czemuś | | na kogoś/na coś |

Защо, когато ти свършат аргументите, или се почувстваш леко засегнат, така агресивно преминаваш в атака, изместваш темата? — Dlaczego, kiedy kończą ci się argumenty albo czujesz się dotknięty, tak agresywnie przechodzisz do ataku?

◆ отбивам // ^cотбия ~; ◆ отразявам // ^cотразя ~; ◆ парирам ~

odpierać//odeprzeć^c atak

Полския полк отбива атаката на немците и преминава в контраатака, разбивайки немските части. — Polski pułk odpiera atak Niemców i przechodzi do kontrataku, rozbijając niemieckie oddziały.

◆ подемам//подема ~; ◆ предприемам //предприема ~ | срещу някого/срещу нещо | [przypuszczać//przypuścić atak]^d, przedsięwziąć atak | na kogoś/na coś | | przeciwko komuś/przeciwko czemuś | ; [atakować//zaatakować]^e | kogoś/coś |

САЩ предприе атака срещу обекти на Ал Кайда в Багдад. — USA przypuściło atak na obiekty Al-Kaidy w Bagdadzie.

◆ провеждам//проведа ~ | [срещу някого/срещу нещо] | ^f przeprowadzać//przeprowadzić atak [przeciwko komuś/przeciwko czemuś | | na kogoś/na coś] | ^f

Страната проведе атака срещу вярата и дейността на евангелските църкви. — Państwo przeprowadziło atak przeciwko wierze i działalności Kościołów ewangelickich.

^a **Verbalizers** in aspectual pairs are written in a new line after a caro mark (◆)

^b Synonymous verbal units are put in a line separated with a semicolon.

The lines of the verbalizers are in **alphabetical order** in accordance with the first verbalizer.

^c **Aspectual verb forms** are separated with a double slash (//)

^d After every line of synonymous Bulgarian verbalizers are given their **Polish equivalents**

^e If there are any **synthetical polish equivalents** (in the form of a verb) they are introduced after a semicolon.

^f After the verbalizers, between vertical lines (| |) are given the **exponents of all the implied objective / nominalized predicative arguments** in the form of indefinite pronouns.

It is worth noting that the graphical concept of the entries is, in the current stage of work, a minor issue and will be modified after all the material is analyzed. The above shows only a preliminary concept of the appearance of the entries.

5 Normative and prescriptive criteria

There are several normative issues related to the use of VNCs. When the first studies of the subject were undertaken, there arose voices among scientists that the usage of such analytical constructions is a clear manifestation of the downfall of the language norm. M. Kniaginina claims, for example that constructions of the type: verb + noun are : „ the result of a pitiful stylistical search supposed to satisfy the tendency for analitism” [6, p.150]. Therefore it doesn't surprise us that while working with VNCs one sometimes encounters an example that isn't corresponding to his language intuition. It is the result of the fact that with the growing frequency of VNCs the connectivity of the verbalizers is broadening. What we mainly encounter here are constructions created under the influence of the connectivity in synonymous and close in meaning constructions e.g.: **dokonać zaliczki* (see *dokonać wpłaty*), **wydać rezultaty* (see *wydać plony / owoce*). Żmigrodzki correctly emphasizes that the attempts of reducing this tendency are done in vain, because the differences are purely normative and not categorial. See [8].

There exist several VNCs incorrect from the normative point of view, for example constructions of the type:

- *ulec poprawie* (*ulegać* can co-occur only with nouns denoting unfavorable processes e.g. *ulec pogorszeniu*);
- *wykazać się ignorancją / brakiem kompetencji* (the verbalizer *wykazać się* denotes abstract nouns meaning positive features e.g. *wykazać się wyrozumiałością*);
- *dokonać kradzieży / włamania* (*dokonać* denotes noble acts e.g. *dokonać czynu bohaterkiego* or acts sanctioned by the public opinion e.g. *dokonać analizy*. See [3]).

In the dictionary we follow the rule of abiding Polish language correctness, so constructions considered incorrect will not be analyzed. The only deviation from this rule are VNCs with the verbalizer *dokonać* referring to actions forbidden by the law (e.g. *dokonać morderstwa*). Such units will be labeled as official. Their special treatment is due to them being deep-rooted in the official and judicial language, and omitting them would make translating official texts difficult.

translated by Jakub Banasiak

Bibliography

- [1] Anusiewicz, J. (1978). *Konstrukcje analityczne we współczesnym języku polskim*, Zakład Narodowy im. Ossolińskich, Wrocław: 67–101.
- [2] Bogusławski, A. (1978), *Jednostki języka a produkty językowe. Problem tzw. orzeczeń peryfrastycznych*, In Szymczak, M. (Ed.) *Z zagadnień słownictwa współczesnego języka polskiego*, Zakład Narodowy im. Ossolińskich, Wrocław: 17–30.
- [3] Buttler, D., Kurkowska, H., Satkiewicz, H. (1986). *Kultura języka polskiego. Zagadnienia poprawności leksykalnej*, PWN, Warszawa.

- [4] Jędrzejko, E. (1992). *Słownictwo tzw. analityczne w opisie leksykalnym (propozycja opisu i klasyfikacji)* In Markowski, A. (Ed.) *Opisać słowa*, Elipsa, Warszawa: 50–61.
- [5] Jędrzejko, E., Loewe, I., Żmigrodzki, P. (Eds) (1998). *Słownik polskich zwrotów werbo-nominalnych. Zeszyt próbny*, Energeia, Warszawa.
- [6] Kniaginina, M. (1963). *Struktury opisowe — znamienna cecha stylu dziennikarskiego*, Język Polski XLIII, Wydawnictwo UJ, Kraków: 148–157.
- [7] Lewicki, A.M. (1976). *Wprowadzenie do frazeologii syntaktycznej. Teoria zwrotu frazeologicznego*, Wydawnictwo Uniwersytetu Śląskiego, Katowice.
- [8] Żmigrodzki, P. (2000). *Właściwości składniowe analitycznych konstrukcji werbo-nominalnych w języku polskim*, Wydawnictwo Uniwersytetu Śląskiego, Katowice.
- [9] Żmigrodzki, P. (2000). *Z zagadnień składni wewnętrznej polskich zwrotów werbo-nominalnych. Dystrybucja i funkcja czasowników posiłkowych*. Polonica 20, Wydawnictwo PAN, Warszawa: 83–99.

EWA MICZKA

Université de Silésie, Katowice

LES STRUCTURES SITUATIONNELLES ET INFORMATIONNELLES DE DISCOURS

Abstract. “**Situational and Information Structures of Discourse**”

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 [3]. 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 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.

Keywords: Situational structures, information structures, theme, rheme, cognitive event, frame.

1 Introduction

L’objectif de cet article est de décrire les relations entre deux types de structures — situationnelles et informationnelles — qui interviennent dans la compréhension de discours.

Nous allons donc tout d’abord expliquer ce que nous comprenons par l’acte d’interpréter un discours — lui-même conçu comme texte plongé dans un contexte de communication.

Deuxièmement, nous allons présenter la conception des structures situationnelles qui résulte de l’application des notions de cadre de l’expérience introduite

par E. Goffman, [3] et d'événement cognitif proposée par R. Langacker [5], à la modélisation du processus de compréhension de discours.

Troisièmement, nous allons définir les structures informationnelles de discours et décrire leur interaction avec les structures situationnelles.

2 Construire une représentation discursive

Dans la perspective méthodologique adoptée dans notre travail [8], [9], [10], l'acte d'interpréter un discours consiste à construire sa représentation mentale qui comprend les réponses aux questions engendrées par six domaines: **informationnel**, autrement dit **thématico-rhématique**, **ontologique**, **fonctionnel**, **axiologique**, **énonciatif et du domaine métatextuel**, appelé aussi **domaine de conventions de genre**. Nous admettons qu'en essayant de comprendre un discours le lecteur crée sa représentation en puisant dans ses connaissances concernant les schémas textuels et discursifs, et aussi l'inventaire des modèles de situations stéréotypées. Ainsi, l'acte interpréter un discours implique un effort cognitif de la part du lecteur qui cherche à trouver les réponses aux questions portant sur :

1. le domaine thématique dans lequel le discours est situé, son thème global et les thèmes partiels qui sont dérivés directement du thème global,
2. la fonction discursive dominante,
3. le registre énonciatif,
4. le type d'univers construit dans le discours,
5. l'orientation axiologique de discours,
6. le degré de ressemblance entre les structures interprétées et les schémas textuels et discursifs déjà connus,
7. la situation modélisée dans le discours.

3 Structures situationnelles de discours

Les psycholinguistes P. Coirier, D. Gaonac'h et J-M. Passerault [2, p. 218], proposent de définir le **modèle de situation comme représentation mnémorique épisodique des événements, des actions, des personnages et de la situation décrite dans le discours**. Nous proposons de préciser la notion de **modèle de situation en le définissant en tant que cadre de l'expérience**. Nous avons ainsi repris [8], [9], [10], la notion introduite par E. Goffman dans l'ouvrage *Cadres de l'expérience* qu'il définit comme schéma interprétatif nécessaire pour identifier un événement quelconque [3, p. 30].

Dans l'ouvrage *Wykłady z gramatyki kognitywnej*, R. Langacker [5, p. 116] présente la notion d'**événement cognitif** conçu comme un schéma de conceptualisation de toutes les situations dans lesquelles peuvent se trouver les objets perçus, de toutes les relations qui peuvent exister entre eux et de tous leurs changements possibles. Il distingue ensuite **sept catégories d'événements cognitifs** : **existence**, **événement**, **action**, **sensation**, **possession**, **déplacement** et **transmission** [11, p. 116–125]. En combinant les approches de E. Goffman et R. Langacker, nous pouvons dire que dans chaque discours il est possible de retrouver au moins un cadre

de l'expérience. Autrement dit nous pouvons répondre à la question concernant la situation modélisée dans le discours, et indiquer la catégorie à laquelle ce cadre appartient. Il faut ajouter que les structures situationnelles de discours peuvent unir deux ou plusieurs types d'événements cognitifs p.ex: la description d'un arbre active le schéma d'existence quand l'auteur indique sa localisation ou énumère ses traits et le schéma de possession quand il procède à la décomposition de l'objet décrit.

L'événement cognitif du premier type — existence — conceptualise l'état et implique un seul rôle archétypique — celui de patient auquel on attribue un trait ou une localisation. On peut aussi indiquer la catégorie à laquelle il appartient ou se limiter à constater son existence. Les discours qui reflètent ce type de structures situationnelles ce sont par ex: les descriptions de lieux ou de produits (sauf — dans ce dernier cas- les parties de discours concernant leur composition).

Le schéma du second type — l'événement conceptualise le processus qui implique — dans le rôle de patients — la présence des choses ou être vivants. Nous pouvons retrouver les structures situationnelles fondées sur ce schéma dans les discours parlant p.ex.: de processus physiologiques ou phénomènes atmosphériques.

Le schéma d'action (R. Langacker l'appelle aussi événement cognitif canonique) relie par une chaîne d'actes au moins un agent et un patient. Il est toujours situé dans un lieu et temps, et implique une cause qu'on peut déterminer et une ou des conséquences prévisibles, comme dans les discours parlant des actes humains pilotés p.ex: manifestations sportives, délits criminels ou œuvres charitables.

L'événement cognitif appartenant au type suivant — celui de **sensation** concerne les actes de perception, les états intellectuels ou émotionnels. Ce schéma prévoit deux rôles archétypiques; de patient et de percepteur. Les discours décrivant les mécanismes de perception: visuelle, auditive, olfactive ou tactile basent sur ce schéma événementiel.

Le schéma de possession conceptualise plusieurs relations: entre un tout et sa/ses partie(s), entre la cause et l'état qu'elle entraîne, la relation de posséder une/des chose(s) concrète(s) et un/des objet(s) abstraits, et aussi la relation de parenté. Quand dans un discours nous retrouvons les descriptions des parties p.ex: d'un appareil, d'un bâtiment, d'une institution, ou d'une plante, nous identifions en même temps les structures situationnelles fondées sur le schéma de possession.

Le schéma de déplacement — plus complexe que les schémas précédents — active soit le schéma d'événement soit le schéma d'action situés dans la configuration source — piste — fin. Les discours qui reflètent ce type de structures situationnelles parlent p.ex: des mouvements conscients et voulus des agents humains ou, au sens abstrait, de l'acquisition du savoir conçue comme déplacement impliquant source, piste et fin.

Le schéma de transmission unit un de trois schémas: de possession, d'événement ou d'action au schéma de déplacement. Il implique quatre rôles: d'agent, de receveur, de patient et de fin. C'est le cas p.ex: des discours parlant de l'héritage génétique et du transfert des fonds.

4 Structures informationnelles de discours et leur lien avec les structures situationnelles

Nous définissons les structures informationnelles de discours en tant que structures thématico-rhématiques hiérarchisées. Ce modèle base sur la conception aristotélicienne de thème en tant qu'objet dont on parle dans une phrase, un paragraphe ou dans tout un discours. Le rhème est constitué par ce qu'on dit à propos de l'objet-thème.

Les thèmes phrastiques distingués grâce à l'application du test de négation — nous adoptons ici la proposition de A. Bogusławski [1] — constituent le premier niveau de la structure thématique. Le niveau supérieur est occupé par **les thèmes partiels**, autrement dit thèmes de groupes phrastiques. Le thème partiel — dérivé directement du thème global de discours, englobe les thèmes d'au moins deux ou plusieurs phrases dans le discours. Nous avons formulée cette idée de niveau intermédiaire dans la structure thématique de discours dans nos travaux antérieurs en développant successivement [6], [7], [8] et [9], le catalogue de procédures qui permettent d'identifier les thèmes partiels et de relations qui organisent la structure thématique au niveau supraphrastique.

La constitution du thème partiel peut se fonder sur plusieurs types de relations observées entre les thèmes de phrases dans le discours. Tout d'abord, cette décision peut s'appuyer sur la relation d'identité référentielle entre les expressions linguistiques qui renvoient au même objet-thème. Mais il arrive souvent que les thèmes de phrases et le thème partiel qui les recouvre sont liés par les relations beaucoup plus complexes. Le classement de relations sémantiques proposé par M. E. Winston, R. Chaffin et D. Herrman [12, p. 421] permet de préciser que le lien entre les thèmes phrastiques et le thème hiérarchiquement supérieur — thème partiel — peut consister en:

1. inclusion taxinomique — conçue en tant que relation entre une catégorie (représentée dans ce cas-là par le thème partiel) et son ou ses exemplaires (représentés par les thèmes de phrases),
2. inclusion mérologique qui relie un tout à ses parties,
3. inclusion topologique,
4. possession,
5. attribution.

La partie rhématique des structures informationnelles de discours reflète la structuration hiérarchique de thèmes. Ainsi, tous les rhèmes de phrases qui sont englobées par le même thème partiel constituent **l'ensemble rhématique** subordonné à ce thème. Chaque ensemble rhématique — donc un ensemble de rhèmes appartenant aux phrases qui dépendent du même thème partiel — peut être structuré par un des ordres suivants: spatial, temporel, axiologique, taxinomique ou mérologique, ou, éventuellement, par la combinaison de deux ou plusieurs de ces types d'organisation interne.

Nous allons nous concentrer maintenant sur les relations entre deux types de structures discursives: situationnelles et informationnelles. Nous croyons que ces

relations sont très importantes dans l'interprétation de discours en permettant au lecteur de reconnaître non seulement les relations entre les thèmes de phrases successives, mais aussi de formuler des hypothèses sur les structures informationnelles globales et de répondre ainsi aux questions concernant le thème de discours et sa segmentation en thèmes partiels recouvrant les groupes de phrases.

Dans cet article nous allons nous concentrer sur le niveau le plus haut de la structure informationnelle — celui de thème global. Il paraît qu'il faut prendre en considération au moins deux relations possibles entre le thème global et les structures situationnelles. Dans le premier cas **le thème global correspond à l'événement cognitif** — un de ses sept types — activé dans le discours. Mais il peut arriver que **le thème global est constitué par un un des éléments constitutifs de l'événement cognitif**: p.ex: un des rôles impliqués par le schéma événementiel, temps, lieu, cause, conséquence, source, piste, fin.

Les exemples qui suivent présentent ces deux possibilités. Dans le cas des exemples n° 1 et 2 les thèmes globaux — explicitement donnés dans les titres — correspondent à l'événement cognitif considéré comme un tout. Les textes n° 3 et 4, par contre, introduisent les thèmes globaux qui sont des éléments sélectionnés des structures événementielles de discours.

4.1 Le thème global correspondant à l'événement cognitif

Les structures situationnelles du texte n° 1 combinent deux types d'événements cognitifs. Les cinq premières phrases et la dernière (contient l'opinion à laquelle pourrait souscrire chaque œnophiliste) sont organisées par le schéma d'existence en décrivant un état — comparé à une maladie — dans lequel se trouve le patient — un œnophiliste. Les phrases n° 6, 7 et 8 reflètent, par contre, un autre schéma — celui de possession. Dans les phrases n° 6 et 7 on observe la relation mérologique entre un tout — l'Association nationale d'œnologie et ses parties représentées soit par ses membres (phrase n° 6), soit ses sections régionales (phrase n° 7). La phrase n° 8 introduit la relation entre les propriétaires — collectionneurs et les choses possédées — étiquettes de bouteilles de vin. Le thème global de discours — signalé dans le titre — œnologie — correspond donc au premier des événements cognitifs présents dans les structures situationnelles du discours analysé.

Texte n° 1

Œnologie — la petite soif du collectionneur

(1) L'œnophiliste ne dissimule pas sa maladie: au contraire, (2) il ... l'affiche!
(3) De foire en exposition, il se meurt pour la petite étiquette de bouteille de vin qui manque encore à sa collection.

(4) Condition sine qua non pour devenir un œnophiliste averti: aimer le vin.
(5) Sans pour autant perdre la faculté de prononcer le mot, un rien barbare! (6) Créée en février 1986 par deux Bordelais, vignobles oblige, l'Association nationale d'œnologie (A.N.O.) compte aujourd'hui quelque 500 membres, tous prêts à des démoniaques sacrifices pour dénicher la perle rare qui fait défaut à leur cave miniature. (7) Elle s'est récemment dotée d'un réseau de onze sections régionales,

calqué sur les principaux lieux viticoles. (8) Grâce à un système d'échanges, courant les foires au vin où se multiplient les concours de la plus belle étiquette, certains collectionneurs possèdent plus de 200 000 pièces, dont de véritables oeuvres d'art, souvent signées de grands noms. (9) „L'étiquette”, paraît-il, „est à la bouteille ce que le vêtement est à la femme”...

Dans le texte n° 2 les structures situationnelles basent sur deux schémas — de déplacement — les phrases n° 1 et 2 — et d'action, celui-ci étant réalisé par les dernières phrases de discours. Rappelons que le schéma de déplacement active soit le schéma d'événement soit le schéma d'action situés dans la configuration source — piste — fin. Dans le cas du texte analysé le patient — ce rôle est attribué à la chouette cheveche — est la victime d'une série d'événements qui peuvent mener à l'extinction de toute l'espèce. On peut donc reconstruire la source représentée par l'état initial où la chouette n'était encore en danger et la fin — qui correspondrait à sa disparition. Les dernières phrases reflètent le schéma d'action dans lequel les agents humains (“des spécialistes de l'animal aux yeux d'or”, “de nombreux agriculteurs”) unissent leurs efforts pour sauvegarder cet oiseau “en voie de disparition”. Il faut souligner que le thème global “le chant de la chouette” résulte de la substitution opérée sur un des éléments de l'expression figurée “chant du cygne” qui se réfère originellement, selon la définition de Lexis, à “la dernière œuvre d'un poète, d'un musicien, etc., d'un génie prêt de s'éteindre”. Le syntagme nominal “le chant de la chouette” renvoie à la situation dramatique des oiseaux en voie de disparition — leur sort est, selon l'auteur, similaire à celui des artistes évoqués par l'expression originale. Nos pouvons donc constater que ce thème global correspond au premier des événements fondant les structures situationnelles de discours — le déplacement.

Texte n° 2

Le chant de la chouette

(1) La chouette cheveche est en voie de disparition dans les campagnes normandes.

(2) L'homme et le modernisme sont les deux seuls responsables de la mort de ces oiseaux nocturnes. (4) Des spécialistes de l'animal aux yeux d'or ont décidé de protéger cet oiseau charmeur et proposent d'installer, dans des endroits propices, des nichoirs où la chouette pourra élever sa progéniture. (5) De nombreux agriculteurs ont répondu présents à cette initiative et (6) en autorisent la pose dans leurs exploitations. (7) Chouette alors!

4.2 Le thème global correspondant à un des éléments constitutifs de l'événement cognitif

Les structures situationnelles du texte n° 3 basent sur un seul schéma — celui d'existence qui est activé deux fois. Tout d'abord, les phrases de la 1^{ère} à la 5^e localisent le premier patient — dans ce rôle apparaît une aubépine, et énumèrent ses traits. Dans la dernière phrase le même schéma est actualisé étant appliqué, cette fois-ci, à un autre patient — la nature. Le thème global — l'arbre le plus vieux du

monde — correspond donc au patient du schéma événementiel d'existence dans sa première application.

Texte n° 3

L'arbre le plus vieux du monde

(1) Au village Saint-Mars-la-Futaie, dans la Mayenne, à l'ombre de l'église romane plantée en plein milieu de la commune, se dresse une aubépine. (2) Cette aubépine, de source sûre, multiséculaire, est pour les habitants bien plus âgée. (3) Il semblerait que son origine remonte au III^e siècle.

(4) Cet arbre représente sans aucun doute l'un des plus beaux trésors écologiques français. (5) Une imposante aubépine qui ne fleurit qu'en mai, et qui a la particularité d'année en année d'apparaître rose ou blanche. (6) La nature a ses secrets...

Dans le cas de l'exemple suivant, les structures situationnelles combinent trois schémas — celui de déplacement qui organise la première partie de discours (phrases de la 1^{ère} jusqu'à la 4^e), les schémas d'existence et de possession réalisés par les deux dernières phrases n°5 et 6. Le premier schéma permet de conceptualiser la série des actions des agents humains — les Lorrains — grâce à laquelle il était possible de passer de l'état initial — une situation démographique et économique désastreuse — à l'état final — développement de l'industrie viticole dans la région. La première partie de la phrase n°5 dans laquelle on attribue un trait au patient — “cette liste est incomplète” reflète le schéma d'existence, tandis que la suite de la phrase n°5 et toute la phrase n°6 activent le schéma de possession parce qu'on y énumère les parties de la liste en question. Il est donc clair que le thème global — la Lorraine viticole — correspond à un des éléments constitutifs du schéma de déplacement activé en tant que premier — c'est un lieu où se déroule toute cette série d'action.

Texte n° 4

La Lorraine viticole

(1) Après avoir été mise à mal par la désertification des campagnes et une épidémie de phylloxéra, la Meuse retrouve un véritable enthousiasme à développer son „industrie” viticole.

(2) De la plaine des Vosges aux rives de la Moselle, les Lorrains tentent de réaliser un rêve en redevenant une grande région de production aux spécificités locales, tout comme Alsace. (3) Mais le vœu se concrétise déjà avec l'obtention d'une appellation VDQS (vin délimité de qualité supérieure). (4) Cette renaissance nous permet de déguster des vins au caractère solide, plein de ressources, des blancs au goût d'amandes grillées recommandés pour accompagner les poissons et des pinots noirs corsés.

(5) Toutefois, cette liste est incomplète si l'on ne mentionne pas la „piquette” ou „piquette”, une boisson rafraîchissante obtenue en rinçant le moult avec de l'eau. (6) A ne pas oublier non plus le ratafia, sorte de vin cuit qui ne manque pas de personnalité.

4. Conclusion

Dans cet article nous avons présenté la conception des structures situationnelles fondée sur deux notions — la notion de cadre de l'expérience tirée des réflexions d'E. Goffman [3] sur les mécanismes de l'interprétation des événements de la vie quotidienne et celle d'événement cognitif formulée par R. Langacker [5]. Après avoir défini les structures situationnelles comme séquence de cadres de l'expérience — qui peuvent véhiculer un de sept types d'événements cognitifs distingués par R. Langacker — et les structures informationnelles comprises comme structures thématico-rhématiques hiérarchisées, nous avons réfléchi sur les relations entre ces deux types de structures discursives. Nous nous sommes concentrée sur le niveau le plus haut de la structure informationnelle — celui de thème global. Nous avons distingué deux relations entre le thème global et les structures situationnelles. Premièrement, le thème global peut correspondre à l'un des événements cognitifs activés dans le discours. C'est le cas des textes n° 1 et 2. Deuxièmement, le thème global peut être constitué par un des éléments de l'événement cognitif, p.ex: un des rôles impliqués par le schéma événementiel (patient, agent, receptrice, percepteur), temps, lieu, cause, conséquence, source, piste ou fin. Ce second cas est illustré par les textes n° 3 et 4. Evidemment la liste de relations possibles entre les structures situationnelles et informationnelles n'est pas finie. Il faudrait surtout réfléchir sur les cas où le thème global est plus général que les événements que nous pouvons identifier dans le discours et se réfère p.ex: à tout un domaine cognitif dans lequel un ou plusieurs schémas activés dans le discours sont situés.

Bibliographie

- [1] Bogusławski, A. (1977). *Problems of thematic-rhematic structure of sentences*. Państwowe Wydawnictwo Naukowe, Warszawa.
- [2] Coirier, P., Gaonac'h, D., and Passerault, J.-M. (1996). *Psycholinguistique textuelle. Approche cognitive de la compréhension et de la production des textes*. Armand Colin, Paris.
- [3] Goffman, E. (1991). *Les cadres de l'expérience*. Editions de Minuit, Paris.
- [4] Langacker, R. (1987). *Foundations of cognitive grammar. T. 1: Theoretical prerequisites*. Stanford University Press, Stanford.
- [5] Langacker, R. (1995). *Wykłady z gramatyki kognitywnej*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.
- [6] Miczka, E. (1993). Les structures supraphrastiques dans le texte. analyses et procédures. *Neophilologica*, 1993(9):41–60.
- [7] Miczka, E. (1996). Rola kategorii ponadzdaniowych w procesie rekonstrukcji tekstu. In Dobrzyńska, T., editor, *Tekst i jego odmiany*, pages 41–52, Warszawa. Instytut Badań Literackich PAN.
- [8] Miczka, E. (2000). Structures textuelles en tant qu'expression des catégories conceptuelles-organisateur d'expérience. *Neophilologica*, 2000(14):36–52.
- [9] Miczka, E. (2002). *Kognitywne struktury sytuacyjne i informacyjne w interpretacji dyskursu*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

- [10] Miczka, E. (2007). L'application des notions de cadre de l'expérience et d'événement cognitif à l'analyse de discours — cas du fait divers. *Neophilologica*, 2007(19):138–146.
- [11] Tabakowska, E. (2001). *Kognitywne podstawy języka i językoznawstwa*. Universitas, Kraków.
- [12] Winston, M. E., Chaffin, R., and Herrman D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 1987(11):417–444.

EWA GWIAZDECKA

Université de Varsovie, Warsaw, Poland

QUELLE DESCRIPTION POUR LE PRÉVERBE POLONAIS ?

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 and quasi-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 understood as an application (in the formal sense) of a predicate to its arguments. Moreover, 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. We argue that in the majority of the cases, verbal prefix encodes an achievement understood as a closure of the right boundary of the *process* associated 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, but according to the meaning of the verbal prefix, it can indicate “achievement of the beginning”, “achievement of the end phase”, etc.

Keywords: aspect; Polish verbal prefix; preposition; achievement; state; process; event; topology; semantics.

1 Le dictionnaire aspectuel des verbes

La description de l’aspect perfectif engendré par le préverbe polonais pose plusieurs problèmes. Le lien diachronique et sémantique avec la préposition fait que le préverbe, en plus des changements aspectuels, introduit très souvent des modifications de signification dans le verbe de base. Au surcroît, pratiquement chaque préverbe peut participer à la formation d’une *paire aspectuelle* et le choix en est déterminé lexicalement :

czytać — ***przeczytać***
pisać — ***napisać***
robić — ***zrobić***
siwieć — ***osiwieć***, etc.

Il arrive également que les verbes possèdent plusieurs correspondants aspectuels construits avec des préverbes divers : *zółknąć* — **po***zółknąć* — **z***zółknąć* 'jaunir'.

Tenant compte de ces difficultés, W. Cockiewicz et A. Matlak ont conçu *le dictionnaire structural et aspectuel du polonais* [1]. Le but de cet ouvrage est de présenter le système dérivationnel du verbe tout en montrant ses relations aspectuelles. Pour ce faire, Cockiewicz adopte une théorie de l'aspect [2] dans laquelle il oppose une action qui dure (imperfectivité) à une action terminée (perfectivité). Cette dernière est soit naturellement menée à son terme (perfectivité *au sens strict* : *przeczytać, napisać*), soit artificiellement interrompue (perfectivité *au sens large* : formations délimitatives engendrées régulièrement avec *po-* : *poczytać, popisać*).

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 de correspondant imperfectif, comme dans :

czytać → *przeczytać* → **przeczytywać*.

Dans ce cadre théorique, le réseau dérivationnel du verbe *pisać* 'écrire' se présente comme suit (la flèche montre la direction de dérivation et les deux-points indiquent les relations aspectuelles) :

pisać	→:	napisać		
	→:	popisać		
	→:	dopisać	→:	dopisywać
	→:	odpisać	→:	odpisywać
	→:	opisać	→:	opisywać
	→:	podpisać	→:	podpisywać
	→:	przepisać	→:	przepisywać
	→:	przypisać	→:	przypisywać
	→:	rozpisać/się	→:	rozpisywać
	→:	wpisać/się	→:	wpisywać
	→:	wypisać/się	→:	wypisywać
	→:	zapisać/się	→:	zapisywać
	→:	spisać	→:	spisywać

La première position du réseau est occupée (dans 75 % de cas) par un verbe imperfectif simple. Dans la deuxième colonne, nous trouvons les formations préfixées. Remarquons que, au sens de l'aspect adopté par les auteurs, seuls *napisać* et *popisać* sont considérés comme aspectuels. La troisième position est réservée aux verbes imperfectifs engendrés par un suffixe (ici : *-ywa-*). Les formations perfectives itératives 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 de la 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 ici utilisent les opérateurs topologiques (cf. *infra*) et font référence à la Grammaire Applicative et Cognitive [3] qui est une extension de la Grammaire Universelle de Shaumjan [4]. Cette grammaire s'articule sur trois niveaux de représentation :

- (i) le premier niveau, directement observable par le linguiste décrit les configurations morpho-syntaxiques d'une langue;
- (ii) le second niveau analyse des opérations logico-grammaticales en dégagant une structure applicative de la langue : (structure opérateur/opérande et non plus l'ordre syntagmatique);
- (iii) le troisième niveau représente les significations des unités lexicales en les décomposant en relations et opérations primitives.

Bien que chacun de ces niveaux possède une certaine autonomie, ils sont imbriqués par des procès de synthèse et d'analyse qui peuvent être décrits à l'aide de formalismes applicatifs (Grammaires Catégorielles et la logique combinatoire de Curry [5]).

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 [6] et dans la *théorie des lieux abstraits* développée par J.-P. Desclés et son équipe [7]. 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 [8] appelle la notion, plutôt que de privilégier la catégorie spatiale, nous allons parler d'un *lieu abstrait*. Par exemple, la préposition polonaise *do* + *gén* 'à', 'jusqu'à' marque la frontière d'un lieu dans :

- (a) *biegła do domu* (espace)
- (b) *czytał do wieczora* (temps)
- (c) *tańczyliśmy do upadłego* (notion)

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 :

- (d) *Anna biegła przez ulicę* → *Anna przebiegła ulicę*
'Anna courait à travers la rue' → 'Anna a traversé la rue (en courant)'
- (e) *Jan czekał przez (całą) noc* → *Jan przeczekał noc*
'Jan attendait jusqu'au matin' → 'Jan a attendu jusqu'au matin'
- (f) *Maria czytała książkę (przez pewien czas)* → *Maria przeczytała książkę*
'Maria lisait un livre (pendant un certain temps)' → 'Maria a lu le livre d'un bout à un autre'

Le préverbe possède donc une signification provenant de la préposition. Cet opérateur modifie souvent la signification du verbe, mais il marque aussi l'aspect perfectif. La question qui se pose concerne le choix du métalangage susceptible 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 [9], nous allons distinguer trois aspects : *état*, *événement* et *processus*. Nous utilisons les mêmes concepts, à des niveaux de représentation distincts, pour définir l'aspect lié aux marqueurs grammaticaux et l'aspectualité associée à la signification des prédicats lexicaux.

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) :



Fig. 1. Aspect ETAT

L'aspect *événement* représente une situation prise dans sa globalité. Il se réalise sur un intervalle fermé.



Fig. 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.



Fig. 3. Aspect PROCESSUS

Le processus qui a atteint sa 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 et au-delà de laquelle il pourrait encore continuer.

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 [10] ou Mourelatos [11] où les concepts aspectuels sont organisés hiérarchiquement.

Les trois aspects de base (**ASP**) s'appliquent à toute la relation prédicative (et non à un verbe seul) 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 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 :

PROC_I (DIT (**ASP_J** (Prédicat, Terme1, Terme2...)) I) & (**I REL J**)

avec : **J** — intervalle relatif à la relation prédicative aspectualisée,

I — intervalle relatif au processus d'énonciation.

Dans ce modèle, il serait possible de considérer le préverbe aspectuel comme un opérateur qui s'applique à un *processus* pour créer un *événement achevé*. Cependant, une telle description ne serait pas complète, car elle « écraserait » la dimension sémantique de la relation prédicative supprimant la distinction entre *napisać list*, *dopisać list*, *popisać list*, *przepisać list*, etc. En effet, pour expliquer certains phénomènes, nous aurons besoin de décomposer un événement grammatical engendré par le préverbe, c'est-à-dire de considérer les zones qualitatives liées au lexique.

En faisant 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*

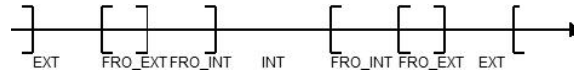


Fig. 4. Zones aspecto-temporelles

Si, à titre d'exemple, l'on considère le prédicat lexical *gnić* 'pourrir' et les formations préverbaux *nadgnić* 'commencer à pourrir', *zgnić* 'pourrir complètement', *przegnić* 'pourrir à travers'; *nadgnić* renvoie à la zone du début (frontière intérieure), *zgnić* marque la zone résultative et *przegnić* indique le dépassement de la frontière extérieure qui se situe vers la fin.

Remarquons que ces zones sont associées au *processus* en cours, dont les phases ne sont pas équivalentes entre elles. Chacune de ces phases (sauf l'intériorité associée à un état d'activité) peut être accomplie ou achevée en renvoyant à un événement qualitatif.

3 L'achèvement spatial

Comment, avec les outils présentés, modéliser les opérations du préverbe? Prenons quelques exemples de verbes de mouvement.

- (1) *Agata jechała*^{IMPF} **przez** *miasto*
 se déplacer à travers ville.ACC
 'Agata allait par de la ville...'
- (2) *Agata przejechała*^{PERF} (*przez*) *miasto*
 à travers-se déplacer (à travers) ville.ACC
 'Agata a traversé la ville'
- (3) *Agata jechała*^{IMPF} **do** *miasta*
 se déplacer jusqu'à ville.GEN
 'Agata allait jusqu'en ville'
- (4) *Agata dojechała*^{PERF} **do** *miasta*
 jusqu'à-se déplacer jusqu'à ville.GEN
 'Agata est arrivée en ville'

Dans (1) la préposition *przez* 'à travers' détermine le lieu spatial *miasto* 'ville' 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 'ville', 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 'ville' 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 'ville'. 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.

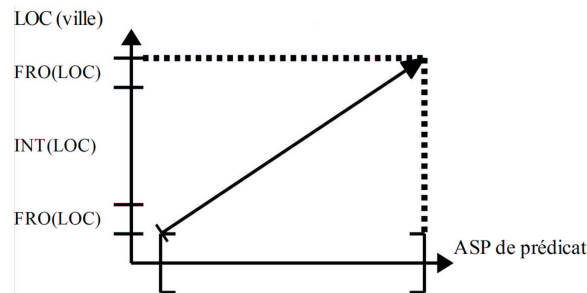


Fig. 5. Przejechać przez miasto

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 de l'objet direct [12]. 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'accomplissement¹.

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 :

¹ Comparons : (a) *Il a bu son verre de vin* et (b) *Il a bu du vin*. Dans (a) le processus est achevé, il ne peut plus se poursuivre, alors que dans (b) l'article partitif nous indique que l'action de boire pourrait continuer. Le processus est donc accompli.

- (5) *Agata zbudowała^{PERF} dom*
 prev-construire maison.ACC
 ‘Agata a construit la maison’
- (6) *Jan napisał^{PERF} list*
 prev-écrire lettre.ACC
 ‘Jan a écrit la lettre’
- (7) *Anna zjadła^{PERF} kolacj-ę*
 prev-manger dîner-ACC
 ‘Anna a mangé son dîner’

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, dans lequel le préverbe est un opérateur qui engendre la fermeture d’un processus sous-jacent et qui, en même temps, agit sur l’objet.

Les exemples traditionnellement considérés comme des modalités d’action, s’expliquent dans le cadre du même modèle :

- (8) *Anna dojadła^{PERF} kolacj-ę*
 jusqu’à-manger dîner-ACC
 ‘Anna a terminé son dîner’
- (9) *Agata odbudowała^{PERF} 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 *jeść*. En effet, *do-* marque la modification de l’objet dans une phase finale de sa “disparition”. Dans l’exemple suivant, nous analysons *od-* comme “une réponse à une situation donnée”. 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 précis de la *reconstruction* que le processus est achevé.

Avec les mêmes outils, nous pouvons décrire les formations délimitatives créées 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, etc.). 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.

5 Transition d'un état à un autre

Nous avons jusqu'à maintenant analysé les verbes qui marquent le changement agissant sur le lieu spatial ou sur l'objet. Prenons des exemples où les modifications concernent un élément T¹ (sujet) de la relation prédicative.

(10) *Jan wyłysiał*^{PERF}
prev-devenir chauve

'Jan est devenu chauve'

(11) *Anna schudła*^{PERF}
prev-maigrir

'Anna a maigri'

Traditionnellement (10) et (11) sont analysés comme résultatifs. En effet, dire *Jan wyłysiał* 'devenu chauve' implique l'état résultant *jest łysy* 'est chauve', mais nous pouvons inférer également un état antérieur *nie jest łysy* 'n'est pas chauve'. La transition d'un état à un autre est exprimée par un processus souvent lexicalisé (*łysieć, chudnąć, siwieć...*). Lorsque le préverbe s'applique à ce processus de transition, il ferme sa borne droite créant ainsi un événement et un état résultatif qui lui est concomitant. Cependant, il semble que le polonais se focalise non pas sur le résultat (celui-ci est seulement impliqué), mais sur le moment-même de l'achèvement. L'exemple suivant, tiré du corpus IPI PAN montre une succession d'événements :

Malowała powoli, portret robiła wprost latami. Model się zestarzał, wyłysiał, ożenił, schudł, musiał pozować, chciał czy nie chciał, chyba że umarł.

Nous pouvons analyser les formations inchoatives de la même manière :

(12) *Darek pokochał Mari-ę*
prev-aimer Maria-ACC
Darek est tombé amoureux de Maria'

(13) *Jan znienawidził ojc-a*
prev-haïr père-ACC
Jan a commencé à haïr son père'

En effet, *pokochać* 'tomber amoureux', *znienawidzić* 'commencer à haïr' *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 polonais apparaît comme le marqueur d'un achèvement. Est-il possible de faire émerger une représentation abstraite associée à

cette opération ? Un des buts de la Grammaire Applicative et Cognitive est de faire apparaître des schèmes grammaticaux et les invariants du langage. Pour ce faire, le linguiste doit reconstruire un ordre applicatif de l'énoncé, où les unités du langage se présentent comme des opérateurs et des opérands. Dans cette organisation qui est supposée sous-jacente à un ordre syntagmatique, un opérateur s'applique à un opérande pour former un résultat qui, par la suite, pourra fonctionner soit comme un opérateur soit comme un opérande. Les formalismes applicatifs du lambda-calcul et celui de la logique combinatoire permettent de composer les opérateurs entre eux.

Dans ce cadre applicatif, montrons les opérateurs relatifs à l'achèvement engendré par le préverbe polonais. En effet, cette opération met en jeu plusieurs opérateurs aspectuels qui se situent sur les niveaux différents de représentations. Donnons quelques éléments de l'analyse dans le contexte d'une relation prédicative binaire.

Considérons un prédicat P_2 comme un opérateur qui s'applique successivement à ses deux termes : $((P_2T^2)T^1)$. Nous allons introduire un opérateur aspectuel **ASP3** qui représente les propriétés aspectuelles inhérentes à ce 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° :

$$4. \text{PROC}\{\text{DIT}[\mathbf{ASP1} ((\mathbf{ASP2}(\mathbf{ASP3}P_2)T^2)T^1)]S^\circ\}$$

L'expression (4) 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, soit la trace de la composition formelle entre les opérateurs **ASP2** et **ASP3**. Il s'ensuit que l'aspectualité dans $P_2T^2T^1$ porte toujours sur un opérande, c'est-à-dire, l'argument T^2 . 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 dans le cadre de la logique combinatoire voir [13].

7 Conclusion

Nous avons présenté une analyse du préverbe polonais en tant que le marqueur d'un achèvement. Cette opération qui porte toujours sur un argument se présente comme "une complétude" temporelle et qualitative de l'action. En termes topologiques, il s'agit d'une fermeture du *processus* inhérent au prédicat lexical 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 établi plusieurs zones de l'achèvement en définissant cette opération plus largement que "la fin naturelle de l'action".

Il est clair qu'au stade actuel cette étude est loin d'être complète. Par exemple, les préverbes qui indiquent la quantité et la mesure (*na-*, *po-*) nécessitent une analyse à part. D'un autre côté, il semble que les compositionnalités sémantiques entre le préverbe et *się* 'se' (*przespacerować się* 'se faire une promenade', *przespać się* 'se faire un somme'...) et le préverbe *po-* et *sobie* (*poczytać sobie* 'se faire une lecture'), 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 polonais.

Bibliographie

- [1] Cockiewicz, W., Matlak A. (1995). *Strukturalny słownik aspektowy czasowników polskich*. Kraków : Uniwersytet Jagielloński.
- [2] Cockiewicz, W. (1992). *Aspekt na tle systemu słowotwórczego polskiego czasownika i jego funkcyjne odpowiedniki w języku niemieckim*. Kraków : Uniwersytet Jagielloński.
- [3] Desclés, J.-P. (1990). *Langages applicatifs, langues naturelles et cognition*. Paris : Hermès.
- [4] Shaumyan, S. K. (1977). *Applicative Grammar as Semantic theory of Natural Language*, Chicago : Chicago University Press.
- [5] Curry, H. B., Feys, R. (1958), *Combinatory logic*, vol. 1, North Holland, Amsterdam.
- [6] De Glas, W. (1991). Locological spaces: knowledge representation in an intensional setting. In *Proceedings of the third COGNITIVA symposium*, 229–337. Amsterdam : North-Holland Publishing Co.
- [7] Desclés, J.-P. (2006). Opérations métalinguistiques et traces linguistiques. *Colloque en Hommage à Antoine Culioli*. Centre International de Cerisy, septembre 2006.
- [8] Pottier B. (1995). *Sémantique générale*. Paris : Presses Universitaires de France.
- [9] Desclés, J.-P. (1980). Construction formelle de la catégorie grammaticale de l'aspect. In *La notion d'aspect*, David, J., Martin, R. (éds). Paris : Klincksieck. 198–237.
- [10] Vendler, Z. (1967). Verbs and Times. In *Linguistics in philosophy*. Ithaca : Cornell University Press, 97–121.
- [11] Mourelatos, A. (1978). Events, states and processes. *Linguistics in Philosophy* 2, 415–34.

- [12] Kiparsky, P. (1998), Partitive case and aspect. In *The projection of Arguments: Lexical and Compositional Factors*. Butt, M, Geuder W. (eds). Stanford: CSLI Publications, 275–269.
- [13] 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.