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

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

Metalanguage and Encoding Scheme Design for Digital Lexicography

Innovative Solutions for Lexical Entry Design in Slavic Lexicography

MONDILEX Third Open Workshop Bratislava, Slovakia, 15–16 April, 2009

Proceedings

Radovan Garabík (Ed.)

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

Metalanguage and Encoding Scheme Design for Digital Lexicography
Bratislava, Ľ. Štúr Institute of Linguistics, 2009.

The volume contains contributions presented at the Third open workshop "Metalanguage and encoding scheme design for digital lexicography", held in Bratislava, Slovakia, on 15–16 April 2009. The workshop is organized by the international project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources,* Capacities – Research Infrastructures (Design studies for research infrastructures in all S&T fields) EU FP7 programme.

Editor of the volume**: Radovan Garabík**

ISBN 978-80-7399-745-8

Contents

Foreword....7

Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian....9
Ivan Derzhanski, Natalia Kotsyba

Establishing Links between Natural Languages and the Universal Dictionary of Concepts....27
Viacheslav Dikonov

Lexical Database of the Experimental Bulgarian–Polish Online Dictionary....36
Ludmila Dimitrova, Rumyana Panova, Ralitsa Dutsova

Towards a Unification of the Classifiers in Dictionary Entries....48
Ludmila Dimitrova, Violetta Koseska-Toszewa, Joanna Satoła-Staśkowiak

MULTEXT-East Morphosyntactic Specifications: Towards Version 4....59
Tomaž Erjavec

Design of a New Slovak-Czech Lexical Database....71
Radovan Garabík, Jana Špirudová

Experience with Building Slovak Electronic Lexical Database....77
Ján Genči

Development of a Russian Tagged Corpus with Lexical and Functional Annotation....83
Igor Boguslavsky, Leonid Iomdin, Tatyana Frolova, Svetlana Timoshenko

Slovak Medical Terminology – Is a Worldwide Interoperability in Medicine Possible?....91
Oskár Kadlec

Russian Dictionary Base – First Steps....99
Karel Pala, Adam Rambousek, Maria Khokhlova, Victor Zakharov

Form, Its Meaning, and Dictionary Entries....105
Violetta Koseska-Toszewa

On the Meaning of Verbal Forms and Its Net Representation....112
Violetta Koseska, Antoni Mazurkiewicz

General Architecture and Lexical Entry Structure of the Polish-Ukrainian Electronic Dictionary....119
Natalia Kotsyba, Igor Shevchenko

To a Question about Semantic Syncretism in Old Russian Language and Its Reflection In Modelling Semantics of an Old Russian Word....133
Nekipelova Irina

Morphosyntactic Specifications for Polish. Theoretical Foundations. Description of Morphosyntactic Markers for Polish Mouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)....140
Roman Roszko

Theory of Lexicographic Systems. Part 1....151
Volodymyr A. Shyrokov

A Knowledge-rich Lexicon for Bulgarian....168
Kiril Simov

Non-Technical Computer Thesaurus versus Specialized Computer Thesaurus....177
Olena Siruk

Définition d'un prototype général de bases de données (étude des langues slaves de l'Ouest dans une visée multilingue)....183
Patrice Pognan

Foreword

This volume contains articles presented at the Third open workshop "Metalanguage and Encoding scheme design for digital lexicography" of the MONDILEX project. The workshop is organized by the international project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources*, Capacities – Research Infrastructures, developed under EU FP7 programme. The workshop, organized by Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, is held on 15–16 April 2009 in Bratislava, Slovakia.

The main purpose of this workshop is to study and outline innovative solutions for lexical entry design in Slavic lexicography and to present solutions for choosing and using a metalanguage in Slavic multilingual dictionaries and for designing an encoding scheme, studying how its design can best serve digital lexicography and natural language processing, as well as other related fields.

We hope the workshop results will be useful to lexicographers, computer linguists and linguists in general.

Ludmila Dimitrova, Radovan Garabík

Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian*

Ivan A Derzhanski[1] and Natalia Kotsyba[2]

[1] Institute for Mathematics and Informatics, Bulgarian Academy of Sciences
[2] Institute of Slavic Studies, Polish Academy of Sciences

Abstract. Comparative studies in theoretical linguistics and the production of bi- and multilingual dictionaries and tagged corpora, especially of closely related languages, can benefit from the use of a common, crosslinguistically consistent tagset which reflects the unity of grammatical categories to the greatest extent. As a case in point, the project MULTEXT-East developed tagsets for several Slavic languages and laid the foundations of the creation of a common Slavic tagset. Close scrutiny reveals, however, that it suffers from a number of inconsistencies and design flaws, which can have an adverse effect on its use in comparative work. In this paper we will suggest some amendments to MULTEXT-East v.3 (and v.4), and discuss what will have to be done in order for the remaining Slavic languages to be covered as well, with a focus on Polish, Ukrainian and Belarusian.

1 Introduction

Comparative studies in theoretical linguistics and the production of bi- and multilingual dictionaries and tagged corpora, particularly digital ones, can benefit from the use of a common, crosslinguistically consistent morphological tagset reflecting the structural, etymological and semantic unity of grammatical categories to the greatest extent. This is especially desirable in the case of closely related languages.

The project MULTEXT-East (MTE [3]) housed a classic endeavour to construct a foundation for creating tagsets for Eastern European languages (as well as one Western European language, namely English, which served as the hub language of the project). Version 3.0 covers 11 languages, with three more added in Version 4, to wit [4]:

- Indo-European:
  - Slavic:
    - East: (v. 4) RUSSIAN
    - West: CZECH, SLOVAK
    - South:
      - Western:
        - Slovenian: SLOVENE, RESIAN[1]
        - Serbo-Croat: CROAT, SERBIAN
      - Eastern: BULGARIAN, (v. 4) MACEDONIAN
  - non-Slavic: ENGLISH, ROUMANIAN, (v. 4) PERSIAN
- Uralic: ESTONIAN, HUNGARIAN

The seven Slavic tagsets in v.3 use 13 of the 14 parts of speech defined in the common tagset, with a total of 72 features and 263 values.

The project is generally acknowledged as having been very successful, and some of the MTE tagsets have become *de facto* standard for the respective languages. It is therefore a natural starting point for further work in this field.

* The study and preparation of these results have received partial funding from the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

[1] This is the Resian sub-dialect of the Slovene language of Bela/San Giorgio, Italy. Resian and standard Slovenian are mutually unintelligible due to archaisms preserved in Resian but not in contemporary Slovenian and to Italian-induced innovations in Resian grammar (including prepositive definite and indefinite articles).

Close scrutiny reveals, however, that the MTE system of tagsets for Slavic languages has a number of shortcomings which can have an adverse effect on its use in comparative work and its potential for extension to cover the remaining languages of the branch:

- On several occasions the same phenomenon in different languages is handled in different ways. For example, attributive participles are classified as verb forms in Bulgarian, but as adjectives in the other six Slavic languages in v.3, although there is no structural, semantic or etymological reason for such a discrepancy.[2] The four tagsets for Czech, Slovene, Russian and Bulgarian assume four different attitudes to the treatment of short and full forms of adjectives, where the actual semantic divergence might justify two.
- There are redundant values, such as 'transgressive' and 'gerund' (values of the feature VForm of the part of speech Verb), which refer to the same category, but the former is used in the tagsets for Czech and Slovak and the latter for Bulgarian and Serbian.
- Some terms are interpreted in unlike ways in different tagsets. Within the part of speech Numeral the type multipl[icativ]e is defined, but to the Czech tagset a multiple numeral is an adverbial one (*dvakrát* 'twice'), whereas to the Slovene tagset it is adjectival (*dvojen* 'double').
- Some solutions are not extensible. In Czech the 2nd person singular present tense form of the copula *jsi* can be cliticised as *-s* on certain non-finite verb forms and pronouns, and its presence is indicated by the positive value of the binary feature Clitic_s of the parts of speech Verb and Pronoun. Essentially the same phenomenon exists in Polish, but it involves four cliticised forms of the copula (1sg *-m*, 1pl *-śmy*, 2sg *-ś*, 2pl *-ście*), and they float more freely (the host can be any content word, e.g. *świniaś* 'thou art a pig', *dobryś* 'thou art good'), so the solution chosen in MTE for Czech can't be applied to Polish.

Excessively faithful adherence to grammatical tradition creates more awkwardness in the marking. This is especially conspicuous in the part of speech Pronoun. According to the traditional classification, personal and possessive pronouns are separate types, but reflexive pronouns are a single type. Thus in Czech *tobě* 'to thee' and *tvůj* 'thine' have different values of the feature Type (personal and possessive, respectively), whereas *sobě* 'to oneself' and *svůj* 'one's' are of the same Type (reflexive) and differentiated through the additional feature Referent_Type, although the relation is obviously the same in the two cases.

Some peculiarities can be explained by the need to keep the system compact because of the limitations of computing power a decade ago, a likely motivation for the designers to reuse the features as much as possible, even at the cost of linguistic adequacy. Now these concerns are no longer relevant.

In this paper we will examine MTE's treatment of the Slavic languages already covered and discuss what will have to be done in order for the rest of the branch, especially Polish, Ukrainian and Belorussian, to be treated as well.[3] In so doing we will focus on linguistic adequacy and crosslinguistic consistency, but will also aim for a concise tagset.

---

[2] Some of this is rooted in differences between national grammatical traditions. That they have often been followed is understandable, but comparative work requires a theoretical common ground, the lack of which defeats the purpose of a common tagset, so some traditional propositions will have to be sacrificed. (If the information is retained in whatever form, it will be a straightforward matter to convert it to the traditional form.) We are not aware of any post-MTE work aimed at bringing the various MTE tagsets closer to one another.

[3] We will not be concerned here with non-Slavic languages. Their coverage is particularly problematic, because so is the question of identifying matching grammatical categories when the languages aren't (closely) related. One of MTE v.3's most perplexing choices is that it uses the same binary feature Definiteness of the part of speech Verb to indicate, in Bulgarian, that a participle bears a definite article (*говорилите* 'the ones who talked'), and in Hungarian, that a finite form of a transitive verb has a definite 3rd person direct object (*tanulom* 'I learn it'). Thus two totally dissimilar (not to mention unrelated) phenomena are handled alike merely because their names in the respective grammatical traditions happen to mean the same. In MTE v.4 the tagset for Persian encodes izafet as Case=genitive (i.e., practically the opposite!) in an effort to avoid introducing a language-specific feature.

2 General remarks

The working definition that a word is a maximal uninterrupted sequence of letters stands in good stead most of the time, but there are several morphemes and clitics which form a graphic whole with their hosts in the standard orthographies (forms of the copula, the emphatic particle *-że* in Polish and *-ž* in Czech, prepositional markers of degrees of comparison), and some multi-word sequences might count as lexical units, but this technique should be used sparingly, and the matter relegated to syntax wherever possible.

2.1 Definiteness

Bulgarian has developed a synthetic definite article through the fusion of a form of a word belonging to one of the nominal parts of speech and a postpositive demonstrative pronoun. It is a peculiarity of the written norm that with singular masculine nouns ending in a consonant (as well as singular masculine forms of words of the other parts of speech) the article has two forms, full and short, originally stemming from different dialects but coexisting in the standard, being artificially assigned to different functions (according to the current norm, the full form is nominative and the short form oblique[4]).

The MTE tagset for Bulgarian maintains the feature Definiteness with the four values no (no article), yes (unique form of the definite article), full_art (full form of the definite article) and short_art (short form of the definite article). This makes it appear as though the distinction between the two forms of the article were on a par with its presence or absence. In fact these are features of different orders: the short and the full forms are varieties of the article, not its alternatives. We would propose two features, Article (no, yes) and DefForm (full, short).

Most Slavic languages (including Bulgarian) preserve the distinction between the full and the short form of the adjective, though typically only in a small part of the paradigm.[5] This can also be encoded through the feature DefForm (rather than Definiteness or Formation, as in MTE v.3 for the South Slavic languages and Czech respectively). The system would then look as follows:

Article	DefForm	Bulgarian (*як* m. ‘yak’)	Bulgarian (*яка* f. ‘collar’)	Bulgarian (*як* adj. ‘strong, sturdy’)	Ukrainian (*ярий* ‘violent’)
–	–				*ярий* (m.)
–	short				*яра* (f.); *яр*і (pl.)
–	full				*ярая* (f.); *ярії* (pl.)
no	–	*як; якове*	*яка; яки*	*яка* (f.)*; яки* (pl.)	
no	short			*як* (m.)	
no	full			*яки(й)* (m.)	

[4] Another norm existed during the rule of the Bulgarian Agrarian Popular Union (1921–23), when the choice of the full or short form of the article was based on euphonic rather than syntactic grounds (it depended on whether the following word began with a vowel or a consonant).

[5] In Serbo-Croat and Slovene the long forms are used as definite in all genders, numbers and cases, which justifies their encoding through a positive value of the feature Definiteness (or Article).

In Russian only the short nominative case forms are productive; they are used predicatively, as a general rule to express a temporary rather than permanent quality (*он весел* ‘he is in a cheerful mood’ vs *он весёлый* ‘he has a cheerful character’). However, short oblique case forms survive in numerous collocations (*среди бела дня* amidst white:GEN[SHORT] day:GEN ‘in broad daylight’). The situation is similar in Czech.

In Bulgarian only the masculine singular has a long form in *-и* (archaic *-ий*), used as a vocative (*драги съседе* ‘dear neighbour!’), appellative (*Петър Велики* ‘Peter the Great’), or (in archaic and poetic usage) definite (*равнините, набраздени с наший плуг* ‘the plains furrowed by our plough’). The MTE v.3 tagset for Bulgarian does not account for this form.

Ukrainian has lost the short masculine singular forms of all but 31 adjectives (an exhaustive list is given in [20]) and restricted the full feminine, neuter and plural forms to poetic speech.

yes	–	*яковете*	*якàта; якúте*	*якàта; якúте*	
yes	short	*якà*		*якúя*	
yes	full	*якъ́т*		*якúят*	

In Macedonian the norm supports three forms of the article distinguished by distance, and in MTE v.4 they are encoded as values of Definiteness (proximal, yes, distal). Strictly speaking, they call for a separate feature, Distance (proximate, neutral, distal), since the presence of any article should be opposed to indefiniteness, but DefForm and Distance can be unified for practical convenience.

Article	DefForm	Distance	Bulgarian (*як* m. ‘yak’)	Macedonian (*јак* m. ‘yak’)
no	–	–	*як*	*јак*
yes	short	–	*якà*	
yes	full	–	*якъ́т*	
yes	–	proximal		*јаков*
yes	–	neutral		*јакот*
yes	–	distal		*јакон*

2.2 Clitic_s

This feature is only defined for verbs and pronouns in Czech. As said before, it should be eliminated, because it is too specific, and can’t be extended to the parallel phenomenon in Polish.

3 Noun

3.1 Type

Currently gerunds (deverbal nouns) are encoded as common nouns. Since they are very frequent in Polish, it seems expedient to add a type for them, with the additional features Aspect and Negation relevant only to gerunds. The latter would enable *celebrowanie* ‘celebrating’ and *niecelebrowanie* ‘not celebrating’ to count as forms of the same lexeme [15:46].

3.2 Class

Noun class in Slavic is an interplay of gender and animacy. All Slavic languages have the same system of three genders (masculine, feminine and neuter). In addition, inflexion and agreement often draw a line between live beings and everything else or between human beings and everything else. In Polish and Sorbian both distinctions are relevant (the former in the singular and – in Sorbian – the dual, the latter in the plural); many accounts of Polish grammar handle them by distinguishing three masculine genders (human, animal and inanimate), but this leads to massive syncretism, because in fact the differences only affect a few forms each, and is not readily extensible to other languages (in Russian, for example, animacy is orthogonal to gender in the plural). It seems more advantageous to maintain three features: Gender (m, f, n), Human (yes, no) and Animate (yes, no).[6] Here is how the forms of the Polish cardinal numerals ‘1’ and ‘2’ in all genders and cases can be encoded. Note especially the rows where either Human or Animate is neutralised, but not both.

[6] The idea of encoding the Slavic generalised gender category through a combination of gender and animacy features was also expressed in [13–14], though stipulating a feature with further subdivisions (‘animacy’ includes ‘inhumanity’ and ‘humanity’ with two values). In our proposal there are a total of four values, including the contradictory combination of ‘human and inanimate’, but this is a low price to pay for the simplification of the general feature structure of the tagset, and it actually saves rules: in [9] it is shown that the entire paradigm of the Polish demonstrative pronoun *ten* ‘this’ can be described by 34 rules in a five-gender system, but in ours only 31 are needed.

Gender	Human	Animate	Case	Polish
m	–	–	n	*jeden*
m	no	no	a	
m	–	yes	a	*jednego*
mn	–	–	g	
mn	–	–	d	*jednemu*
mn	–	–	i, l	*jednym*
n	–	–	n, a	*jedno*
f	–	–	n	*jedna*
f	–	–	a, i	*jedną*
f	–	–	g, d, l	*jednej*
m	yes	yes	n	*dwaj*
m	yes	yes	n, a	*dwóch, dwu*
–	–	–	g, l	
–	–	–	d	*dwom, dwu*
–	–	–	i	*dwoma*
m	no	–	n, a	*dwa*
n	–	–	n, a	
f	–	–	n, a	*dwie*
f	–	–	i	*dwiema*

In Polish some masculine human nouns are formally demoted to non-human to express derogation (*te*/**ci pijaki* 'these:NONHUM/*HUM drunkards'); these can be encoded as masculine animal.[7] With other nouns of the same class occasional conversion to the wrong class is used to express a certain attitude. Some authors have suggested introducing Disparagement as a formal feature of the noun [7]. This is unworkable, however, because which form is neutral and which is disparaging depends on the lexeme, and agreement is with humanness, not with disparagement (cf. neutral *ci profesorowie* 'these professors', *te chłopaki* 'these lads', disparaging *te profesorzy*, *ci chłopacy*).

A common gender is also expedient for words that can be masculine as well as feminine whilst retaining the same inflexion (Bulgarian *родина* 'relative, kins[wo]man', Russian *сирота* 'orphan'). On the other hand, if a noun inflects in different ways (or not at all when feminine, as Polish *doktor* 'doctor'), this should be considered a pair of homonymous lemmata, with the homonymy resolved in the oblique cases.

3.3 Case

The original Slavic case system, preserved intact in most languages, contains seven cases (nominative, accusative, dative, genitive, instrumental, locative, vocative).

In Russian some nouns have two genitive or two locative forms with different meanings. Since these nouns are few, and the distinctions appear nowhere else in the grammar, introducing extra cases seems counterproductive. It is better to have an extra feature, CaseForm (first, second), whose value will select the correct subcase when needed, and be undefined most of the time.[8]

[7] When such a word is a subject, the predicate is masculine human (*Te pijaki przyszli* 'These:NONHUM drunkards came:HUMAN'). This is merely an instance of semantic agreement, which occurs in other Slavic languages also (Russian *Последний человек уволилась* 'The last:M person [= woman] resigned:F'), has an occasional character, and is outwith the scope of tagging.

[8] The proposed Russian tagset for MTE v.4 introduces the feature Case2 (p 'partitive', l 'locative'). This confines the choice to two possibilities with necessarily pre-defined cases, which is too restrictive, especially given that the locative in Ukrainian can even have three forms for the same word (*на водії*, *на водію*, *на водієві* 'on the driver'), cf. [19].

Case	CaseForm	Russian
n	–	*чай* ‘tea’, *молоко* ‘milk’, *снег* ‘snow’, *вода* ‘water’
g	–	*молока: цвет, чашка* ~ ‘the colour, a cup of milk’
g	first	*чая: цвет* ~ ‘the colour of tea’
g	second	*чаю: чашка* ~ ‘a cup of tea’
l	–	*воде: увидеть кольцо, красоту в* ~ ‘see beauty, a ring in the water’
l	first	*снеге: увидеть красоту в* ~ ‘see beauty in the snow’
l	second	*снегу: увидеть кольцо в* ~ ‘see a ring in the snow’

The same technique can be used for other instances of forms of the same case distinguished by usage, e.g.:

- the dative and locative singular of masculine nouns in Czech, which have the ending *-ovi* if the word is last in its phrase and *-u* otherwise (*bratrovi* ‘to the brother’, *bratru Janovi* ‘to Brother John’), and the similar alternation *-ові* ~ *-у* in Ukrainian, partly motivated by euphony (*панові Карпові Микитовичу Ковалеві* ‘to Mr Karp Mykytovych Kovalev’ [21:190]);
- the locative of monosyllabic Ukrainian nouns, where the ending *-у* tends to render a more specific meaning than *-і* (*муха в меді* ‘a fly is in the honey’, *зварено на меду* ‘cooked with honey’ [21:192]);
- the genitive of masculine nouns in Belarusian and Ukrainian, which has the ending *-а* for count nouns and *-у* for mass nouns, with some nouns assuming either depending on the interpretation (Bel. *пераезда* ‘of the [place for] crossing’, *пераезду* ‘of the [act of] crossing’; Ukr. *барви листопада* ‘the colours of leaf-fall’, *першого листопаду* ‘on the 1st of November’ [21:195]).[9]

This phenomenon is not to be confused with variability in the use of case, which is not restricted to the noun form, e.g., accusative in Ukrainian: *пасти (чорні) бики*$_{\text{ACC=NOM}}$, *пасти (чорних) биків*$_{\text{ACC=GEN}}$ ‘herd (black) bulls’ or *писати (довгий) лист*$_{\text{ACC=NOM}}$, *писати (довгого) листа*$_{\text{ACC=GEN}}$ ‘write a (long) letter’.

Russian, Slovak, Slovene and Lower Sorbian have lost the vocative case except for a few fossilised forms (*боже*, *bože* ‘god!’), which may be encoded as vocative forms of the nouns, as can Russian colloquial vocatives formed by truncation (*мам* ‘mum!’, *Вань* ‘Vanya!’). Categorising concordant adjectives etc. as vocative case forms (as *môj* in Slovak *môj bože* ‘my god!’), however, appears superfluous.

3.4 Additional features

All Slavic languages have pluralia tantum nouns (Bulgarian, Russian *клещи* ‘pliers’), consequently the tagset needs a way of marking this, as they have some syntactic peculiarities, such as cooccurrence with collective numerals (Russian *двое часов* ‘two clocks’ vs *два часа* ‘two hours’). It might be possible to do this by an additional value of the feature Gender, but for those languages that don’t collapse all genders in the plural, gender features (possibly reduced[10]) for pluralia tantum nouns are also essential (Serbian *макази* f. pl.t. ‘scissors’, *кљешта* n. pl.t. ‘pliers’; Slovene *anali* m. pl.t. ‘annals’, *gosli* f. pl.t. ‘fiddle’, *vrata* n. pl.t. ‘door’), which means that a separate feature will be needed.

As said earlier, the features Aspect (imperfective, perfective) and Negation (no, yes) should be added at least for Polish, where gerunds are especially frequent and *nie-* ‘non-’ is productively prefixed to them.

[9] In Belarusian this is actually an innovation, an effect of the incursion of the Russian genitive ending *-а* into the language in the second third of the 20th century and its rivalry with the originally ubiquitous *-у*, although the ensuing opposition of count and mass nouns is different from the distribution of the two genitives in Russian.
In present-day standard Ukrainian *першого листопаду* is considered incorrect ([18:53–54], [19]).

[10] Or conventional: e.g., in the IPI—PAS corpus of Polish pluralia tantum nouns that are not masculine human (and thus are fully ambiguous between masculine non-human, neuter and feminine) are labelled as neuter.

4 Verb

4.1 Verb form

Verb forms include the following:

- Original finite forms, typically inflecting within each tense only for person and (verbal) number, although Upper Sorbian also distinguishes gender in the dual, Slovene does likewise (although the feminine/neuter forms are considered obsolete), and Resian has a distinction of courtesy in the 2nd person plural.

  The following three tables display forms of the verb 'be'.

Person	Number	Gender	Human	Courtesy	Resian	Slovene	U Sorbian
1	dual	–	–	–	*swa*	*sva*	*smój*
1	dual	f, n	–	–		**sve*	
2, 3	dual	–	–	–	*sta*	*sta*	*stej*
2, 3	dual	m	yes	–			*staj*
2, 3	dual	f, n	–	–		**ste*	
2	plural	–	–	–		**ste*	*ście*
2	plural	–	–	no	*sta*		
2	plural	–	–	yes	*stë*		

- Erstwhile perfect participles that are only used predicatively and have effectively become finite past-tense indicative forms. They only inflect for number and gender.

Number	Gender	Russian
singular	m	*был*
singular	f	*была*
singular	n	*было*
plural	–	*были*

- Past participles (termed pseudoparticiples in [15]) used mostly as complements of an occasionally omitted copula in analytic forms of perfect tenses, the conditional mood or the passive voice, inflecting for (nominal) number (including collective in Resian) and nominal class. These are encoded as VForm=participle.

Number	Gender	Human	Animate	Resian	Czech	Polish	U Sorbian
singular	m	–	–	*bil*	*byl*	*był*	*był*
singular	f	–	–	*bila*	*byla*	*była*	*była*
singular	n	–	–	*bilu*	*bylo*	*było*	*było*
dual	–	–	–				*byłoj*
dual	m	–	–	*bila*			
dual	f, n	–	–	*bili*			
plural	–	–	–				*byli*
plural	m	–	–	*bili*			
plural	m	–	yes		*byli*		
plural	m	yes	–			*byli*	
plural	m	–	no		*byly*		
plural	m	no	–			*były*	*byłe*
plural	f	–	–	*bile*	*byly*	*były*	*byłe*
plural	n	–	–	*bile*	*byla*	*były*	*byłe*
collective	m	–	–	*bile*			

- Adverbial participles (gerunds as they are called in MTE's tagset for Bulgarian, or transgressives by the name used in the West Slavic tradition), uninflecting except in Czech, where they have retained number and gender: *nesa* (sg. m.), *nesouc* (sg. f./n.), *nesouce* (pl.) 'carrying'. These two values of the feature VForm should be unified; we would propose the label 'r' (because the part of speech Adverb is marked 'R').
- An invariable impersonal, originally an adverbial form of the past passive participle (in Polish, Ukrainian and Belorussian). For this we would propose the label 't', reminiscent of one of the suffixes.
- Finite forms of moods other than the indicative.
- Infinitive, invariable.[11]
- Supine, ditto (only in Slovenian, Resian and Lower Sorbian, though formerly in Czech as well).

Attributive participles, inflecting for number, gender and case or definiteness, are considered adjectives in several but not all tagsets in MTE. We believe this is right, and should be followed for all languages. The assumption that fully inflected participles are verb forms entails that the entire paradigm of the adjective is a proper part of the paradigm of the verb. This runs afoul of the proposition that the adjective and the verb are entities of the same order (parts of speech). Intuitively, too, Russian *читающего* 'reading:SG.M.GEN' is a form of the lemma *читающий* 'reading (present participle)', not of the lemma *читать* 'read'. And the argument (of a syntactic nature) that clause-forming participles have verbal government should not be considered relevant to morphological analysis.[12]

The tagset for Resian includes a subjunctive, but this category contains merely the 2nd person imperative forms, which are used as a subjunctive mood for all persons.

The tagsets for the other languages except Bulgarian include a conditional marker, inflecting for person and number in Czech and Serbo-Croat as in Polish and Upper Sorbian, uninflecting in Slovak, Slovene, Macedonian and Russian as in Ukrainian, Belarusian and Lower Sorbian.[13]

The IPI—PAS corpus of Polish (IPIC [7]) introduces a separate subcategory within the part of speech Verb for the so-called agglutinants, i.e., bound cliticised forms of the copula. The form *-s* of Czech *jsi* (2nd \ person singular form of the copula) calls for the same treatment.

VForm	Tense	Person	Number	Polish	Czech
indicative	present	1	singular	*jestem*	*jsem*
indicative	present	2	singular	*jesteś*	*jsi*
indicative	present	1	plural	*jesteśmy*	*jsme*
indicative	present	2	plural	*jesteście*	*jste*
bound	–	1	singular	*-m*	*-ch*
bound	–	2	singular	*-ś*	*-s*
bound	–	1	plural	*-śmy*	*-chom*
bound	–	2	plural	*-ście*	*-ste*

4.2 Aspect

Aspect is a category common to all Slavic languages, although not reflected in all tagsets in MTE. It would be desirable for the aspect called progressive to regain its usual name, imperfective. An ambivalent aspect might be more widely recognised (biaspectual verbs are numerous in Bulgarian, for example).

[11]The Bulgarian (truncated) infinitive has recently become obsolete, but can occur in texts: *недей казва* 'don't say', *можете ли каза* 'can you say' (now more commonly *недей да казваш*, *можете ли да кажете*).

[12]Neither is it consistently appealed to: Czech and Slovak attributive participles are clause-forming, but are encoded in MTE as qualificative adjectives; Bulgarian or Russian participles are no different.

[13]The Bulgarian conditional *бих*, *би* etc. are encoded in MTE as aorist tense forms of the verb *бъда* – a perfective counterpart of the imperfective copula *съм* –, although the forms *бидох*, *биде* etc. are better candidates for such encoding; in the contemporary language *бих*, *би* have no perceivable relation to the aorist.

4.3 Tense

MTE v.3 supports present, future, past, aorist, imperfect and pluperfect. The undifferentiated past tense is based on participles in the East Slavic languages or on the collapse of the aorist of perfective verbs and the imperfect of imperfective verbs into a single so-called preterite tense in Sorbian (a pronounced tendency in Macedonian as well).

Aspect	Tense	Person	Number	Gender	Bulgarian	Russian	U Sorbian
imperfective	imperfect	2, 3	singular	–	*ядеше*		
imperfective	past	2, 3	singular	–			*jědźeše*
imperfective	past	–	singular	masculine		*ел*	
imperfective	aorist	2, 3	singular	–	*яде*		
perfective	imperfect	2, 3	singular	–	*изядеше*		
perfective	past	–	singular	masculine		*съел*	
perfective	past	2, 3	singular	–			*zjě*
perfective	aorist	2, 3	singular	–	*изяде*		

The pluperfect is only introduced in the tagsets for Croat and Serbian, for no evident reason, as no Slavic language has a synthetic pluperfect.

4.4 Other features

Many (though not all) Russian verbs have a 1st person plural inclusive, formally present tense, form with hortative semantics: *идёмте* (imperfective), *пойдёмте* (perfective) 'let us (you:PL and I) go'. This could be encoded as a 1st person plural form of a special mood (verb form, e.g. 2nd imperative, as in the National Corpus of the Russian Language); however, structurally it is not the mood but the person (a combination of *-м* '1st pl.' and *-те* '2nd pl.') that makes it exceptional. Such a form should either have a special value (inclusive) of the feature Person or be treated as an agglutinative compound of a 1st person plural verb form and the bound particle *-те* (also found in *вот* 'here you are!', *ну* 'well!' with an addressee for whom the 2nd person plural is used).

For Polish the feature Vocalicity (voc, nvoc) has been added in IPIC to separate the cliticised forms of the copula with a buffering vowel (*-em*, *-eś*) or without one (*-m*, *-ś*).

IPIC also introduces the feature Agglutinativity (agl, nagl) for accounting for some problems of wordhood [15].[14] It has a positive value for past tense forms of verbs (pseudoparticiples) that require a bound clitic (*gniotł-em* 'I kneaded') and a negative one for their self-sufficient counterparts (*gniótł* 'he kneaded'). The same technique might be used for Czech singular imperatives which have a bound form before the particle *-ž* (*buď* 'be!', but *budi-ž* 'be thou now').

5 Adjective

5.1 Type

MTE v.3 recognises adjectives of three types: qualificative, possessive and ordinal (actually relative, a mistranslation of the Slovenian term *vrstni*). All attributive participles in all languages except Bulgarian are categorised as qualificative adjectives, ignoring voice and tense. However, it would be desirable to preserve this information by introducing a new type of adjective, participle, and voice, tense and aspect as features relevant only to participles. The table below displays the Bulgarian adjective *дъвчащ* 'chewing (of sweets)' as well as all participles formed from the verb *дъвча* 'chew':

[14] In the formalism used in the IPIC tagset [7] binary features typically have values of the type (‹value›, n‹value›); in MTE's notation these can always be rendered as (yes, no).

PoS	Type	Aspect	Tense	Voice	Bulgarian
Adjective	qualificative	–	–	–	*дъвчащ*
Adjective	participle	imperfective	present	active	*дъвчещ*
Adjective	participle	imperfective	aorist	active	*дъвкал*
Adjective	participle	imperfective	aorist	passive	*дъвкан*
PoS	VForm	Aspect	Tense	Voice	Bulgarian
Verb	participle	imperfective	imperfect	active	*дъвчел*

Furthermore, since exclusively predicative adjectives (e.g., Slovak *dlžen* ‘obliged’) are treated as regular adjectives, predicative participles (including such as are used as past tense forms of verbs, alone or with conjugated forms of a copula) should be too.

It would be advantageous to also move ordinal (and other adjective-like) numerals and some types of pronouns to the part of speech Adjective, again distinguishing them by type, so as to relieve the other parts of speech of the strictly adjectival features.[15]

Type	Czech
qualificative	*dobrý* ‘good’
possessive	*matčin* ‘mother’s’
ordinal numeral	*pátý* ‘fifth’
specific numeral	*dvojí* ‘double, twofold’

IPIC distinguishes two further types of adjectives: preadjectival (the first halves of compounds such as *biało-czerwony* ‘white-and-red’) and postprepositional (the content words in expressions of the type *po polsku* ‘in Polish’, only used following the preposition *po*). The former is advisable since it would be impractical to provide all compounds in the dictionary; the latter are better classified as adverbs.

5.2 Degree

Degree (positive, comparative and superlative[16]) is defined for all Slavic languages except Bulgarian, where it has been decreed that the degree markers *по-* (comparative) and *най-* (superlative), both linked to the adjective or adverb by a hyphen in the current orthography, might better be treated as separate words (Particles of type comparative). While fully functional, this decision separates the Bulgarian superlative *най-* from its counterparts in the other languages (*nej-* in Czech, *naj-* elsewhere, all prefixed to the comparative form and written as one word); then again, this may be justified by the fact that in Bulgarian both degree markers can also be used with other parts of speech and expressions, although then separated by a space in writing (*по юнак* ‘more of a hero’, *най ми е жал* ‘I regret most’). In Macedonian the same markers are written as a solid word together with the adjective or adverb (*подолг* ‘longer’, *најмногу* ‘most’), and MTE v.4 treats the whole as a form inflected for degree.

In the Ukrainian Grammatical Dictionary [20], the source of morphological information for Ukrainian, degree was disposed of, comparative and superlative adjectives and adverbs are recorded as separate lexemes with corresponding lemmata. Rules for extracting information on degree and redirecting non-positive units to their lemma were designed and implemented in the project UGTag [6], enabling information on degree to be encoded for Ukrainian.

[15] Some national traditions actually call for this: ‘Numerals in Slovene can function as nouns, adjectives or adverbs, and are in grammars described as subtypes of these categories. The above classification runs counter to the established practice and is missing an important syntactic distinction’ [4:205].

[16] Also elative for Slovene, Resian and Serbian and diminutive for Resian, though no examples are provided.

5.3 Additional features

The feature Negation (no, yes) should be added at least for Polish with its regularly formed participles.
For Sorbian the feature Owner_Gender would have to be borrowed from the part of speech Pronoun, to encode the gender of the noun from which a possessive adjective is derived, as such a noun can have concordant modifiers (Upper Sorbian *stareje žoniny syn* 'the old woman's son', Lower Sorbian *našogo nanowe crjeje* 'our father's shoes' [8]).

PoS	Type	Owner_Gender	Gender	Number	Case	Upper Sorbian
Adjective	qualificative	–	feminine	singular	genitive	*stareje*
Adjective	possessive	feminine	masculine	singular	nominative	*žoniny*
Noun	common	–	masculine	singular	nominative	*syn*

6 Pronoun

6.1 Type

Traditional Slavic grammars acknowledge nine types of pronouns (personal, possessive, reflexive, demonstrative, interrogative, relative, indefinite, negative and general). The system is partly inconsistent: some pairs of pronouns of the same type (both reflexive, interrogative, etc.) stand in the same relation with one another as a personal and a possessive pronoun, and many pronouns fit the criteria for membership in more than one class (Ukrainian *свій* 'one's [own]' could be classified as both reflexive and possessive, *хтозна-чий* 'who knows whose' as indefinite and possessive, *хтозна-який* 'heaven knows what kind of' as indefinite and demonstrative, etc.).

It appears that personal and possessive pronouns can be conflated (because there have to be other means for handling this kind of opposition anyway, as between 'who' and 'whose'), and reflexive pronouns can be unified with them (as a special value of Person[17]).

MTE v.3				Our proposal		
Type	Person	Referent_type	Czech	Type	Person	Referent_type
p	2	–	*tobě*	p	2	p
s	2	–	*tvůj*	p	2	s
x	–	p	*sobě*	p	x	p
x	–	s	*svůj*	p	x	s
q	–	(p)	*kdo*	q	–	p
q	–	(s)	*čí*	q	–	s

In general these features refer to the meaning of pronouns and should be dealt with at the level of semantics. The developers of UGD [20] divide traditional pronouns into pro-nouns and pro-adjectives (pro-adverbs, too, in Russian National Corpus project); the designers of IPIC [7] refer to pro-adjectives as ordinary adjectives, while pro-nouns are singled out as a class. We would favour encoding pro-adjectives as several types of adjectives and preserving pro-nouns as a separate class.

6.2 Referent_Type and Syntactic_Type

These two features appear redundant, as a personal (possessive) value of Referent_Type correlates with a nominal (adjectival) value of Syntactic_Type.

The Bulgarian tagset doesn't use Syntactic_Type at all, but employs two unique values of Referent_Type: attributive and quantitative. The first of these allows distinguishing, e.g., attributive *какъв* 'what kind of' from possessive *чий* 'whose'. The words categorised as quantitative pronouns (*колко* 'how many/much', *няколко* 'several', *толкова* 'this many/much') correspond to numerals distinguished by

[17]This would not work, obviously, if English with its person-marked reflexives were restored to the system.

values of the feature Class (interrogative, indefinite, demonstrative) in Czech and Slovak, and the Slovene and Resian tagsets don't identify them in any way. The choice seems to be a matter of economy. Handling these words as pronouns takes advantage of the numerous types of pronouns already defined, and treating them as numerals facilitates their classification by type of numeral (e.g., Czech cardinal *kolik* 'how many', ordinal *kolikátý* 'number what', multiplicative *kolikrát* 'how many times'; Bulgarian has fewer such types, but it needs a way of distinguishing *колцина* 'how many [people]' from *колко* 'how many/much', although MTE v.3 provides none).

6.3 Additional features

In all East and West Slavic languages personal pronouns of the 3rd person have forms starting with /n/ instead of /j/, typically employed when the pronouns are objects of prepositions. For this phenomenon IPIC uses the feature Postprepositionality (praep, npraep), a practice which should be emulated. Also, in Upper Sorbian the pronoun *što* 'what?' has the same form in the accusative except after a preposition, where *čo* substitutes; this can be encoded in the same way.

Type	Gender	Human	Number	Case	Postprep	Upper Sorbian
personal	masculine	no	singular	accusative	no	*jón*
					yes	*njón*
interrogative	neuter	no	singular	accusative	no	*što*
					yes	*čo*

It should be noted, however, that the condition of the use of these forms vary somewhat across languages: in Russian they are optionally used after comparative degree forms (*ниже них ~ ниже их* 'below them, lower than they'), in Ukrainian the conditions depend on the dialect. For this reason it may be advisable to give the feature a less binding name (one motivated by the form rather than the function).

7 Numeral

7.1 Type and Form

All languages distinguish cardinal and ordinal numerals; also, in MTE v.3 collect[ive]s are introduced for Serbian, and multipl[icativ]es and special[18] numerals for all seven languages except Resian and Bulgarian. On the whole the systems of numerals are made to look more different than most of them actually are.

The Bulgarian masculine personal numerals are handled as Type=cardinal Form=m_form in MTE v.3. In a common tagset this language-specific value would be superfluous, thanks to the feature Human.

Gender	Human	Bulgarian	
m	yes	*двама*	'2'
m	no	*два*	
fn	–	*две*	

7.2 Class

For Polish the feature Accomodability (congr 'agreeing', rec 'governing') has been added in IPIC to identify the structural relation between the cardinal numeral and the noun (attribute–head or head–complement, respectively): *Przyszli dwaj chłopcy* 'Two:CONGR boys:PL.NOM came:PL.HUM', *Przyszło dwóch/ dwu chłopców* 'Two:REC boys:PL.GEN came:SG.N'. This can be encoded here through the feature Class, introduced in MTE v.3 in order to account for the different syntactic distribution of the cardinal numerals (esp. in Czech):

[18] Or specific, denoting a number of kinds of substances.

Gender	Human	Class	Polish	
m	yes	definite	*dwóch*, *dwu*	‘2’
		definite2	*dwaj*	
m	no	definite2	*dwa*	
n	–	definite2		
f	–	definite2	*dwie*	
m	yes	definite	*trzech*	‘3’
		definite34	*trzej*	
m	no	definite34	*trzy*	
f, n	–	definite34		
m	yes	definite	*pięciu*	‘5’
m	no	definite	*pięć*	
f, n	–	definite		

8 Adposition

8.1 Type

Slavic languages tend to only have prepositions. In Russian a few prepositions (*вопреки* ‘contrary to, notwithstanding’, *назло* ‘to spite’, *ради* ‘for the sake of’, *спустя* ‘after, later’) can be used postpositively; Sorbian *dla* ‘because of’ is more often a postposition than a preposition (Upper Sorbian *špatneho wjedra dla ~ dla špatneho wjedra* ‘because of the bad weather’; Lower Sorbian *chórosći dla ~ dla chórosći* ‘due to illness’, cf. German *krankheitshalber*). These should be undefined as to Type.

8.2 Case

In linguistic theory an adposition’s subcategorisation of an object in a certain case is no different from the subcategorisation of a verb. Tagsets don’t usually encode transitivity features for verbs, so introducing such a feature for prepositions amounts to an inconsistency. In practice, too, since in Slavic languages many prepositions can govern more than one case, the case syncretism common in nouns entails massive ambiguity in the tagging of prepositions.

We contend that no such feature ought to have been introduced into the morphological tagset. We would keep it only for the reason that its use is a widespread practice.

8.3 Additional features

Typically the object of a preposition, if a pronoun, must be a full (stressed) form. But there are exceptions. In Bulgarian the object of a few prepositions can be expressed as a dative (possessive) clitic[19] as well as a full accusative form (*помежду им* or *помежду тях* ‘between them’, but only *между тях* dto.). In Upper Sorbian the 1st person singular pronoun appears as a clitic after polysyllabic prepositions (*přećiwo mi* ‘against me’, *pola mje* ‘by me’, but *ku mni* ‘towards me’, *za mnje* ‘for me’). These peculiarities of the prepositions can be encoded by an additional feature.

It would be advisable to borrow the binary feature Vocalicity from the part of speech Verb for extended forms of prepositions (Bulgarian *във ~ в* ‘in’, Russian *передо ~ перед* ‘before’, Polish *ku ~ k* ‘towards’, Upper Sorbian *wote ~ wot* ‘from’, etc.), used in specific (morpho)phonological conditions.

[19]The MTE tagset for Bulgarian marks the short dative forms of the pronouns (*ми* ‘to me’, …, *им* ‘to them’) doubly as Type=personal Case=dative and Type=possessive, which is in conformity with the traditional descriptions, but redundant (especially since the use of a dative clitic as an adnominal possessive marker in Bulgarian is not an accident, but an areal feature shared with other languages of the Balkans).

In several languages adpositions optionally merge with some pronouns, yielding such compounds as Czech *zaň ~ za něho* ‘for him’, *proč ~ pro co* ‘for what’, Slovene *zate ~ za tebe* ‘for thee’, Polish *przezeń ~ przez niego* ‘because of him’, Upper Sorbian *mojedla ~ dla mnje* ‘because of me’, Lower Sorbian *mójogodla ~ dla mnjo* dto. (cf. German *meinetwegen*). It is best to treat these as agglutinative compounds, so as not to lose information about either the adposition or the pronoun.

9 Conjunction

Forms such as Czech *abych* ‘that I would’, *kdybyste* ‘if you would’ might also be treated as compounds (following the path suggested by their Polish counterparts *abym*, *gdybyście*) rather than as conjunctions inflected for person and number as in the MTE v.3 Czech tagset. (Conjunctions are, after all, supposed to be an invariable part of speech.) This would make for greater consistency across languages.

10 Predicative

Uninflecting words (and some collocations) which are restricted to being complements of copulative verbs are recognised as a separate part of speech in several reference grammars and tagsets of various Slavic languages. This appears superfluous: as we argued in [2], such items are adverbs no less than predicative adjectives (English *glad*, Russian *paд* dto.) are adjectives. However, attributivity/predicativity may be introduced as an additional feature for the purposes of syntactic analysis.

11 Conversion of existing formats for Polish and Ukrainian to an MTE-like format

Resources for morphological processing of Polish and Ukrainian have been developed independently from the project MTE in Poland and Ukraine, respectively. Morphological information is encoded in the form of grammatical dictionaries that allow for both analysing and synthesising word forms. The granulation of grammatical information there and the formats of recording it differ considerably from the core MTE tagset. Grammatical categories and values overlap (are one-to-one relations) only in part; some of them have to be decomposed into finer ones, and new categories/values need to be assigned to all relevant lexemes in a grammatical dictionary. On the other hand, grammatical dictionaries contain information that is not necessary for MTE-like tagging. There are two possible levels of introducing changes into Polish and Ukrainian grammatical sources. This can be done at the level of conversion of tagged texts, or directly in the dictionary source files. The former option is chosen for Polish, since the source files are not available for processing and development. The latter option has been chosen for Ukrainian, and additional grouping of lexemes is done within UGTag [6], which foresees the creation of a morphological tagger for Ukrainian with the possibility of adding new words from tagged texts, unrecognised by the tagger. One possible output format of UGTag will be an MTE-like tagged text.

As for Belarusian, a grammatical dictionary for it is under development now on the basis of an extensive orthographic dictionary [11], and suggestions concerning its design and compatibility with MTE-like tagging format can be taken into account, so that no further conversion will be required.

The tagsets for Polish (IPIC) and Ukrainian (UGD) were brought together within the PolUKR project with the aim of creating a common tagset for the parallel corpus of those languages [5]. The criterion of minimal information loss was used, although the common tagset is not a pure arithmetic sum of the two tagsets; rather, it was based on the pattern of IPIC, as it was easier this way to adjust the search program Poliqarp for the needs of PolUKR. Since MTE-like tagging is becoming a standard now, it was decided to bring the PolUKR tagset to conformity with it.

Here is a fragment of the conversion table IPIC/PolUKR → MTE v.3/4 (111 dictionary positions):

Ukrainian term	Polish term	English term	PolUKR tag	MTE tag (fragment)	example
частка-вигук	partykuło-przysłówek	particle-adverb	qub	Q	*niech*
вставні слова	dyskursyw	discourse markers	dsc	Q	*властиво*
інфінітив	bezokolicznik	infinitive	inf	V, VForm=n	*спатоньки*
безособова форма	forma -no/-to	impersonal form	imps	V, VForm=t	*rozpoczęto, robiono*
дієприслівник	imiesłów przysłówkowy	adverbial participle	part	V, VForm=r	
недоконаний дієприслівник	imiesłów przysłówkowy współczesny	simultaneous adverbial participle	pcon	V, VForm=r, Tense=p	*роблячи, robiąc*
доконаний дієприслівник	imiesłów przysłówkowy uprzedni	anterior adverbial participle	pant	V, VForm=r, Tense=a, Aspect=e	*зробивши, zrobiwszy*
дієприслівник минулого часу	imiesłów czasu przeszłego	simultaneous past participle	ppast	V, VForm=r, Tense=a, Aspect=p	*робивши, *robiwszy* (rare)
загальний	ogólny	common (general) noun	gnoun	N, Type=c	*шахи*
власна назва	nazwa własna	proper name	propnoun	N, Type=p	*Сколе*
пейоративний іменник	rzeczownik deprecjatywny	disparaging (depreciative) noun	depr	N, Animate=y, Human=n	*profesorzy*
займенник-іменник 1-2 особа	zaimek 1-2 osoba	1st- or 2nd-person pronoun	ppron12	P, Type=p, Person=(1|2)	*я, ти*
герундій	gerundium	gerund	ger	N, Type=g	*robienie, nierobienie niezrobienie*
займенник-іменник 3 особа	zaimek 3 osoba	3rd-person pronoun	ppron3	P, Type=p, Person=3	*він, вони*
займенник себе	zaimek siebie	pronoun ‘self’	siebie	P, Type=x	*себе*

And a fragment of the correspondence table MTE v.3/4 → IPIC/PolUKR (332 positions):

category	attribute	value code	value name	IPIC/PolUKR equivalent
Adjective(A)	Aspect	e	perfective	(pact|pass)&aspect=perfective
Adjective(A)	Aspect	p	progressive	(pact|pass)&aspect=imperfective
Adjective(A)	Voice	a	active	pact&aspect=perfective
Adjective(A)	Voice	p	passive	pass&aspect=perfective
Adverb (R)		R		adv|adjp|pred
Verb(V)	VForm	i	indicative	fin|praet|bedzie
Verb(V)	Tense	p	present	fin&aspect=imperf
Verb(V)	Tense	f	future	bedzie|(fin&aspect=perf)

Two sets of XML morphosyntactic specification files for Polish and Ukrainian have been prepared: specifications compatible with the most recent, still unreleased version of MTE (v.4), also based on [10][20], and specifications following from the suggestions formulated in this article.

A fragment of the XML specification file for Ukrainian compatible with the MTE-4 proposal for Russian:

```
<row role="attribute">
  <cell xml:lang="en" role="position">6</cell>
  <cell role="name" xml:lang="en">Case2</cell>
  <cell xml:lang="en" role="values">
   <table>
    <row role="value">
     <cell role="name" xml:lang="en">genitive</cell>
     <cell role="code" xml:lang="en">g</cell>
    </row>
    <row role="value">
     <cell role="name" xml:lang="en">dative</cell>
     <cell role="code" xml:lang="en">d</cell>
    </row>
    <row role="value">
     <cell role="name" xml:lang="en">locative</cell>
     <cell role="code" xml:lang="en">l</cell>
    </row>
   </table>
  </cell>
 </row>
```

The same fragment for Ukrainian according to our proposals:

```
<row role="attribute">
  <cell xml:lang="en" role="position">6</cell>
  <cell role="name" xml:lang="en">CaseForm</cell>
  <cell xml:lang="en" role="values">
   <table>
    <row role="value">
     <cell role="name" xml:lang="en">first</cell>
     <cell role="code" xml:lang="en">1</cell>
    </row>
    <row role="value">
     <cell role="name" xml:lang="en">second</cell>
     <cell role="code" xml:lang="en">2</cell>
    </row>
    <row role="value">
     <cell role="name" xml:lang="en">third</cell>
     <cell role="code" xml:lang="en">3</cell>
    </row>
   </table>
  </cell>
 </row>
```

[20]We would like to express our gratitude to Tomaž Erjavec for his advice and especially for directing us to the archives of the mailing list for MTE–Russian, which proved a valuable resource for our work on the XML specifications.

12 Conclusions and recommendations

We realise that the suggested modifications entail a need of modifying, or even retagging, corresponding text files in various MTE languages. This should be undertaken only after general agreement on the tagset is achieved among its developers. We do hope that the proposed changes will evoke a wide discussion, and that a common ground will eventually be found.

In its current state the MTE tagset includes information from different levels of language description: purely morphological, derivational, syntactic and semantic. Syntactic and semantic analysis and tagging are further necessary steps in language description, and principles of tagging for them should be developed. The layer of derivation is significant for (semi)automatic lexicon development. This is why the currently encoded information about levels other than the morphological one (such as valency for prepositions or classification of pronoun types) should also be redistributed in the future.

References

[1] Broda B., Piasecki M. and Radziszewski A. (2008). Towards a Set of General Purpose Morphosyntactic Tools for Polish. P*roceedings of Intelligent Information Systems, Zakopane, Poland, 2008*. Institute of Computer Science–PAS.

[2] Derzhanski I. and Kotsyba N. (2008). The category of predicatives in the light of the consistent morphosyntactic tagging of Slavic languages. In *Lexicographic Tools and Techniques: Proceedings of the MONDILEX First Open Workshop*, pages 68–79, Moscow: IITP–RAS.

[3] Dimitrova L., Erjavec T., Ide N., Kaalep H.-J., Petkevič V., Tufiş D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING–ACL '98*, pages 315–319, Montréal, Québec, Canada.

[4] Erjavec, T. (ed.) (2004). *MULTEXT-East Morphosyntactic Specifications: Version 3.0*. Ljubljana.

[5] Kotsyba N., Shypnivska O. and Turska M. (2008). Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). In *Proceedings of Intelligent Information Systems, Zakopane, Poland, 2008*. Institute of Computer Science–PAS.

[6] Kotsyba N., Mykulyak A., Shevchenko I. (to appear). UGTag: morphological analyzer and tagger for Ukrainian language.

[7] Przepiórkowski A. and Woliński M. (2003). A Flexemic Tagset for Polish. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003.*

[8] Sadock, J. (1985). Autolexical syntax: A proposal for the treatment of noun incorporation and similar phenomena. Natural Language and Linguistic Theory, 3, 379–439.

[9] Sauvet G., Włodarczyk A. and Włodarczyk H. (2007). Morphological data exploration using the SEMANA platform: Feature granularity problem in the definition of Polish gender. Lecture slides: ‹http://www.celta.paris-sorbonne.fr/anasem/papers/miscelanea/PolishGender.pps›.

[10] Sharoff S., Kopotev M., Erjavec T., Feldman A., and Divjak D. (2008). Designing and evaluating a Russian tagset. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris, ELRA.

[11] Shevchenko I., Kotsyba N., Kurshuk K. (to appear). Towards the Creation of a Belarusian Grammatical Dictionary.

[12] Turska M. and Kotsyba N. (2007). Polish-Ukrainian Parallel Corpus and its Possible Applications. In *Proceedings of the International Conference 'Practical Applications in Language and Computers', 7–9 April 2005, Łódź*. Peter Lang GmbH.

[13] Włodarczyk, H. (2007). Relewantność cech HUM, ANIM i LOC w gramatyce języka polskiego. Presentation at *The 4th CASK Initiative—Workshop at the Jagiellonian University*, 17–21 April 2007.

[14] Włodarczyk, H. (2008). Pierwsze studium przypadku: problem ziarnistości definicji rodzaju w języku polskim. Presentation at the Institute for Slavic Studies—Polish Academy of Sciences, 14 April 2008.

[15] Woliński, M. (2004). System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica XII, 39–54.

[16] Бірала А. Я., Булахаў М. Г., Жыдовіч М. А., Жураўскі А. І., Карнеева-Петрулан М. І., Крыўчык В. Ф., Лапаў Б. С., Мацкевіч Ю. Ф. (1957). *Нарысы па гісторыі беларускай мовы. Дапаможнік для студэнтаў вышэйшых навучальных устаноў*. Мінск.

[17] Ломтев, Т. П. (1956). *Грамматика белорусского языка*. Минск.

[18] Сцяцко, П. (2002). *Культура мовы*. Мінск: Тэхналогія.

[19] Шевченко, І. В. (1996). Алгоритмічна словозмінна класифікація української лексики. Мовознавство №4–5, 40–44.

[20] Шевченко И. В., Широков В. А., Рабулець А. Г. (2005). Электронный грамматический словарь украинского языка. In *Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей», 27 июня–2 июля 2005 года, Меганом, Крым, Украина*, pages 124–129.

[21] Шерех, Ю. (1951). *Нарис сучасної української літературної мови*. Мюнхен: «Молоде життя».

Establishing Links between Natural Languages and the Universal Dictionary of Concepts

Viacheslav Dikonov

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

Abstract. This article explains how to create dictionaries, which would link the vocabularies of any chosen natural languages with the Universal Dictionary of Concepts [3] and the pivot language UNL. All languages linked with the Universal Dictionary of Concepts become automatically linked with each other at the semantic level of word senses. The article describes the minimal requirements for the contents of such dictionary, explains the principle of data exchange and suggests a possible procedure of producing the dictionaries by merging already existing common lexicographic resources.

1 Introduction

The Universal Dictionary of Concepts (UDC) [3] is the definitive repository of concepts forming the lexicon of the Universal Networking Language (UNL) [4]. The UNL language enables computers to record the meaning of a natural language text, store and exchange semantic information in a standardized form. UNL has many potential applications. For example, it can serve as a pivot language for automatic translation or facilitate unambiguous search in multilingual environments.

There are several linguistic processors developed in different countries, which support the UNL language[1]. Systems which translate text into UNL (enconversion) are called UNL converters. UNL Deconverters are systems that perform the reverse operation (deconversion) and turn UNL documents into texts in some natural language. The list of languages already having a UNL deconverter includes English, Russian, French, Spanish, Arabic, Japanese and more. UNL represents the meaning of a text as a graph joined by semantic relations. The graphs can be visualized and their visual form is intuitively understandable.

The basic elements of UNL and UDC are concepts. Concepts are understood as abstract semantic units more or less equivalent to word senses commonly distinguished by explanatory dictionaries. However, concepts are not bound to concrete words or idiomatic phrases of any particular language. All concepts have their origin in natural languages and should be supported by some linguistic source or a practical need.

Each concept is unambiguously represented by a Universal Word (UW) [2,3,4]. Every UW stands for one and only one concept. Any new concepts receive their own unique UWs. It is possible for technical reasons to have several UWs for one concept (strict synonyms) but such situation is undesirable and should be avoided if possible.

UDC consists of three parts: the repository of concepts, a semantic network establishing relations between concepts, and a number of local dictionaries establishing links between concepts and words or expressions of natural languages. Every language should have its own local dictionary. UDC will be a free public resource constantly developed by the UNL community and any other interested parties.

[1] The projects of making a UNL enconverter and deconverter for the Russian and English languages have received funding from the Russian Foundation for Basic Research (RFBR) under grant agreements 08-06-00367 and 08-06-00344.

2 Local dictionaries

2.1 What is a local dictionary?

Local dictionaries as a whole are one of the key elements of the UNL infrastructure enabling the intermediary language to perform its function of capturing and recording the semantics of any natural language text. Each local dictionary provides a lexical interface between a single natural language and UDC. Any lexicographic resource that describes the polysemy of words of any natural language by linking them with UWs of UDC will qualify as a local dictionary in terms of UDC. Local dictionaries can be used by UNL converters and deconverters to perform automatic or semi-automatic conversion between a natural language text and its semantic representation in UNL.

The exact content of a local dictionary is determined by peculiar properties of the natural language it describes. It is hardly possible to set a rigid standard in this area, but certain common guidelines and principles are essential for interoperability.

A local dictionary can be used for:

1. making the graphical form of the UNL semantic graphs more intuitive for a casual reader or author, who wants to verify the semantic representation of his work
2. semantic markup of corpora, disambiguation of keywords for performing search in UNL or multilingual environment, other cases when lexical disambiguation is necessary
3. finding relations between words of different languages to produce translation dictionaries automatically
4. UNL conversion and deconversion, automatic translation.

Each of the four uses sets different and progressively greater quality and content requirements for a local dictionary. Every new dictionary can be developed gradually through a process of iterative refinement that would make it increasingly bigger, better and more useful. The entry level can be low enough to allow practical use of a bare minimal local dictionary which is just a list of word lemma and UW pairs.

2.2 Levels of quality

The first of the four uses listed earlier is the least demanding. There are specialized software tools to visualize and edit UNL graphs in order to post-correct any errors of an automatic converter. The UWs of UNL are rather long and less familiar to a novice user, so some editors provide an option to display translations instead of the UWs. It helps to see words of a different human language inside the nodes of the graph to quickly assess the quality of lexical disambiguation and spot important errors. Even an incomplete or autogenerated preliminary version of a local dictionary might serve this purpose as soon as it is free from obvious errors. Figure 1 shows an example of a very simple but already useful local dictionary.

Word	**Universal Word**
сказать	say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing)
сказать	tell(icl>narrate>do,cob>uw,agt>person,obj>uw,rec>person)
сказать	say(icl>order>do,agt>volitional_thing,obj>uw,rec>volitional_thing)
сказать	say(icl>imagine>do,agt>person,obj>uw)
человек	person(icl>abstract_thing,equ>personality)
человек	one(icl>unit>thing)
человек	mankind(icl>homo>thing,equ>world)
человек	human(icl>hominid>thing,equ>homo)

Fig. 1: A fragment of a minimalistic Russian local dictionary

The second goal, i.e semantic markup of corpora, is much more demanding from the point of view of dictionary's coverage, correctness and precision. At the same time, the dictionary can still be a simple list of word-UW pairs, supplemented with definitions and examples. The existence of several local dictionaries in UDC makes it possible to retrieve definitions of the concepts in different languages, as shown in Figure 2. The English local dictionary already contains definitions and examples for all concepts in the current version of UDC and POS classes of the linked words are easily deductible from the UWs[2].

Word	**Universal Word**
человек	man(icl>person,equ>human,ant>animal)
	человеческое существо // отряд в пятьдесят человек
	a human being // a hundred men died
человек	person(icl>abstract_thing,equ>personality)
	совокупность черт характера // приятный человек
	the personality of a human being // a nice person
человек	one(icl>unit>thing)
	всякий, любой человек // человек никогда не должен себя ронять
	any person as representing people in general // one should never be complacent
человек	mankind(icl>homo>thing,equ>world)
	человеческая цивилизация // человек шагнул в космос
	all of the living human inhabitants of the earth // one giant leap for mankind
человек	human(icl>hominid>thing,equ>homo)
	биологический вид // человек умелый
	the genus homo // the evolution of humans
человек	man(icl>subordinate>person,equ>agent,pos>person)
	зависимое лицо // человек Путина
	a male subordinate or agent // our man in Habana

Fig. 2: A fragment of the Russian local dictionary with definitions and examples from two local dictionaries

The third possible goal of matching words of multiple natural languages for automated construction of translation dictionaries represents a whole new level of requirements. A very detailed and precise description of polysemy is needed to establish correct translation pairs. Some additional information, such as pragmatic usage tags, e.g. *poet*, *archaic*, *informal*, good definitions and examples in all matched languages, becomes mandatory. Other types of information typically provided by translation dictionaries include morphological and grammatical features, phonetic transcription, sample sentence structures, etc. Figure 3 shows local dictionary entries containing enough data to fill a typical translation dictionary entry and how they combine.

To achieve good results, the coverage and degree of precision should be comparable for all languages involved and sufficient to establish correct translation pairs.

Finally, the fourth and most important use of a local dictionary is automatic translation (MT) through UNL conversion and deconversion. Different linguistic processors set different standards for their dictionaries. Usually such applications favor generalization of word senses to lessen the complexity of dictionaries and disambiguation procedures employed at the stage of syntactic analysis. On the other hand, automatic translation requires full morphological and grammatical information as well as knowledge about combinatorial potential of the word.

[2] All UWs have specific descriptors corresponding to parts of speech provided by the *icl* relation: *do, be, occur* – verbs, **thing, person, animal* etc. – nouns, *adj* – adjectives, *how* – adverbs, *how* in combination with an *obj* constraint – prepositions.

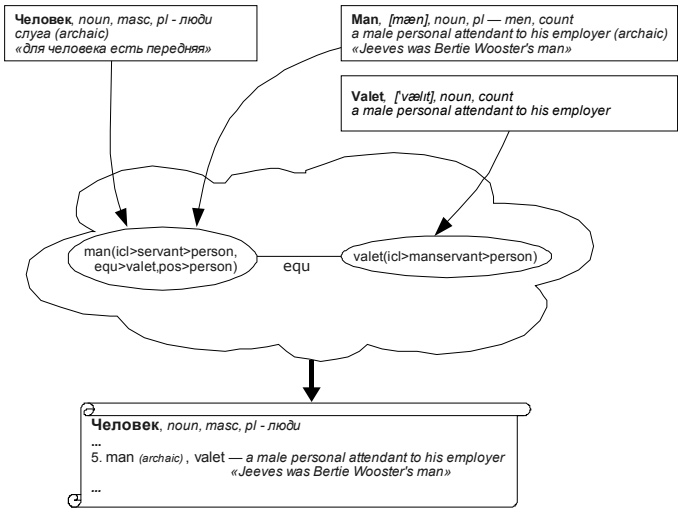


Fig. 3: Russian and English local dictionary entries linked through UDC provide data for construction of a translation dictionary

2.3 Data Exchange

Since local dictionaries are optional parts of UDC and most of them are going to be maintained separately by independent teams, there will be no technical requirements for the storage format or a prescribed set of tools. Instead, there will be a requirement to maintain compatibility of data with UDC and ensure regular reciprocal data exchange. It means that all local dictionaries must synchronize with each new release of UDC to accommodate to any changes in the UW set. At the same time, any changes in a local dictionary that result in adding new concepts or changes of relations between concepts must be submitted to UDC.

Each local dictionary must be machine readable. UDC is going to be stored in an SQL database table, so the local dictionaries should be ready to export and import data in Unicode in a compatible table form either as CSV or XML. The exact technical description of the exchange format does not exist yet. It is going to be designed together with the Internet infrastructure for the UNL dictionary following the availability of the first public release.

All local dictionaries must export at least one data field containing lemmas of the words or expressions associated with UWs of UDC. This field and any additional fields with extra kinds of data are called public. All public data fields involved into the data exchange process need to be marked in a standard and consistent way across all local dictionaries, but their contents may be language specific. A dictionary may contain certain data not relevant to the UNL and UDC project or excluded from the data exchange. Such fields are called private.

We consider it a good practice to keep a copy of every local dictionary that would include all public data fields in the central public database as a safety and informational measure. It will make editing of UNL graphs more convenient by enabling on-the-fly switch from UWs to words of any desired language and help to rebuild any local dictionary in the event of data loss or if the original team ceases to exist.

3 Making of a local dictionary

3.1 General steps

The process of making a local dictionary includes several steps. Some of them can be automated or significantly simplified by re-using existing lexicographic resources and merging their data. The steps are:

1. Identification of word senses (concepts) of a target natural language and definition assignment.
2. Matching of the word senses of the natural language with existing UWs.
3. Creating new UWs for concepts that could not be matched exactly.
4. Linking the new UWs into the semantic network of UDC. It can be done in parallel with stage 3.

This work is quite similar to creation of a Wordnet for the target language. Languages that already have a Wordnet with a good ILI linking it with recent versions of the Princeton Wordnet will have a substantial advantage. Most of the UWs in the current version of UDC are prepared on the basis of Princeton Wordnet [1] v2.1 and can be traced back to the corresponding synsets. UDC will maintain its links with Wordnet to simplify data migration in both directions. Any new and edited UWs, which have their counterparts in Wordnet, should be included in the UDC-Wordnet list of correspondences. Each concept added to UDC will be tagged with its source language. All concepts will also carry a tag with the list of languages that have an exactly matching word sense. The semantic network of UDC will include all links and hierarchy provided by Wordnet and extend it with any missing relations. The combination of these measures will make it possible to extract a Wordnet-type resource for any linked language from UDC.

3.2 Matching word senses and UWs

The list of word senses and their definitions for a chosen language is usually available in the form of an explanatory dictionary[3] while the list of UWs will be provided by UDC. Each UW already has a supposedly self-explanatory name, a definition in English and sometimes an English example. At the current stage of development there are about 200 000 UWs covering the lexicon of the English language and all of them use English words as headwords. It is possible to use a translation dictionary to find English translations of a word. UWs with headwords matching the English translations of the chosen word create a list of candidate UWs for each word sense.

The next step is matching the word senses of each word with candidate UWs. (Fig.4).

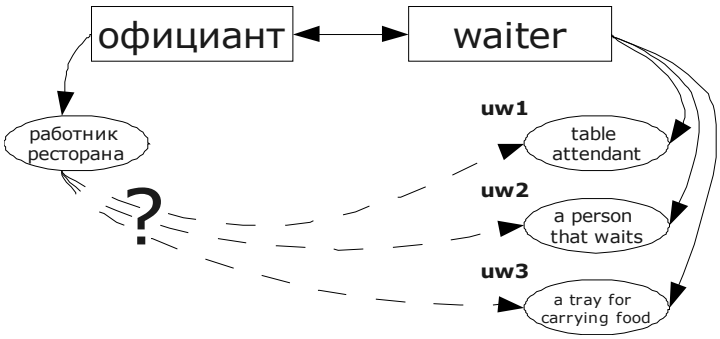


Fig. 4: The word sense matching problem

[3] If no explanatory dictionary is available, as it might happen with some less studied minority languages, there are other ways to identify word senses, e.g. by using text corpora or translation dictionaries.

The choice based on definitions is simple enough for a small number of words but doing it for all UWs is a lot of work. Therefore, it is convenient to have yet another source of information that would help to find certain pairs of word senses automatically. Existing national Wordnets, such as those built for Bulgarian and Czech by the Balkanet project have less coverage than UDC and the Princeton Wordnet, but they provide valuable data for the most frequently used and most polysemous words.

There is a difference between Wordnets and UDC, which becomes evident at this stage. UDC does not treat synsets as monolithic atoms of meaning. Each entry of the dictionary is a single UW. UWs and are still joined by the synonymy relation *equ* into synsets, but UDC permits independent modification of synonyms, recognizing the possibility of subtle differences between them. The synonymy relation is understood as a relation between close but not exactly similar units. Therefore, each synset imported from a Wordnet resource and matched with a set of UWs will produce a set of word-UW pairs (see Fig. 5). Such pairs have high probability of being correct but they must be put to scrutiny as well.

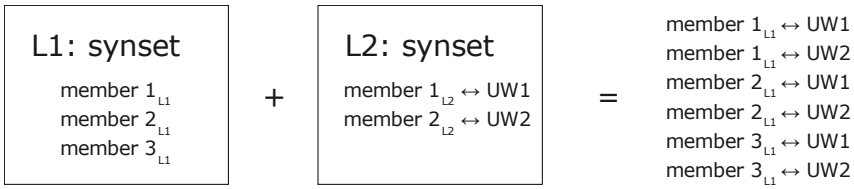


Fig. 5: The result of importing two synsets linked by an ILI

When the process of matching of the word senses with existing UWs is completed, there will be a certain number of word senses left without a matching UW. It is normal, because each language has its own unique conceptual lexicon and it is never fully identical with lexicons of other languages for cultural and historical reasons. The word senses in this list should be added to UDC as new concepts.

3.3 Adding new concepts

Any concept existing in the form of a distinct word sense in any of the linked languages and not found in UDC may and should be added there. A new concept must receive a unique name – a new UW. Local dictionaries cannot reference any UWs not submitted to UDC. Failure to do so may cause incompatibilities between different local dictionaries. There is a standard for UW construction adopted by active UNL centres in Grenoble in 2007[4]. All new UWs submitted to UDC must follow this standard. Malformed UW will be rejected. The designers of the standard can arrange short training courses for those who need to create a large number of UWs.

Every UW consists of a headword and a set of constraints, which describe how the concept represented by the UW is different from the concepts represented by other UWs with the same headword. A constraint consists of a UNL relation and another UW, usually reduced to its headword. The general UW format is:

headword(relation>uw>uw,relation>uw,...)

The headword is usually an English word or phrase. New UWs for concepts related with some previously known concept must be derived from an existing UW by adding or changing constraints. The new constraints must reflect the difference between the new concept and the old one. For example, the first of the following three UWs stands for a general concept of entering into a marriage. The other two are its hyponyms describing two aspects of the action differentiated by some languages, including Russian.

[4] The full description of the standard and detailed guidelines for constructing new UWs are described in a special manual [2]. The manual is still being updated in parallel with the refinement on the initial set of UWs. This work should be completed in summer 2009.

marry(icl>do,agt>person,obj>person) "заключать брак"
*marry(icl>do,agt>*man*,obj>*woman*) "жениться"*
*marry(icl>do,agt>*woman*,obj>*man*) "выходить замуж"*

If the new concept is culture-specific and has no hypernym in English, we can use the native word transliterated into Latin and supplement it with constraints that would link it with the nearest commonly known class of objects.

tarator*(icl>soup(icl>food)>matter)*
lapot*(icl>footwear>..,equ>bast_sandal,com>russian_peasantry)*

UW constraints convey only a minimal amount of information required for identification of concepts. There are three types of constraints: ontological, semantic and argument.

Ontological constraints reflect the most important links between concepts: hypernymy (icl), meronymy (pof), instantiation (iof).

*tongue(**icl**>concrete_thing,pof>body) , madrid(**iof**>city)*

Semantic constraints are used to show the difference between several concepts associated with one headword: synonymy (equ), antonymy (ant), association (com).

*ably(icl>how,**equ**>competently,**ant**>incompetently,com>able)*

Argument constraints reflect the semantic frame of the concept: agent (agt), object (obj), second object (cob), source (src) ...

*buy(icl>get>do,agt>person,**obj**>thing,**cob**>thing,**src**>thing)*

More detailed information about the relations between UWs is going to be stored in the semantic network of the Universal Dictionary of Concepts.

3.4 Linking of concepts into the semantic network

All new concepts should be linked into the semantic network of UDC to maintain integrity of the common dictionary. Linking a concept requires answering several questions, which are usually addressed at the time of construction of a new UW:

1. What is an immediate hypernym or hypernyms of the new concept?
2. What are the immediate hyponyms of the concept?
3. Are there any exact synonyms?
4. Are there any antonyms?
5. What is the semantic argument frame of the concept?

It is possible to create a special software tool to add new concepts to UDC that would provide a wizard interface and reference information to guide the user through the process of creating a new UW and linking it.

3.5 Why linking of new concepts is important

Linking of new concepts extends the semantic network component [3] of UDC. One of its functions is to ensure the ability of UDC and UNL to serve as a pivot for multilingual translation. UDC must always provide a way to find some translation for any word of any supported natural language into any other supported language.

However, objective differences between languages and different approaches towards the degree of granularity and precision of definitions taken by lexicographers will cause situations when different

languages will link to closely related yet different UWs. While Princeton Wordnet sets a common standard it is not always consistent in this aspect. It may happen that some local dictionaries, especially the ones based on richer source data, will go into greater semantic detail while others will link to more general concepts. As a result, some translation equivalents will never be matched (See Figure 6).

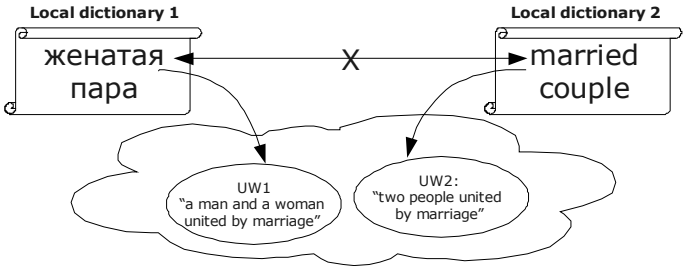


Fig.6: Two words linked to different concepts cannot be matched

A translation for any concept can be found by tracing the ontological (inclusion, instance of, part of) and semantic (synonym of) relations of the semantic network. The rules of finding a translation for a concept that lacks a direct translation into the desired language can be outlined as follows:

1. If a synonym of the concept has a direct translation (member of the same synset), take it.
2. If the concept has immediate hyponyms with translations, choose one of the hyponyms by examining the context e.g . to translate *pedicle* as either *цветоножка* (stem of a flower) or *плодоножка* (stem of a fruit). This is only possible for MT systems.
3. Follow the hypernymy chains until the nearest hypernym with translation is found. If there are several possible paths in the web-like structure, take the shortest one leading to the top parent class specified by the icl restriction of the UW.

The general effect of the third rule applied to an incomplete dictionary resembles the casual speech or speech of an uneducated person, e.g. *give me that thing* (because I do not know its proper name).

4 Summary

This article extends the description of the features and structure of the Universal Dictionary of Concepts in [3]. It shows how to make a local dictionary on the basis of existing lexicographic resources. The advocated incremental manner of development and refinement of a local dictionary allows to obtain some practical result from early steps and find new applications when the quality, content and size become sufficient. The proposed data exchange scheme provides maximum flexibility to the dictionary developers by allowing them to link any suitable resources to UDC regardless of the tools and data formats to used maintain them.

The resulting common multilingual dictionary infrastructure can be used for various linguistic purposes not necessarily related with the development of the UNL project itself. The scheme described in this article is designed to avoid resource fragmentation that became a serious problem in the realm of Wordnets, where multiple projects develop without mutual coordination. Absence of a common data repository for Wordnet-like resources causes huge amounts of useless parallel work. A lot of valuable lexical resources became obscure or simply disappeared after being completed for lack of support and technical maintenance. The Universal Dictionary of Concepts offers a chance to change this situation and accumulate lexicographic data in such way that they will always be readily available to researchers.

References

[1] Fellbaum, C. (1998). WordNet: An Electronic Lexical Database, MIT Press
[2] Boguslavsky. I. (2008). Guidelines for UW construction, manuscript
[3] Boguslavsky I., Dikonov V. (2008). Universal Dictionary of Concepts. In *Proceedings of the First open MONDILEX workshop "Lexicographic Tools and Techniques"*, pages 31–55, Moscow
[4] Web site of the UNL project, `http://www.undl.org`

Lexical Database of the Experimental Bulgarian–Polish Online Dictionary*

Ludmila Dimitrova[1], Rumyana Panova[2], Ralitsa Dutsova[2]

[1] Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences, Sofia

[2] Veliko Tărnovo University & IMI-BAS Master Program, Sofia, Bulgaria

Abstract. In this paper we describe briefly the experimental ongoing version of the Bulgarian–Polish online dictionary. We focus our attention to the lexical database of the dictionary. The starting point for the formal model of lexical database of the dictionary is the CONCEDE model for dictionary encoding. Thus the first Bulgarian-Polish online dictionary will be compatible with other TEI-conformant resources. Some examples from lexical database are presented.

1 Introduction

The base of the first Bulgarian-Polish experimental online dictionary is the ongoing version of the Bulgarian-Polish electronic dictionary [1], [2]. The procedure for selecting the headwords is very simple: we take the headwords from the electronic dictionary. The Bulgarian-Polish electronic dictionary is currently developed in WORD-format in the framework of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI-BAS and ISS-PAS under the supervision of L. Dimitrova and V. Koseska. The current version consists of approximately 20 thousand dictionary entries.

2 Formal model for the Bulgarian-Polish online dictionary encoding

The starting point for the formal model of lexical database (LDB) of the dictionary is the CONCEDE model for dictionary encoding that respect the guidelines of the Text Encoding Initiative Dictionary Working Group (TEI-DWG) [6]. The LDB of the project CONCEDE [4] has standardised and well-understood structure and semantics, and so the first Bulgarian-Polish online dictionary will be compatible with other TEI-conformant resources. With the support of the European Commission the CONCEDE (*Consortium for Central European Dictionary Encoding)* prepared lexical databases for the six Central and East European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene [5]. The first LDB for Bulgarian, more than 2700 lexical entries from the Bulgarian Explanatory Dictionary, based on encoding standards established by the TEI was developed in CONCEDE project [3].

3 Lexical Database

We start to develop the structured LDB taking the recent version of the ongoing Bulgarian-Polish electronic dictionary. This LDB is an entry point to the relational database (RDB) of the Bulgarian-Polish online dictionary. Whenever possible the LDB will generate a new structure of entries for the Polish-Bulgarian online dictionary.

The ***structural tags***, used in the LDB of the Bulgarian-Polish online dictionary, are three: **entry, struc, alt.**

alt: alternation, though generally for use in quite different contexts
entry: dictionary entry
struc: indicates separate independent part in the dictionary entry.

* The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX

The set of ***content tags*** includes the elements:

case: contains grammatical case information given by a dictionary for a given form
conjugation: *a new tag* is added to represent the conjugation of verbs; its structure allows the sub tag **type** for the possible types of conjugations of Bulgarian verbs
def: directly contains the text of the definition
domain: domain
eg: a structure, contains an example, as given in a dictionary, and allows the tags **source** and **q**
etym: a structure, contains etymological information and allows the tags **lang** and **m**, as given in a dictionary
gen: identifies the morphological gender of a lexical item, as given in the dictionary
geo: geographic area
gram: contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, perfect aspect and progressive aspect
hw: the headword; used for alphabetization and indexing, access
itype: indicates the inflectional class associated with a lexical item, as given in a dictionary
lang: language; for use in etymologies (in **etym**)
m: indicates a grammatical morpheme in the context of etymology
mood: contains information about the grammatical mood of verbs, as given in a dictionary
number: indicates grammatical number associated with a form, as given in a dictionary
orth: gives the orthographic form of a dictionary headword
person: indicates grammatical person associated with a form, as given in a dictionary
pos: indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)
q: contains a quotation or apparent quotation
register: register, for type attribute on **usg** tag
source: bibliographic source for a quotation
subc: contains sub-categorization information (transitive/intransitive, countable/non-count, etc.)
time: temporal, historical era, for example, “archaic”, “old”, etc.
type: *a new* subtag in the frame of **conjugation** tag indicates explicitly one of the three types of conjugation of the Bulgarian verbs
tns: indicates the grammatical tense associated with a given inflected form in a dictionary **trans**: contains translation text and related information, so may contain any of the content tags; the principle is that everything under **trans** relates to the target language
usg: contains usage information in a dictionary entry, other than **time**, **domain**, **register** (as these all have their own element), like “dialect”, “folk”, “colloquialism”, etc.
xr: uses to indicate a cross reference with the pointer.

4 Dictionary entry samples

The following samples represent the dictionary entry in XML format and suggest a structure of this dictionary entry in the database of the dictionary to be presented on the Internet. Let us introduce some notation used in the lexical database. We used “’” to mark the accent of the words. The symbol “|” is used to separate the variable part of the word from the main part. The transitive and intransitive verbs should be represented with the corresponding term in the tag **subc**. We introduce “NILL” value in order to represent empty corresponding values.

1) Headword “**притеснение**” */embarrassment/*
притесне’**ни|е, -я** *n* ucisk *m,* udręczenie *n*, uciemiężenie *n*, przygnębienie *n*; kłopoty materialne

```
<entry>
    <hw>притесне’ни|е</hw>
  <alt>
       <orth>-я</orth>
       <num>pl</num>
  </alt>
   <gen>n</gen>
```

```
    <struc type="Sense" n="1">
        <trans>ucisk</trans>
        <gen>m</gen>
        <alt>
            <trans>udręczenie</trans>
            <gen>n</gen>
        </alt>
        <alt>
            <trans>uciemiężenie</trans>
            <gen>n</gen>
        </alt>
        <alt>
            <trans>przygnębienie</trans>
            <gen>n</gen>
        </alt>
    </struc>
    <eg>
        <q>NILL</q>
        <transl>kłopoty materialne</transl>
    </eg>
</entry>
```

2) Headword "поддавам се" /succumb, give way/

подда'ва|м се, -ш *vi*. poddawać się, ulegać, ustępować; **това не се ~ на описание** tego nie da się opisać; **~ ми се нещо** *pot.* coś idzie mi łatwo

```
<entry>
    <hw>подда'ва|м се</hw>
    <pos>v</pos>
    <gram>i</gram>
    <subc>transitive</subc>
    <conjugation>
        <orth>-ш</orth>
        <type>I</type>
    </conjugation>
    <struc type="Sense" n="1">
        <trans>poddawać się</trans>
        <alt>
            <trans>ulegać</trans>
        </alt>
        <alt>
            <trans>ustępować</trans>
        </alt>
    </struc>
    <eg>
        <q>~ това не се ~ на описание</q>
        <transl>tego nie da się opisać</transl>
    </eg>
    <eg>
        <q>~ ми се нещо</q>
        <usg type="register">pot</usg>
        <transl>coś idzie mi łatwo</transl>
    </eg>
</entry>
```

3) Headword "**притежателен**" /*possessive*/

притежа'тел|ен, -на, -но *adi. gram.* dzierżawczy; **~ни местоиме'ния** zaimki dzierżawcze

```
<entry>
    <hw>притежа'тел|ен</hw>
    <alt>
        <orth>-на</orth>
        <gen>f</gen>
    </alt>
    <alt>
        <orth>-но</orth>
        <gen>n</gen>
    </alt>
    <pos>adi</pos>
    <usg type="register">gram</usg>
    <struc type="Sense" n="1">
        <trans>dzierżawczy</trans>
    </struc>
    <eg>
        <q>~ни местоиме'ния</q>
        <transl>zaimki dzierżawcze</transl>
    </eg>
</entry>
```

4) Headword I "**под**" /*under*/, **II** "**под**" / *floor*/

I под *praep.* pod; poniżej; **миньорите работят ~ земята** górnicy pracują pod ziemią; **усмихвам се ~ мустак** uśmiecham się pod wąsem; **държа нещо ~ ключ** trzymam coś pod kluczem; **пет градуса ~ нулата** pięć stopni poniżej zera; **парите са вложени в банката ~ лихва** pieniądze są złożone w banku na procent

II под, -о'ве *m* podłoga *f*

```
<entry n="1">
    <hw>под</hw>
    <pos>praep</pos>
    <struc type="Sense" n="1">
        <trans>pod</trans>
    </struc>
    <struc type="Sense" n="2">
        <trans>poniżej</trans>
    </struc>
    <eg>
        <q>миньорите работят ~ земята</q>
        <transl>górnicy pracują pod ziemią</transl>
    </eg>
    <eg>
        <q>усмихвам се ~ мустак</q>
        <transl>uśmiecham się pod wąsem</transl>
    </eg>
    <eg>
        <q>държа нещо ~ ключ</q>
        <transl>trzymam coś pod kluczem</transl>
    </eg>
```

```
    <eg>
        <q>пет градуса ~ нулата </q>
        <transl>pięć stopni poniżej zera</transl>
    </eg>
    <eg>
        <q>парите са вложени в банката ~ лихва</q>
        <transl>pieniądze są złożone w banku na procent</transl>
    </eg>
</entry>

<entry n=”2”>
    <hw>под</hw>
    <alt>
        <orth>-о’ве</orth>
        <num>pl</num>
    </alt>
    <gen>m</gen>
    <struc type=”Sense” n=”1”>
        <trans>podłoga</trans>
        <gen>f</gen>
    </struc>
</entry>
```

5) Headword “**поддам се**” /succumb, give way/

подд|а’м се, -а’деш *vp.* **v. подда’вам се**

```
<entry>
    <hw>подд|а’м се</hw>
        <pos>v</pos>
        <gram>p</gram>
        <subc>transitive</subc>
        <conjugation>
            <orth>-а’деш</orth>
            <type>I</type>
        </conjugation>
    <xr>подда’вам се</xr>
</entry>
```

5 Relational Database

The model of a relational database is experimentally based on a limited number of studied lexical entries. In the design of the relational database we have provided also the opportunity for translation from Polish to Bulgarian language. That translation will be made only from the main meanings of the Bulgarian headwords. No derivations, phrases or examples will be used for translating from Polish to Bulgarian language.

The relational database is presented on *Figure 1*.

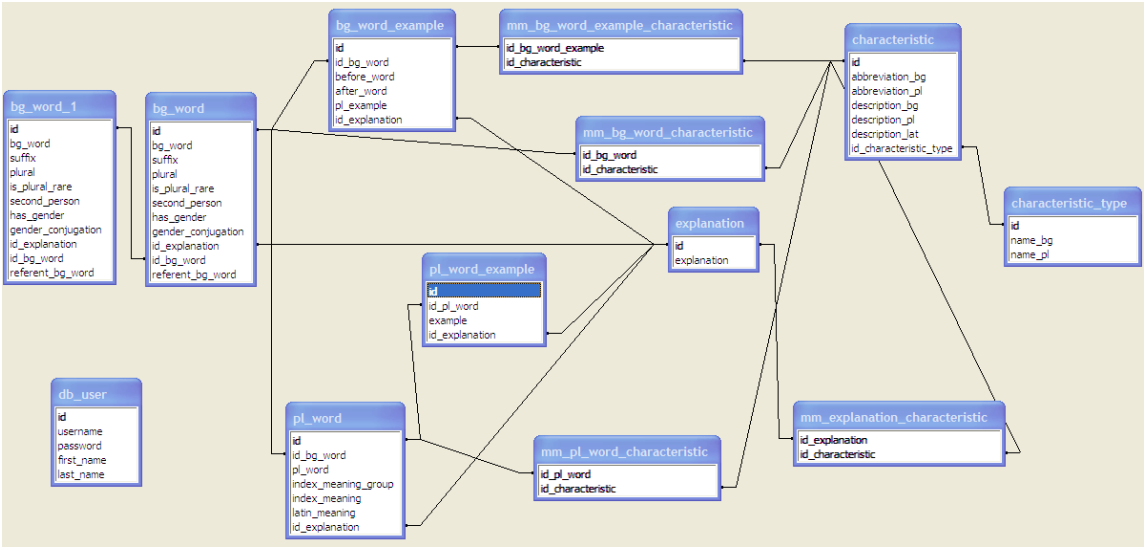


Figure 1: Relational database upon the lexical database of the Bulgarian-Polish-Bulgarian Dictionary

Detailed information on the base units follows.

Table: bg_word

Description: Bulgarian headwords

Field	Type	Null	Default	Comments
id	int(11)	No		Id
homonym_index	int(1)	Yes	*NULL*	Index of the homonym (if null, no homonym exists)
bg_word	varchar(100)	No		Bulgarian headword
suffix	varchar(20)	Yes	*NULL*	Suffix
plural	varchar(20)	Yes	*NULL*	Plural form for a noun
is_plural_rare	int(1)	Yes	*NULL*	Frequency of usage of the plural form for a noun (null – normal, 0 - often, 1 – rare)
conjugation	varchar(20)	Yes	*NULL*	Conjugation form for a verb (2 p., present)
conjugation_type	int(1)	Yes	*NULL*	Type of conjugation for a verb (1, 2 or 3)
has_gender	int(1)	Yes	*NULL*	Whether a noun has feminine and neuter gender
gender_feminine	varchar(20)	Yes	*NULL*	Feminine gender form for an adjective
gender_neuter	varchar(20)	Yes	*NULL*	Neuter gender form for an adjective
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”
id_bg_word	int(11)	Yes	*NULL*	Id of the referent Bulgarian word
referent_bg_word	varchar(255)	Yes	*NULL*	Referent Bulgarian word

Table: bg_word_example

Description: Derivations, phrases or examples of the Bulgarian headwords and their translation in polish

Field	Type	Null	Default	Comments
id	int(11)	No		Id
id_bg_word	int(11)	No		Foreign key to “bg_word”
before_word	varchar(100)	Yes	*NULL*	Text before the headword
after_word	varchar(100)	Yes	*NULL*	Text after the headword
type	int(1)	No		Type of the usage (1 - Derivation; 2 - Phrase; 3 - Example)
pl_translation	varchar(255)	Yes	*NULL*	Polish translation
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

Table: pl_word

Description: Polish headwords

Field	Type	Null	Default	Comments
id	int(11)	No		Id
id_bg_word	int(11)	No		Foreign key to “bg_word”
pl_word	varchar(100)	Yes	*NULL*	Polish headword
sense_index	int(2)	No		Index of the sense
alternative_sense_index	int(2)	No		Index of the alternative sense
latin_translation	varchar(255)	Yes	*NULL*	Latin translation of the word
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

Table: pl_word_example

Description: Examples of the polish headwords

Field	Type	Null	Default	Comments
id	int(11)	No		Id
id_pl_word	int(11)	No		Foreign key to “pl_word”
example	varchar(255)	No		Example in Polish
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

Further improvements will be made when we examine more lexical entries.

6 Web-based Application

The web-based application consists of **administrator and end-user modules**. The administrator module is used to fill in the database and to offer user- friendly interface to the administrators. The idea is that both end-user and administrative parts of the web-based application be bilingual. The following web-based application is experimental, and the structure of the text fields is not permanently determined yet. Changes are possible during the implementation process.

The technologies used for the implementation of the web-based application are Apache, MySQL, PHP and JavaScript. We use free technologies originally designed for developing dynamic web pages with a lot of functionalities. With the help of HTML and CSS we created the designs of both administrative and end user modules. The **administrator module** is intended for the person updating the dictionary. It offers a user-friendly interface for adding, editing, deleting and searching words. The access to the administrative module will be possible only for authorized users. There are possibilities to create more than one user with different passwords and usernames. After the user's password and username have been verified, the user is redirected to the administrative module where there are **several sections** - **section** for entering a new word, **sections** for searching Bulgarian or Polish words, **section** where the user can enter new abbreviations, **section** for setting translations of the user alerts and messages so the user can change the both Polish and Bulgarian translations, **section where** end-users report the missing words. The Help section serves both the administrative and the end users.

Section for entering a new word: from the beginning the user must choose from a combo box what he wants to enter - noun, verb, adjective or any other part of speech (pronoun, conjunction, adverb). Than with the help of AJAX only the corresponding text fields are loaded.

Figure 2: Administrative panel - choosing the type of the word which will be added

When the user wants to add a **new noun** the fields which are necessary for describing nouns are displayed - field for the headword, combo box for choosing the gender of the noun, etc... With the help of AJAX the user has the opportunity to add as many as needed qualified abbreviations like (archaic, dialect, colloquial) or specialized abbreviations like (botanical, chemistry, anatomy, astronomy).

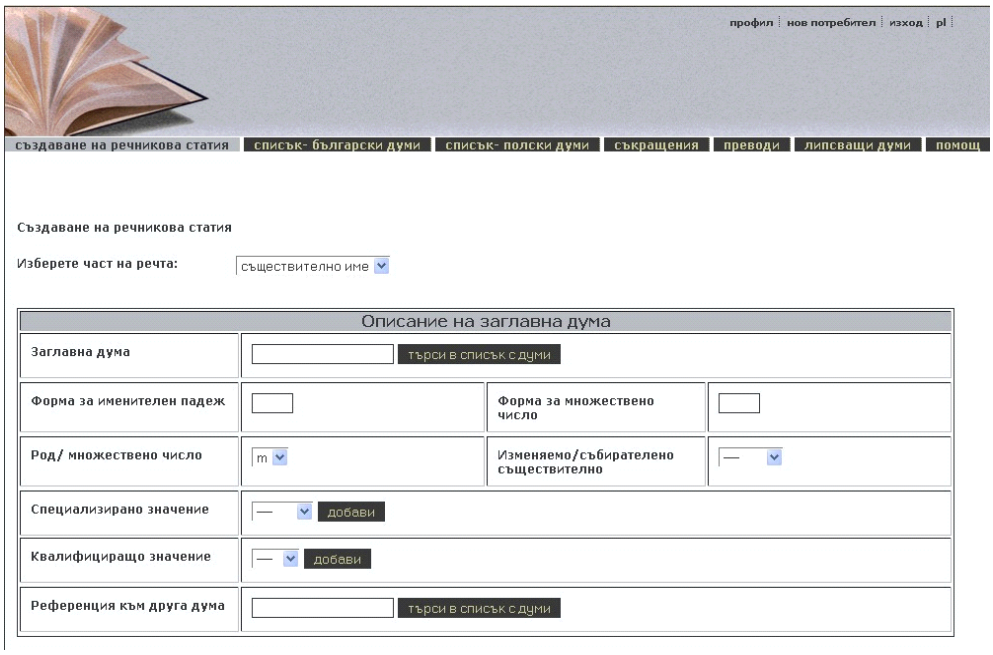


Figure 3: Administrative panel - adding a noun

When the user adds a **new verb** the displayed fields are headword, checkboxes for choosing perfect aspect (vp) or imperfect aspect (vi) of the verb, etc. To display the conjugation of the verb (except showing the conjugation of the verb in 3rd person, singular) we add an extra field where the user can specify the conjugation type. In the help of the administrative module there is an explanation how to determine the conjugation type of any verb.

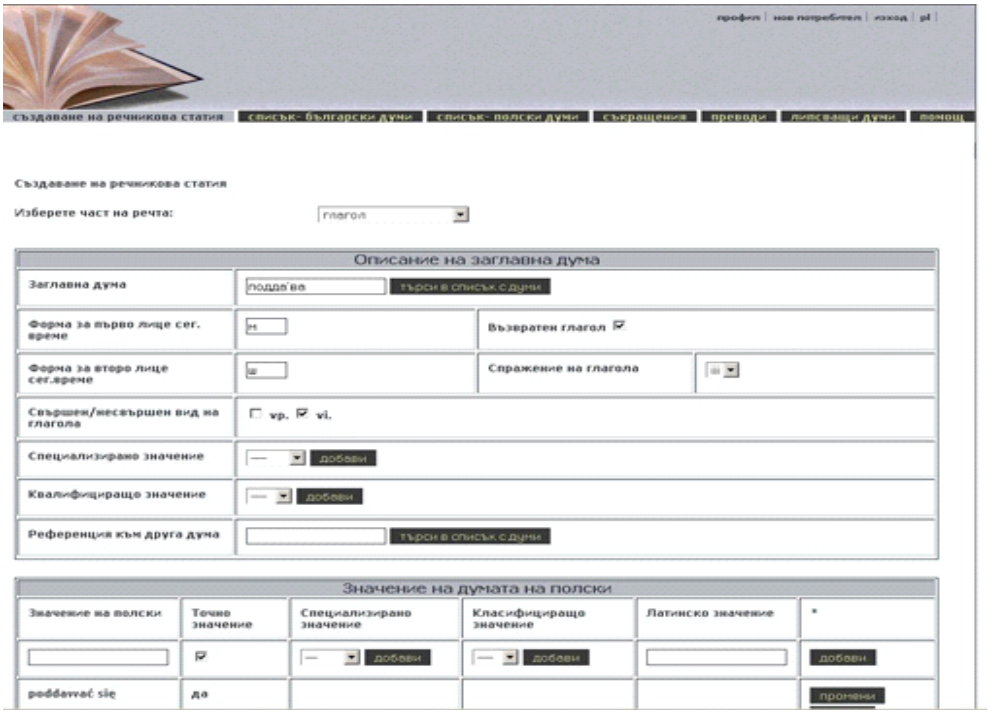


Figure 4: Administrative panel - adding a verb

When adding a **new adjective,** fields specifying the forms for masculine, feminine and neuter are displayed.

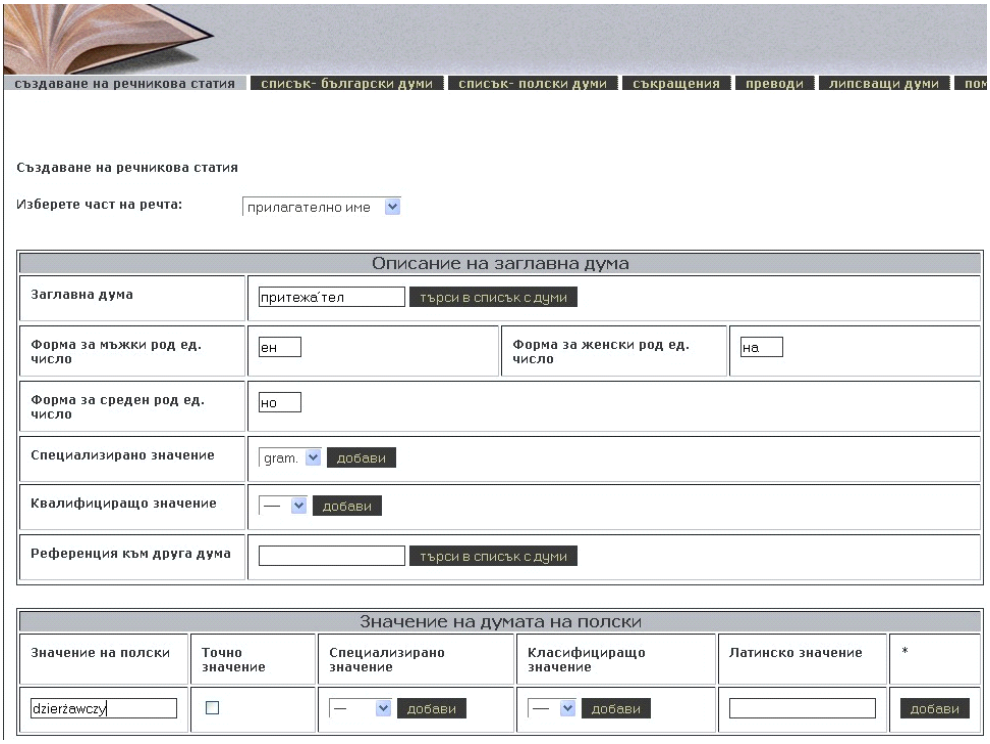


Figure 5: Administrative panel - adding an adjective

There is a common part for each part of speech that ensures the possibility to add unspecified number of derivations, phrases and examples for each headword. At the end of each page for entering headword there is a button “Add derivation / phrase / example”. When the user clicks on it a new window is opened in order to add as many as needed **derivations, phrases and examples** for this headword:

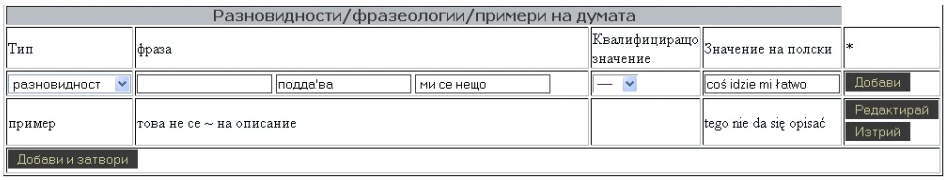


Figure 6: Administrative panel - adding derivations, phrases and examples for the specified headword

Realization of the homonyms in the web-based application: the meanings of the homonyms are entered in the dictionary as different DB records. In the page for entering the words there is a field where the user must specify a homonym index - a number which shows the order of the meanings.
The web-based end-user application is bilingual as well. In this application there are three sections - section for translating a word, information section and section for reporting a missing word. The user can choose the input language (Bulgarian or Polish) and according to it a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Polish characters if they are not supported by the different keyboards.

After making a search for a word on the left site of the screen a list of words, starting from the given entry, are displayed. When clicking on any of these words in the list the translation is visualized in the right frame. If we translate from Bulgarian to Polish, the whole information saved in the RDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized.

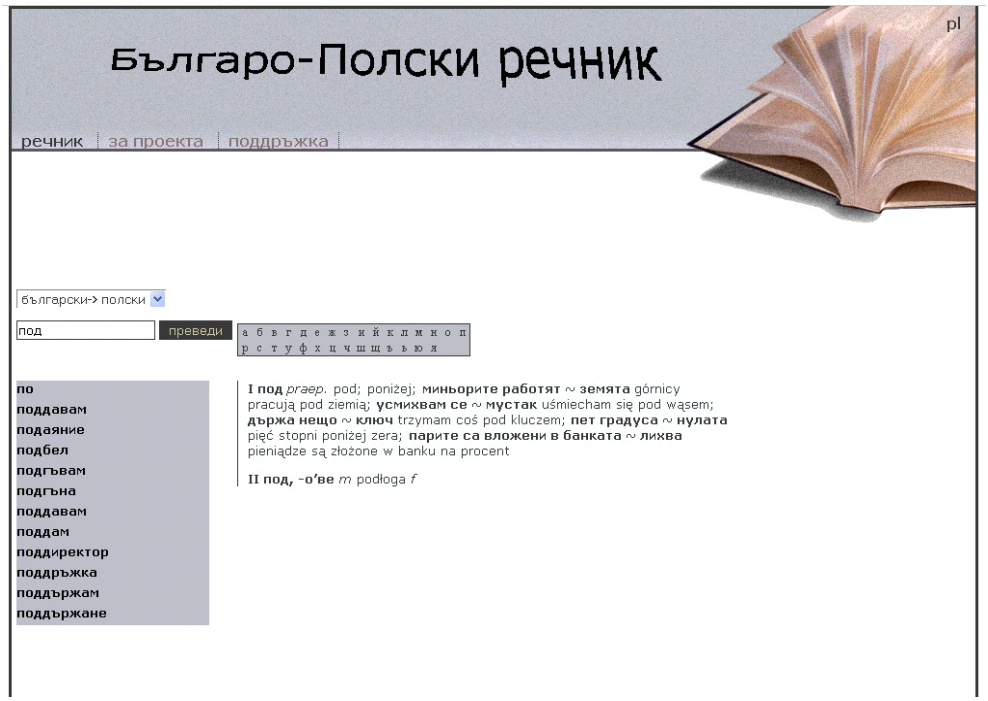


Figure 7: Web page for end users - translation of a Bulgarian word

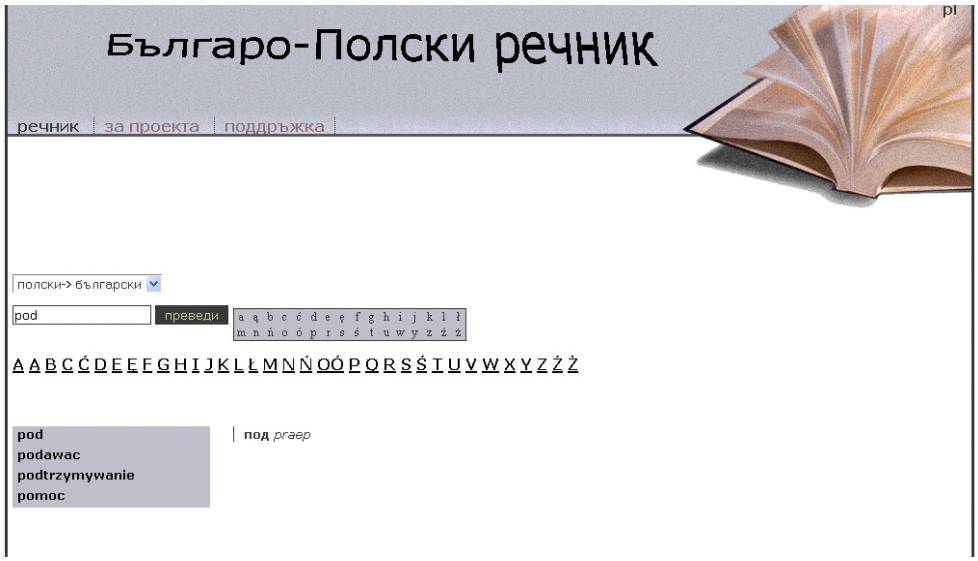


Figure 8: Web page for end users - translation of a Polish word

Both web-based applications have “Help” panels. The end users have the opportunity to report words that are missing in the dictionary into a provided “Contact” form. In this case the administrators will add the reported missing words into the database after.

7 Conclusion

This paper has presented the lexical database of the ongoing version of the first Bulgarian–Polish online dictionary. The formal model of the designed lexical database is CONCEDE model, so the dictionary will be compatible with other TEI-conformant resources.

Due to the limited number of lexical entries taken in consideration, the represented Bulgarian-Polish online dictionary is still at an experimental stage. Further extension of the LDB and RDB will be made.

References

[1] Dimitrova, L., V. Koseska–Toszewa. (2007). Digital Dictionaries – Problems and Features. Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics. 6 July 2007, Sofia, Bulgaria, pages 25–34.

[2] Dimitrova, L., V. Koseska–Toszewa. (2008). Some Problems in Multilingual Digital Dictionaries. International Journal Études Cognitives. SOW, 8, 237–254.

[3] Dimitrova, L., Pavlov, R., Simov, K. (2002). The Bulgarian Dictionary in Multilingual Data Bases. Cybernetics and Information Technologies, 2(2), 33–42.

[4] Tomaž Erjavec, Roger Evans, Nancy Ide, Adam Kilgarriff. (2000). The Concede Model for Lexical Databases. Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00. 355-362, ELRA, Paris.

[5] CONCEDE: `http://www.itri.brighton.ac.uk/projects/concede/`

[6] TEI: `http://www.tei-c.org/index.xml`

Towards a Unification of the Classifiers in Dictionary Entries*

Ludmila Dimitrova[1], Violetta Koseska-Toszewa[2], Joanna Satoła-Staśkowiak[2]

[1] Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences, Sofia
[2] Institute of Slavic Studies,
Polish Academy of Sciences, Warsaw

Abstract. In this paper we continue the discussion of the important problems related to the unification of the classifiers in the electronic dictionary entries, started in [2]. We focus our attention especially to dictionary entries with Bulgarian verbs as headwords. We analyze some examples from ongoing experimental version of the Bulgarian–Polish online dictionary.

1 Introduction

The first Bulgarian–Polish electronic dictionary is being developed in the framework of the cooperation between the Polish and the Bulgarian Academies of Sciences – the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary". 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 will be an entry point to the relational database (RDB) of the Bulgarian-Polish online dictionary. The proposed 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.

2 Classifiers of the Dictionary Entry

As we already wrote [2], [3], one of the main problems of the development of digital dictionaries is the *choice of classifiers* in the dictionary entries. The development of a system of multilingual dictionaries on a basis of bilingual ones requires at first a *unification of the classifiers* in the dictionary entries. The problem turns to the *harmonisation of the classifiers* for various languages, and its solution has to present a *unified selection of classifiers and a standard form of their presentation*.

The comparison of the Bulgarian and Polish material requires an explanation, which is important for the part-of-speech classifiers in the dictionary entries of the cited bilingual electronic dictionary. In the current paper we will mainly analyze the verb entries in both languages.

2.1 Headword in the verb entry

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 "да-construction", which connects the particle "да" 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.

2.2 The phenomenon "transitivity-intransitivity"

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

* The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX

dictionary entry which we consistently follow, makes us confirm the necessity of a classifier reflecting transitivity or intransitivity of the verb [2]. 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).

The authors of "Słownik gramatyczny języka polskiego" [12] propose to exclude the voice category from the explanation of the phenomenon "transitivity-intransitivity". They suggest transitivity and intransitivity to be treated as a syntactic phenomenon. They do not introduce the "voice" category in the description of Polish morphology. Without starting a discussion with them, we must stress that this verbal phenomenon is related to the well-known linguistic fact about the existence of passive participles such as the Polish "chwalony", Bulgarian "хвален" which are frequently used in Polish in nominal constructions, for example *Dziecko często chwalone ma dobre samopoczucie* (an example from the cited "Słownik gramatyczny języka polskiego"). In Bulgarian we have a similar phenomenon, for instance: *Често хваленото дете има добро самочуствие*. The paraphrases of both sentences look alike:

„*Dziecko często chwalone ma dobre samopoczucie*" // *Дете, което е често хвалено, има добро самочувствие*".

In Polish and Bulgarian the verbs which form such passive participles are called transitive. They stand in contrast to the intransitive verbs which do not form such participles, for example in Bulgarian one can say "*Майка му спи*", but there exists no participial **спана*, in Polish "*Matka śpi.*", yet a participial like **spana* is missing.

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 case system, while Polish is a typical synthetic language. It is interesting to note that there exists a third type of classification related to this phenomenon. The above-mentioned authors propose a new classifier (quasi-transitivity). This concerns verbs which are weakly connected to their participle, for instance, *uśmiechnąć się - uśmięchnięty* (in Bulgarian *усмихнат*). In Polish such participles can be formed also from intransitive verbs. That is why this group is called "quasi", for example *Dziewczynka uśmiechnęła się. Uśmiechnęta dziewczynka.* Quasi-transitive verbs exhibit a tendency of exceptions in the classification of transitive and intransitive verbs. If a criterion is introduced such as "in Polish a transitive verb is followed by nouns in accusative case without a preposition", it will verify and clear exceptions from the classification of transitive and intransitive verbs. After *uśmiechnęła się* in Polish there follows no accusative case without preposition. One can not say for example **Dziewczynka uśmiechnęła się kogoś, coś...*, the right sentence is: *Dziewczynka uśmiechnęła się do kogoś, z powodu czegoś...* For this purpose it suffices to place the transitive verbs into a group containing only those which are followed by nouns in accusative case without preposition, such as: *Anna chwali Jasia – Jaś jest chwalony przez Annę.(*Chwali kogo, co?) – Jasia – accusative, animate object, singular. The transitivity of the Polish verb shows that it is always followed by nouns in the accusative case without preposition [12]: 109.

2.3 The "aspect" classifier

The classifier "aspect" of a verb is universally accepted. However we must stress also that the "aspect" classifier in the dictionary entry for a Slavic language is obligatory. The aspect in Slavic languages is a well-formed grammatical category whose meaning boils down to the expression of events – by the perfective aspect – and states – by the imperfective aspect, where we interpret "event" and "state" as described in the net description of temporality in a natural language at the MONDILEX forum [11], [10]. On aspect and the problems of its classification see [8] (in this volume), 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*" [6], according to others the tenses are meanings independent from the meaning of the "aspect" of the verbal form [1].

In languages such as Polish, Czech, Slovak, Ukrainian and Russian, in which "aspect" is a strongly developed semantic and grammatical category, there are few tense forms. This is not the case in South Slavic languages, in which, for example, in Bulgarian, has a high number of tense forms as well as a strongly developed semantic and grammatical category "aspect". As we know, the languages which lack the grammatical category "aspect", such as Latin, French, Italian or Spanish, has a high number of tense forms. As mentioned in [8], there are two distinct tendencies in the South Slavic languages – the first towards reduction of tense forms (Croatian/Serbian), the second one towards reduction or extinction of the aspect. So it should happen in Bulgarian and Macedonian, but does not! The example about the development of the category "aspect" in Bulgarian considered here shows that the development of category "aspect" does not lead to a reduction of the tense forms. Furthermore, as shown by Koseska and Gargov in the second volume of the Bulgarian-Polish Contrastive Grammar, all aspectual-temporal combinations of the verbal form in Bulgarian differ in meaning and are not redundant [9].

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 five Slavic languages in the MONDILEX project. It is an analytic language, and not synthetic (like the rest of the Slavic languages), has not cases (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 five Slavic languages from the MONDILEX project.

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.

2.4. A few short remarks

(1) Gender and number must be specified for the nouns and adjectives because in the two languages these classifiers may vary. For example, the Bulgarian noun "стая" /*room*/ is feminine, while the Polish "pokój" /*room*/ is masculine.

(2) The problem about adverb classification requires a separate study. In the literature on adverbs there are no clear-cut criteria about this part-of-speech.

3 Bulgarian-Polish dictionary entries analysis

Here we give an overview of some dictionary entries from the future Bulgarian-Polish online dictionary. The dictionary entries are divided in two groups, the first containing entries whose headwords belong to the open parts of speech - verbs (incl. verbal forms, esp. Bulgarian participles), nouns, adjectives, adverbs, and the second group comprises closed parts of speech (numerals, pronouns, conjunctions, prepositions, particles and interjections).

We plan to use the CONCEDE model [7] for dictionary encoding that respects the guidelines of the Text Encoding Initiative Dictionary Working Group (TEI-DWG) (TEI). The CONCEDE project (CONCEDE), supported by the European Commission under INCO-Copernicus program, developed a formal model for lexical databases (in the form of an SGML DTD). The lexical databases in accordance with the guidelines of the

TEI-DWG for the six Central and East European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene were developed. In CONCEDE, all dictionaries use common tagset [5]. In the framework of the project the first LDB for Bulgarian, based on encoding standards established by the TEI, was developed [4].

3.1. Lexical database of the Bulgarian-Polish online dictionary

The tagset for LDB of the Bulgarian-Polish online dictionary contains 3 structural tags and a set of content tags. The full list of tags can be found in the Appendix.

(1) The structural tags are:

alt – a tag indicates alternation, though generally for use in quite different contexts,

entry - a tag, contains the dictionary entry,

struc- a tag indicates separate independent part in the dictionary entry:

```
<entry>
                    <alt>...</alt>
                     <struc type=”Sense” n=”1”>…</struc>
                     <struc type=”Sense” n=”2”> …</struc>
                     …
</entry>
```

(2) The set of content tags includes all other tags, among them:

The **hw** tag contains the headword and is used for alphabetization and indexing, access. The **pos** tag indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.):

<hw>**свобод|а’**</hw><pos>noun</pos>.

The **xr** tag uses to indicate a cross reference with the pointer:

<hw>**построя’ва|м**</hw> <xr>**постро|я’**<xr>.

The **orth** tag gives the orthographic form of words (part of word): <orth>-**и’**</orth>.

The **gram** tag contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, perfective aspect and imperfective (progressive) aspect: <gram>imperfective</gram>.

The **subc** tag contains sub-categorization information (transitive/intransitive for verbs, countable/non-count for nouns, etc.): <subc> transitive </subc>.

We suggest new tags, **conjugation** and **type,** to represent the conjugation of verbs -

conjugation: to represent the conjugation of verbs; its structure allows the sub tag **type** for the possible types of conjugations of Bulgarian verbs;

type: a tag in the frame of **conjugation** tag indicates explicitly one of the three types of conjugation of the Bulgarian verbs, for example:

```
<conjugation>
        <orth>-ш</orth>
        <type>I</type>
</conjugation>
```

The **trans** tag contains translation text and related information, everything under **trans** relates to the target language: <trans>**wolność**</trans>.

The **eg** tag forms a structure, contains an example, as given in a dictionary, and allows the tags **source** and **q;** the **q** tag contains a quotation or apparent quotation, the **source -** bibliographic source for a quotation: <eg><q>**-я на учи’лище**</q><trans**> chodzę do szkoły** </trans></eg>.

3.2. Examples

The examples contain the dictionary entry in WORD format and a comment on its classifiers. For verbs in particular we suggest a structure of dictionary entry in the LDB of the Bulgarian-Polish online dictionary.

(1) Verbs (*глаголи, czasowniki*):

(1.1) Entry in WORD-format:

постро|я́, -и́ш *vp.* zbudować; uszeregować, uszykować

Comment:

Verb: build/construct */построя/;* **aspect: perfect** */свършен вид/*, **transitive verb** */преходен/,* **-и**’**ш conjugation II type** */II спрежение/*

LDB structure:

```
<entry>
<hw> построя’</hw>
        <pos>verb</pos>
        <gram> perfect </gram>
        <conjugation><orth>-иш</orth>
        <type>II</type>
        </conjugation>
<alt>
<orth>построя’ва|м </orth>
        <gram>imperfect</gram>
        <conjugation><orth>-ш</orth>
        <type>II</type>
        </conjugation>
</alt>
        <subc>transitive</subc>
<struc type="Sense" n="1">
        <trans>zbudować</trans>
</struc>
<struc type="Sense" n="2">
        <trans>uszeregować</trans>
        <alt><trans>uszykować</trans></alt>
</struc>
</entry>
```

(1.2) Entry in WORD-format:

постро|я́ва|м, -ш *vi*. v. **постро|я́**

Comment:

Verb: build/construct */построявам* **/, aspect: imperfect (progressive)** */несвършен вид/***, transitive verb** */преходен/,* **-ш conjugation III type** */III спрежение/*

LDB structure:

```
<entry>
<hw>построя’ва|м</hw>
        <xr>построя’</xr>
</entry>
```

(1.3) Entry in WORD-format:

ви́жда|м, -ш *vi*. widzieć; **~м се** widzieć się; **~ се** zdaje się, wydaje się; widać

Comment:

Verb: see /*виждам/*, **aspect: imperfect (progressive) /***несвършен вид***/, transitive verb /***преходен***/**, **-ш conjugation III type /***III спрежение***/,** *czas. ndk* **widzieć ~dzę, ~dzisz** *czas. ndk VIIa*; **~ м се** widzieć się; **~ се** zdaje się, wydaje się; widać

LDB structure:

```
<entry>
<hw>ви’ждам</hw>
        <pos>verb</pos>
        <gram>imperfect</gram>
        <conjugation><orth>-ш</orth>
        <type>III</type>
        </conjugation>
        <subc>transitive</subc>
<struc type="Sense" n="1">
        <trans> widzieć </trans>
</struc>
<struc type="Derivation" n="1">
        <orth>~м се</orth>
        <struc type="Sense" n="1">
        <trans> widzieć się</trans>
        </struc>
</struc>
<struc type="Derivation" n="2">
        <orth>~ се</orth>
        <struc type="Sense" n="1">
        <trans> zdaje się </trans>
        <alt><trans> wydaje się </trans></alt>
        </struc>
</struc>
<struc type="Sense" n="2">
        <trans> widać </trans>
</struc>
</entry>
```

(1.4) Entry in WORD-format:

сп|я́, -и́ш *vi*. spać; **~и́ ми се** chce mi się spać, ogarnia mnie senność

Comment:

Verb: sleep /*спя***/, aspect: imperfect (progressive) /***несвършен вид/,* **intransitive verb /***непреходен***/,** **conjugation II type /***II спрежение***/**

LDB structure:

```
<entry>
<hw>сп|я’</hw>
        <pos>verb</pos>
        <gram>imperfect</gram>
        <conjugation><orth>-и’ш</orth>
        <type>II</type>
        </conjugation>
```

```
            <subc>intransitive</subc>
<struc type="Sense" n="1">
            <trans> spać </trans>
</struc>
<struc type="Derivation" n="1">
            <orth>~и́ ми се</orth>
<struc type="Sense" n="1">
            <trans> chce mi się spać </trans>
            <alt><trans> ogarnia mnie senność </trans></alt>
</struc>
</struc>
</entry>
```

(1.5) Entry in WORD-format:

хо́д|я, -иш *vi.* chodzić; kursować; **~и слу́х (мълва́)** *lud.* chodzą słuchy, pogłoski; **-я на учи́лище** chodzę do szkoły; **~я си** odchodzę, idę sobie; **~и ми се на ки́но** mam ochotę pójść do kina; **~я ерге́н** jestem kawalerem

Comment:

Verb: walk, go */хо́дя/*, **aspect: imperfect (progressive)** */несвършен вид/*, **intransitive verb** */непреходен/,* **conjugation III type** */III спрежение/*

LDB structure:

```
<entry>
<hw> хо'д|я </hw>
            <pos>verb</pos>
            <gram>imperfect</gram>
            <conjugation><orth>-и'ш</orth>
            <type>III</type>
            </conjugation>
            <subc>intransitive </subc>
<struc type="Sense" n="1">
            <trans> chodzić </trans>
</struc>
<struc type="Sense" n="2">
            <trans> kursować </trans>
</struc>
<struc type="Phrases"><struc type="Phrase" n="1">
            <orth>~и слух (мълва́) </orth>
            <usg type="register"> lud.</usg>
            <trans> chodzą słuchy, pogłoski </trans>
</struc></struc>
<eg><q>-я на учи́лище</q><trans> chodzę do szkoły </trans></eg>
<eg><q>~я си </q><trans> odchodzę </trans>
<alt><trans> idę sobie </trans></alt></eg>
<eg><q>~и ми се на ки́но </q><trans> mam ochotę pójść do kina </trans></eg>
<eg><q>~я ерге́н </q><trans> jestem kawalerem </trans></eg>
</entry>
```

We remark here that the suggested LDB structure of Bulgarian-Polish dictionary entry is suitable for automated generation of a Polish-Bulgarian dictionary entry. For example, from this one in (1.5), a program could generate automatically the simple structures for the corresponding Polish verbs **chodzić** and **kursować:**

<entry>
<hw> **chodzić** </hw>
<pos>verb</pos>
<struc type="Sense" n="1">
<trans**> хо'д|я** </trans>
</struc>
</entry>
<entry>
<hw> **kursować** </hw>
<pos>verb</pos>
<struc type="Sense" n="1">
<trans**> хо'д|я** </trans>
</struc>
</entry>

All others classifiers for the Polish verbs in these entries, derivations, phrases, examples, etc., should be added additionally!

(1.6) **Participle (*причастие, imiesłow*)**

Entry in WORD-format:

следващ *imiesł. przym.* **1. studiujący** *imiesł. przym.*; **2. idący** *imiesł. przym.*, następujący za kimś, następny

Comment:

Participle: next */ следващ/ imiesł. przym.* **1. studiujący** *imiesł. przym.*; **2. idący** *imiesł. przym.*, następujący za kimś, następny.

(2) Nouns (*съществителни имена, rzeczowniki*):

(2.1) Entry in WORD-format:

хо́ра *pl* ludzie *pl*

Comment:

Noun: **people** /*xopa/ rzecz. l.mn* **(plural) ludzie** *rzecz. l.mn* **(plural)**

(2.2) Entry in WORD-format:

свобод|а́, -и́ *f* wolność *f*, swoboda *f*

Comment:

Noun: **freedom /*свобода*/, -и (plural)** *rzecz. ż* (**gender**) **1.wolność** *rzecz. ż*, **2. swoboda** *rzecz. ż*

(3) Adjectives (*прилагателни имена, przymiotniki*):

(3.1) Entry in WORD-format:

мек adi. miękki; łagodny; **~а дъждовна вода́** miękka deszczowa woda; **~а зима** łagodna zima; **~и съгла́сни** *gram*. spółgłoski miękkie; **~а ша́пка** kapelusz (męski)

Comment:

Adjective: **soft** /*мек/ przym.* **1. miękki** *przym.***; 2. łagodny** *przym.*; **~а дъждовна вода** miękka deszczowa woda; **~а зима** łagodna zima; **~и съгласни** *gram.* spółgłoski miękkie; **~а шапка** kapelusz (męski)

(3.2) Entry in WORD-format:

и́стински adi. prawdziwy; adv. naprawdę, prawdziwie

Comment:

Adjective**: true** /*истински/ przym.* **prawdziwy** *przym.*; *przysłów*. naprawdę, prawdziwie

(4) Adverbs (*наречия, przysłówki*):

(4.1) Entry in WORD-format:

рядко *adv.* rzadko

Comment:

Adverb: **seldom** /*рядко/ przysłów.* **rzadko** *przysłów.*

(4.2) Entry in WORD-format:
ско́ро *adv.* prędko, rychło, szybko**;** niedawno, wkrótce; **мно́го ~ свърших та́я ра́бота** bardzo prędko skończyłem tę pracę; **ще се върна ~** wkrótce wrócę; **ча́с по́-~** czym prędzej
Comment:
Adverb: **soon** */скоро/ przysłów.* **1. prędko** *przysłów.***, 2. rychło** *przysłów.***, 3. szybko** *przysłów.***; 4. niedawno** *przysłów.***, 5. wkrótce** *przysłów.*; **много ~ свърших тая работа** bardzo prędko skończyłem tę pracę; **ще се върна ~** wkrótce wrócę; **час по-~** czym prędzej

(5) Pronouns (*местоимения, zaimki*):
Entry in WORD-format:
не́гов *pron. poss.* jego
Comment:
Pronoun: **his, its /***негов/ zaimek dzierż.* **jego** *zaimek dzierż. r. męski* (**gender**) *D. B.*

(6) Conjunctions (*съюзи, spójniki*):
Entry in WORD-format:
но *coni.* ale, lecz; **не са́мо то́й, ~ и а́з** nie tylko on, ale i ja; **и́скат, ~ не мо́гат** chcą, ale nie mogą
Comment:
Conjunctions**: but /***но/ spójnik* **1. ale** *spójnik*, **2. lecz** *spójnik*; **не са́мо той, ~ и а́з** nie tylko on, ale i ja; **и́скат, ~ не мо́гат** chcą, ale nie mogą

(7) Prepositions (*предлози, przyimki*):
Entry in WORD-format:
пред *praep.* przed; wobec; ~ **университе́та** przed uniwersytetem; **явя́вам се ~ съда́** stoję przed sądem; **вино́вен съм ~ ва́с** czuję się wobec was winny; **вси́чки гра́ждани са ра́вни ~ зако́на** wszyscy obywatele są równi wobec prawa; **оста́на глу́х ~ молби́те му** pozostał głuchy na jego prośby; **и́мам ~ ви́д** mam na uwadze; ~ **ви́д на ...** z uwagi na…
ze względu na…
Comment:
Preposition: **in front of; before; at; to; /***пред/ przyim.* **1. przed** *przyim.***; 2. wobec** *przyim.*;

(8) Particles (*частици, partykuły*):
Entry in WORD-format:
не *partyk. przecząca* nie
Comment:
Particle: **no** *не partyk. przecząca* **nie** *partyk. przecząca*

(9) Numerals (*числителни имена, liczebniki*):
Entry in WORD-format:
четири́ма *num.* czterej; czworo
Comment:
Numeral: **four persons /***четирима/ liczeb.* 1st sense: **czterej;** 2nd sense**: czworo** *liczeb.*

(10) Interjections (*междуметия, wykrzykniki*):
Entry in WORD-format:
о́х! *interi.* o!, och! (na wyrażenie bólu, smutku, radości, zachwytu, zdziwienia itp.)
Comment:
Interjection: **oh /***ox!/ wykrzyk.* **o!, och!** *wykrzyk.* (Explanation: na wyrażenie bólu, smutku, radości, zachwytu, zdziwienia itp.)

4. 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.

References

[1] Andrejchin, L. (1944). Основна българска граматика. София. (in Bulgarian)
[2] Dimitrova, L., Koseska-Toszewa, V. (2008). The Significance of Entry Classifiers in Digital Dictionaries. In Proceedings of the MONDILEX First Open Workshop, Moscow, Russia, 3–4 October 2008, pages 89–97, Russian Academy of Sciences, IITP.
[3] Dimitrova, L., Koseska-Toszewa, V. (2008). Some Problems in Multilingual Digital Dictionaries. SOW, 8, 237–254.
[4] Dimitrova, L., Pavlov, R., Simov, K. (2002). The Bulgarian Dictionary in Multilingual Data Bases. Cybernetics and Information Technologies, 2(2), 33–42.
[5] Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2000). The Concede Model for Lexical Databases. Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00. 355-362, ELRA, Paris.
[6] Ivanchev, S. (1971). Проблеми на аспектуалността в славянските езици. София. (in Bulgarian)
[7] Kilgarriff, A. (1999). Generic encoding principles. CONCEDE Project Deliverable 2.1. University of Brighton, UK.
[8] Koseska – Toszewa, V. (2009). Form, its meaning, and dictionary entries (in this volume)
[9] Koseska, V., G. Gargov. (1990). Bulgarian-Polish Contrastive Grammar, vol. 2. Special Definiteness-Indefiniteness category, Sofia. (in Bulgarian)
[10] Koseska, V., Mazurkiewicz, A. (2009) Net-Based Description of Modality in Natural Language (on the Example of Conditional Modality). Proceedings of the MONDILEX Open Workshop, Kiev, 2–3 February 2009. (be appear)
[11] Mazurkiewicz, A. (2008) A Formal Description of Temporality (Petri net approach). Proceedings of the MONDILEX First Open Workshop, Moscow, Russia, 3–4 October 2008. pages 98–108.
[12] Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz R. (2007) Słownik gramatyczny języka polskiego, Wiedza Powszechna, Warszawa.CONCEDE:
[13] http://www.itri.brighton.ac.uk/projects/concede/
[14] TEI: http://www.tei-c.org/index.xml

Appendix

The structural tags, used in the LDB of the Polish-Bulgarian online dictionary, are three:
entry, struc, alt.

alt: alternation, though generally for use in quite different contexts
entry: dictionary entry
struc: indicates separate independent part in the dictionary entry.

The set of ***content tags*** includes the elements:

case: contains grammatical case information given by a dictionary for a given form
conjugation: ***a new tag*** is added to represent the conjugation of verbs; its structure allows the sub tag **type** for the possible types of conjugations of Bulgarian verbs
def: directly contains the text of the definition
domain: domain
eg: a structure, contains an example, as given in a dictionary, and allows the tags **source** and **q**
etym: a structure, contains etymological information and allows the tags **lang** and **m**, as given in a dictionary
gen: identifies the morphological gender of a lexical item, as given in the dictionary
geo: geographic area
gram: contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, perfect aspect and progressive aspect
hw: the headword; used for alphabetization and indexing, access
itype: indicates the inflectional class associated with a lexical item, as given in a dictionary
lang: language; for use in etymologies (in **etym**)
m: indicates a grammatical morpheme in the context of etymology
mood: contains information about the grammatical mood of verbs, as given in a dictionary
number: indicates grammatical number associated with a form, as given in a dictionary
orth: gives the orthographic form of a dictionary headword
person: indicates grammatical person associated with a form, as given in a dictionary
pos: indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)
q: contains a quotation or apparent quotation
register: register, for type attribute on **usg** tag
source: bibliographic source for a quotation
subc: contains sub-categorization information (transitive/intransitive, countable/non-count, etc.)
time: temporal, historical era, for example, “archaic”, “old”, etc.
type: ***a new*** subtag in the frame of **conjugation** tag indicates explicitly one of the three types of conjugation of the Bulgarian verbs
tns: indicates the grammatical tense associated with a given inflected form in a dictionary **trans**: contains translation text and related information, so may contain any of the content tags; the principle is that everything under **trans** relates to the target language
usg: contains usage information in a dictionary entry, other than **time**, **domain**, **register** (as these all have their own element), like “dialect”, “folk”, “colloquialism”, etc.
xr: uses to indicate a cross reference with the pointer.

MULTEXT-East Morphosyntactic Specifications: Towards Version 4*

Tomaž Erjavec

Department for Knowldege Technologies
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract. The MULTEXT-East standardised and linked set of language resources covers a large number of mainly Central and Eastern European languages and includes harmonised morphosyntactic resources consisting of the specifications, lexica and a parallel corpus. The MULTEXT-East resources, currently at Version 3, are freely available for research use and have been used in numerous studies connected to language technologies. In this paper we concentrate on MULTEXT-East morphosyntactic specifications, which define the features that describe word-level syntactic annotations, and explain their structure in Version 4, currently work in progress. The V4 specifications are planned to cover at least 13 languages and will be encoded in XML, according to the latest version of the Text Encoding Initiative Guidelines, TEI P5. The new encoding enables more flexible language-particular encodings, localisations of feature names and codes, easy generation of derived formats (HTML, tabular, XML libraries), and simplifies the addition of new languages.

1 Introduction

The MULTEXT-East project, (Multilingual Text Tools and Corpora for Eastern and Central European Languages) was a spin-off of the EU MULTEXT project [14]; MULTEXT-East ran from '95 to '97 and developed standardised language resources for six CEE languages [3], as well as for English, the 'hub' language of the project. The main results of the project were lexical resources and an annotated multilingual corpus, where the most important resource turned out to be the parallel corpus – heavily annotated with structural and linguistic information – which consists of Orwell's novel "1984" in the English original and translations.

In addition to delivering resources, a focus of MULTEXT-East was also the adoption and promotion of encoding standardisation. On the one hand, the morphosyntactic annotations and lexica were developed in the formalism used for six Western European languages in the MULTEXT project, itself based on the EAGLES specifications [5]. On the other, all the corpus resources were encoded in SGML, according to the Corpus Encoding Standard [12] and, later, in XML and TEI, the Text Encoding Initiative Guidelines [19].

One of the objectives of MULTEXT-East has been to make its resources available to the wider research community. The resources are distributed on the Web at `http://nl.ijs.si/ME/`. A portion of the resources is freely available for download or browsing; for the rest, the user has to first fill out a Web-based agreement form restricting the use of resources for research. Apart from the data itself, the distribution also contains extensive documentation.

After the completion of the EU MULTEXT-East project, a number of other projects have helped to keep the MULTEXT-East resources up-to-date (e.g., migrating the corpus from SGML to XML) and enabled us to add new languages. At the time of writing, the latest publicly released resources are at Version 3 [7].

The MULTEXT-East resources have been instrumental in advancing the state-of-the-art in language technologies in a number of areas, e.g., part-of-speech tagging [21], inductive learning of lemmatisation rules [9], and word sense disambiguation [13], to mention just a few. The licensing form has been submitted by over 100 organisations, mostly academia, but also industry.

* The study and preparation of these results have received funding from the EU 7FWP under grant agreement 211938 MONDILEX.

The success of the resources is mostly due to the fact that they are freely available for research and that they include basic building blocks for processing a significant range of "novel" languages. As the linguistic markup has also been manually validated and tested in practice, the resources can serve as a "gold standard" which enables other researchers to develop and test their approaches to topics in the language processing. The resources also provide a model which languages lacking basic linguistic resources, such as tagsets, lexica and annotated corpora can link-up to, taking a well-trodden path. This aspect of the resources was unexpected but highly rewarding; this steady addition of new languages also gives impetus to continue working on their general improvement.

Since the release of Version 3 the resources have again been expanded and re-encoded, in preparation for Version 4. New languages have been added and the morphosyntactic specifications have been converted from the LaTeX format to XML [8]. A portion of the resources has also been additionally annotated, e.g., for WordNet word-sense disambiguated nouns [13] in the English "1984" and dependency syntactic structures for the Slovenian "1984" [4].

This paper is devoted to one part of the resources, namely the MULTEXT-East morphosyntactic specifications. The specifications are a document that provides the definition of the attributes and values used by the various languages for word-class syntactic annotation, i.e., they provide a formal grammar for the morphosyntactic properties of the languages covered. The MULTEXT-East specifications define 12 categories (parts-of-speech), and approx. 100 different attributes with 500 values.

The morphosyntactic specifications also define the mapping between feature-structures and morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and for corpus annotation. For example, they specify that MSD `Ncms` is equivalent to the feature-structure consisting of the attribute-value pairs `Category:Noun`, `Type:common`, `Gender:masculine`, `Number:singular`. The specifications furthermore determine which feature-value combinations and MSDs are valid for particular languages. In addition to the formal parts the specifications also contain commentary, bibliography, etc.

Originally, these specifications were released as a report of the MULTEXT-East project but have been revised for both subsequent releases, and have become, if not a standard, then at least a reference for comparison [11]. They currently cover thirteen languages; Table 1 gives an overview, and for each language also specifies its language family, and which version of the MULTEXT-East resources it first appeared or will appear in. Special mention deserve the languages which still have to make their debut in Version 4, namely Macedonian, Persian, and Russian, and, to an extent, Slovene. The development of the Macedonian specification, lexicon and corpus started in 2004, and the resources have already been used as the data for several experiments in tagger [22] and lemmatiser induction [15]. The Macedonian resources comprise the specifications, lexicon, and corpus, which is, however, not yet morphosyntactically annotated. The development of Persian resources also started in 2004, and they currently comprise the specifications and annotated corpus [17]. The Russian specifications [18] are the latest addition, although the (unannotated) corpus has been available since Version 1. The Russian resources thus still lack a lexicon and annotated corpus, although an automatically annotated corpus and tagging models are available independently at `http://corpus.leeds.ac.uk/mocky/`.

Slovene has been a part of the MULTEXT-East resources from the start, however, in Version 4 we plan to significantly revise the specifications and harmonise the lexicon and corpus with them. The Slovene specifications have been extensively used for corpus annotation, esp. of the Slovene reference corpora Fida and its successor FidaPLUS (`http://www.fidaplus.net/`) and in the course of the years various shortcomings of the original proposal have come to light. A recent Slovene project, JOS (Jezikoslovno označevanje slovenščine / Linguistic Annotation of Slovene, `http://nl.ijs.si/jos/`), devoted to corpus annotation has provided the means to revise the specifications, and use them as the basis to (semi)manually annotate two corpora of Slovene [10]. The development of these "JOS" specifications, has, to a large extent, also served as the testing ground for the new MULTEXT-East specifications. In Version 4 we plan to incorporate the JOS specifications into MULTEXT-East.

The rest of this paper is structured as follows: Section 2 details the XML format of the specifications, Section 3 discusses the associated XSLT stylesheets, Section 4 briefly introduces the MULTEXT-East lexica and annotated corpus, and Section 5 gives some conclusions and directions for further work.

Language name	Language family	Added in
English	Germanic	Version 1
Romanian	Romance	Version 1
Russian	East Slavic	Version 4
Czech	West Slavic	Version 1
Slovene	South West Slavic	Version 1/4
Resian	dialect of Slovene	Version 3
Croatian	South West Slavic	Version 3
Serbian	South West Slavic	Version 2
Macedonian	South East Slavic	Version 4
Bulgarian	South East Slavic	Version 1
Persian	Indo-Iranian	Version 4
Estonian	Finno-Ugric	Version 1
Hungarian	Finno-Ugric	Version 1

Table 1. Languages covered by the morphosyntactic specifications.

2 The format of the specifications in V4

In this section we give some background in the area of standardisation of multilingual morphosyntactic specifications, and detail their structure and encoding for MULTEXT-East Version 4.

The concepts expressed in MULTEXT-East specifications go back to the EAGLES guidelines from the early '90. The EU project EAGLES, the Expert Advisory Group on Language Engineering Standards, was instrumental for advancing the field of standardisation of language resources in a multilingual setting, and tackled corpora, spoken resources, lexica etc. as well as morphosyntactic descriptions and their specifications [2, 6].

But while the EAGLES compared a large number of proposals and gave general recommendations for encoding morphosyntactic descriptions, it did not provide explicit common specifications for a set of languages which could be mapped into morphosyntactic descriptions as used in lexica and corpora. This did, however, happen in the EU MULTEXT project, where the format of the specifications was concretised [1] for six EU languages (Italian, German, Spanish, French, Dutch, and English). The complete morphosyntactic specifications of MULTEXT were written as a LaTeX document, where the common tables are plain ASCII in a strictly defined format. The MULTEXT proposal also divided the features it defined into "general" and language specific ones. The first are taken to be used by most MULTEXT languages, while the second were those that were felt to be needed to describe the specifics of particular languages and their pre-existing resources.

MULTEXT-East adopted the MULTEXT format, except that it re-defined the language particular features to accommodate the radically different, mainly inflectional properties of the MULTEXT-East languages, and substituted the MULTEXT languages with the MULTEXT-East ones. The two proposals thus cannot be trivially combined, as they share only a subset of the attributes.

The complete MULTEXT-East morphosyntactic specifications consist of the following parts:

1. introductory matter: preface, background, organisation of the proposal, bibliography
2. common part: attribute-value tables for each category with notes
3. language particular parts for each language

The MULTEXT specifications, in particular, the attribute-value tables of the common part, should be interpreted as defining feature-structures, a well-known linguistic representation formalism, where a feature-structure consists of a set of attribute-value pairs. The common tables thus correspond to the definition of attribute- value pairs (e.g., that there exists, for Nouns, an attribute `Type`, which can have the values `common` or `proper`), while an MSD corresponds to a fully-specified feature-structure. But in MULTEXT there was no automatic way (piece of software) provided for converting the MSDs to feature-structures or vice-versa, or for checking the consistency of the specifications. For this reason MULTEXT-East soon developed a (Perl) program, which could expand, on the basis of the common tables in the

specifications, MSDs into a plain text feature-structures or check the validity of an MSD for a given language.

Having the document formatted in LaTeX and the formal parts written as ASCII tables had the virtue of simplicity but was problematic for at least two reasons. As mentioned, ad hoc programs were needed to validate MSDs against the specifications, or to internally validate the specifications. As the years passed, it was also becoming increasingly difficult to add new languages in a controlled fashion, due to the brittleness of the plain text format, and to the inter-dependencies and redundancy between the tables. What was needed was a formal specification for the tables that would enable their validation, extension, rendering on the Web or paper, or conversions into other formats.

2.1 Using the TEI

The Text Encoding Initiative `http://www.tei-c.org/` is an international consortium, whose primary function is to maintain the TEI Guidelines, which set out a vocabulary of elements useful for describing text for scholarly purposes. The Guidelines use XML encoding and are written as a set of XML schemas (element grammars) with accompanying documentation. In MULTEXT-East V3 we used Version P4 for encoding of the corpora, while in V4 we use of the most recent published version, TEI P5 [20].

The are a number of advantages of using TEI for encoding. TEI documents are written in XML, which brings with it the possibility of validation of document structure, a wealth of supporting software and related standards. Of these, the most important is the XML transformation language, XSLT, which allows writing scripts (stylesheets) that transform XML documents into other, differently structured XML documents, or into HTML as well as, indirectly, into a printable version in, say, PDF. The XSLT standard is nowadays generally supported, e.g., we find it implemented in most Web browsers. The MULTEXT-East specifications come with a number of XSLT transforms, which help in authoring or displaying the specifications; they are further discussed in Section 3.

TEI is also general enough to encode the non-normative parts of the specifications, e.g., the introductions, notes, references, etc. The TEI also provides, amongst other software, a sophisticated set of XSLT stylesheets and associated components for converting TEI documents into HTML and PDF. These stylesheets, developed by Sebastian Rahtz and freely available via the TEI homepage, cover a large number of TEI elements, and also perform tasks such as generating the table of contents, splitting (large) TEI documents into several HTML files (while preserving cross-links), giving each HTML a project defined header and footer, etc.

Finally, the MULTEXT-East parallel and MSD annotated corpus was already encoded in TEI; by encoding the specifications in TEI as well, this gives an easy way to directly integrate the corpus with the specifications, leading to simple validation of the corpus annotations or conversion between corpus MSDs and their feature-structure representations. This can be extremely useful for querying the corpus, as it enables e.g., the selection of word tokens based on particular features.

For these reasons the V4 specifications are written in TEI P5, as one XML document (which does not mean they have to be in one file), with the idea that this is the single document which needs to be maintained and to which new languages are added in a controlled fashion. The structure should therefore be amenable to hand editing, minimally redundant, contain as much as possible of structured commentary and references, with the formal parts having a transparent structure.

2.2 The common part of the specifications

This section gives more detail about the structure of the common part of the specifications in TEI. The common part of the specifications contains:

1. A table giving all the languages of the specification. For each language the table also gives its language family, ISO 836 code, and a link to its description in the Ethnologue database.
2. A table giving the (part-of-speech) categories of MULTEXT-East (12) together with their one-letter codes. The derived HTML of the specifications (so called display version) additionally contains the number of attributes defined for each category and which languages distinguish them.

3. For each category, the common table, defining attributes and their values for the category. For attribute they also specify its position in the MSD string, and for each attribute-value pair, a one letter code for the MSD string. For each such pair, the table also lists the languages that the attribute-value is valid for.
4. A table of all defined attributes, with the categories they are defined for, and their position in the MSD string (in display version only, and automatically generated from the XML source).
5. A table of all defined values, with the attribute/categories they are defined for, their code in the MSD string, and the languages that distinguish this attribute-value pair (in display version only, and automatically generated from the XML source).

Figure 1 gives an example from the TEI source, while Figure 2 gives the display view; the latter is, on purpose, quite similar to the tables in MULTEXT-East V3. The master TEI is, however, more logically oriented: the first row defines the category and gives the languages it is appropriate for while the following rows each define an attribute, with the values given in a subordinate table.

2.3 The language particular specifications

The specifications contain, for each language, also a language particular part. These parts can have a minimal structure, just giving the authors and repeating the common tables, but reduced to the categories and attribute-value pairs that are in fact used by the language. They can also be quite complex and can contain some or all of the following divisions:

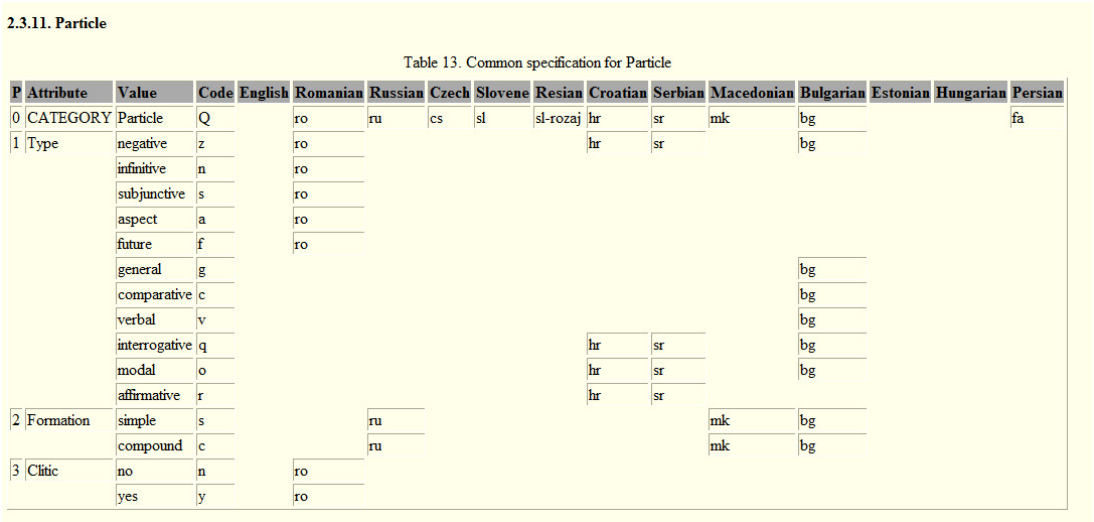
- Introductory matter, e.g., language description; background of the language specifications; bibliography.
- Then, for each category:
  - The language particular table, which can be automatically derived form the common table, but also modified from it, as will be further described below. Furthermore, the tables can also contain localisation information, i.e., the names of the categories, attributes, their values and codes in the particular language, in addition to English. This enables keeping the feature-structures and MSDs either in English, or in the language in question.
  - Notes on the category itself or on the attributes and values used.
  - Combinations of attribute-values (feature co-occurrence restrictions), which in a regular-expression-like syntax limit the possible combinations of attribute-values. These restrictions can also contain examples of usage. It should be noted that these combinations have not yet been operationalised, i.e., it is not possible to directly use them to validate MSDs.
  - A list of lexical MSDs, which should contain all the valid MSDs for the category. This is present only in the display view and automatically extracted from the full MSD index.
- The MSD index, which should contain all the valid MSDs for the language. Each MSD can be furthermore accompanied by explicatory information, i.e., its decomposition into feature-values, examples of usage, and its translation. This index is the authority for the MSD set for the language, and is valuable for MSD validation.

As an example of how a language particular table can look in Version 4, we give the JOS table for Slovene Nouns in Figure 3. The table gives identical information as the (Slovene selected) common tables, except that all information is also translated/localised to Slovene.

In MULTEXT and MULTEXT-East V3 the attribute-value definitions, together with MSD mapping information (i.e., the attribute position and the attribute-value code) were simply copied from the common tables. In MULTEXT-East V4 we take a more flexible position, where a language particular section can have a looser connection to the common tables – in fact, it could be a completely different specification, matching to the MULTEXT-East common one only in form. Of course, in this case any sensible mapping from the language particular specification to the common MULTEXT-East ones become very difficult, if not impossible. However, there do exist sensible compromises between the trivial mapping of MULTEXT and MULTEXT-East V3 and a completely unconstrained one.

```
<div type="section" xml:id="msd.Q">
  <head>Particle</head>
  <table n="msd.cat" xml:id="msd.cat.Q">
    <head>Common specification for Particle</head>
    <row role="type">
      <cell role="position">0</cell>
      <cell role="name">CATEGORY</cell>
      <cell role="value">Particle</cell>
      <cell role="code">Q</cell>
      <cell role="lang">ro</cell>
      <cell role="lang">sl</cell>
      <cell role="lang">cs</cell>
      ...
    </row>
    <row role="attribute">
      <cell role="position">1</cell>
      <cell role="name">Type</cell>
      <cell>
        <table>
          <row role="value">
            <cell role="name">negative</cell>
            <cell role="code">z</cell>
            <cell role="lang">ro</cell>
            <cell role="lang">bg</cell>
            <cell role="lang">hr</cell>
            <cell role="lang">sr</cell>
          </row>
          <row role="value">
            <cell role="name">infinitive</cell>
            <cell role="code">n</cell>
            <cell role="lang">ro</cell>
          </row>
          <row role="value">
            <cell role="name">subjunctive</cell>
            <cell role="code">s</cell>
            <cell role="lang">ro</cell>
          </row>
          ...
        </table>
      </cell>
    </row>
    <row role="attribute">
      <cell role="position">2</cell>
      <cell role="name">Formation</cell>
      ...
    </row>
    ...
  </table>
  ...
</div>
```

Fig. 1. Example of a MULTEXT-East common table: start of definition for Particle.



2.3.11. Particle

Table 13. Common specification for Particle

P	Attribute	Value	Code	English	Romanian	Russian	Czech	Slovene	Resian	Croatian	Serbian	Macedonian	Bulgarian	Estonian	Hungarian	Persian
0	CATEGORY	Particle	Q		ro	ru	cs	sl	sl-rozaj	hr	sr	mk	bg			fa
1	Type	negative	z		ro					hr	sr		bg			
		infinitive	n		ro											
		subjunctive	s		ro											
		aspect	a		ro											
		future	f		ro											
		general	g										bg			
		comparative	c										bg			
		verbal	v										bg			
		interrogative	q							hr	sr		bg			
		modal	o							hr	sr		bg			
		affirmative	r							hr	sr					
2	Formation	simple	s			ru						mk	bg			
		compound	c			ru						mk	bg			
3	Clitic	no	n		ro											
		yes	y		ro											

Fig. 2. Example of a common tabels in HTML: Particle.

The one we plan to adopt for the Slovene specification in Version 4 is exemplified by the JOS specification, where the tables will be aligned to the MULTEXT-East common ones in all respects, except for the attribute positions. This means that the feature-structure set of both will be identical, but not the MSDs. The reason for this is that MULTEXT-East has to cater for attributes of all languages, so language specific attributes (or those added to the specifications at a later date) wind up at the end of the string, leading to unwieldy MSDs, such as `Gppspe--n-----d`. This MSD has a number of hyphens only in order to maintain the position mapping to features, even though the attributes for some of these positions are never used for Slovene. With the freedom to reorder attributes, an individual language can use much shorter and more intuitive MSDs.

3 XSLT stylesheets

An important part of the specifications are the associated XSLT stylesheets, which allow for various transformations over the specifications. The stylesheets are written in XLST V1.0 and documented with XSLTdoc, `http://www.pnp-software.com/XSLTdoc/`. They take the specifications as input, usually together with certain command line arguments, and produce either XML, HTML or text output, depending on the stylesheet.

We provide three classes of transformations, the first ones to help in adding a new language to the specifications themselves, the second to transform the specifications into HTML, and the third to transform or validate a list of MSDs.

3.1 Authoring

The two stylesheets belonging to this class are meant to assist in adding new languages to the specifications, and are the following:

msd-split.xsl makes a template for a language particular section on the basis of the value given to the `-langs` parameter, which should contain a space separated list of ISO language codes. So, to make section for a new language X, which is similar to Y and Z, the stylesheet would be run with -langs 'Y Z' and would produce a section with the union of the attribute-values for these two languages. These new language particular specifications are then corrected by hand.

```
<div type="section" xml:id="msd.N">
<head xml:lang="sl">Samostalnik</head>
<head xml:lang="en">Noun</head>
<table n="msd.cat" xml:id="msd.cat.N">
<head xml:lang="sl">Tabela atributov in vrednosti za samostalnik</head>
<head xml:lang="en">Attribute-Value Table for Noun</head>
<row role="type">
<cell role="position">0</cell>
<cell role="name" xml:lang="sl">samostalnik</cell>
<cell role="code" xml:lang="sl">S</cell>
<cell role="name" xml:lang="en">Noun</cell>
<cell role="code" xml:lang="en">N</cell>
</row>
<row role="attribute">
<cell role="position">1</cell>
<cell role="name" xml:lang="sl">vrsta</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="sl">občno_ime</cell>
<cell role="code" xml:lang="sl">o</cell>
<cell role="name" xml:lang="en">common</cell>
<cell role="code" xml:lang="en">c</cell>
</row>
<row role="value">
<cell role="name" xml:lang="sl">lastno_ime</cell>
<cell role="code" xml:lang="sl">l</cell>
<cell role="name" xml:lang="en">proper</cell>
<cell role="code" xml:lang="en">p</cell>
</row>
</table>
</cell>
</row>
```

Fig. 3. JOS morphosyntactic specifications: start of table for Noun.

msd-merge.xsl takes a language particular specification, and tries to "insert" it into the common specifications. This can mean simply adding the new language flags to existing attribute-value pairs, or adding new values or even new attributes to the common specifications.

3.2 Rendering

Displaying the stylesheets is currently only supported in HTML. This is done in two stages:

msd-spec2prn.xsl generates a "display-oriented" TEI document from the specifications. This means making display-oriented tables and generating the indexes of attributes, values, and MSDs.

msd-prn2html.xsl is a driver file, which calls the standard TEI stylesheets. It takes as input the display-oriented document and produces the HTML equivalent.

3.3 MSD conversion

The stylesheets in this class take a list of MSDs as a parameter, and, on the basis of the given specifications typically convert them to some form of feature-structures. The specifications can be either the MULTEXT-East common ones, or those for a particular language, depending on whether the MSDs are the common or language particular ones.

msd-expand.xsl produces different types of output, depending on the values of its "mode" parameter. It also takes parameters for input language (only MSDs valid for the language will be accepted) and for output language (it can localised to a language, which, of course, must be supported by the specifications). The output is in plain text tabular format, with columns that can be, depending on the value of mode, which is a space separated list of modes, the following:

check only checks the validity of the input MSDs, flagging codes that are illegal for the language – this mode does not combine with the other ones;

id identity transform (with possible localisation);

collate collating sequence, with which it is possible to sort MSDs so that their order corresponds to the ordering of categories, attributes and their values in the specifications;

brief expansion to values only, which the is the most compact feature-expanded format and is meant for short but still readable expansions of MSD; instead of binary values (yes/no), +/− Attribute is written;

verbose expansion to feature-structures (attribute=value pairs) for all attributes defined for the category of the MSD;

canonical expansion to feature-structures (attribute=value pairs) for all defined attributes, regardless of whether they are defined for a particular category or not;

msd-fslib.xsl transforms the MSD list into a XML/TEI feature and feature-structure libraries, suitable for inclusion into MSD annotated and TEI encoded corpora.

The intention isn't to run the above stylesheet whenever a transformation is needed but rather to run them, once the specifications are finished, over the complete set of MSDs to produce the tabular and XML files, which are then made available together with the specifications. To enable simpler processing and to produce output files with useful combinations of expansions, an additional Perl wrapper script is made available with the specifications.

4 Associated resources

Even though this paper is devoted to the morphosyntactic specifications, we also mention associated MULTEXT-East morphosyntactic resources, as without them, the specifications are not of much use. In the first instance this means the MULTEXT-East morphosyntactic lexicons, as it the lexicons that should provide the complete set of MSDs for a language, as well as examples of their usage. A second level resource are MSD annotated corpora, as this grounds the lexicon in contextualised examples of usage.

4.1 MULTEXT-East Lexicons

The MULTEXT-East morphosyntactic lexicons have a simple structure, where each lexical entry is composed of three fields: (1) the *word-form*, which is the inflected form of the word, as it appears in the text, modulo sentence-initial capitalisation; (2) the *lemma*, which is the base-form of the word; where the entry is itself the base-form, the lemma is typically given as the equal sign; and (3) the *MSD*, i.e., the morphosyntactic description, which should be 1) valid according to the specifications and 2) contained in the set of MSDs listed in the lexical list of the language particular sections. It should be noted that this second criterion is to an extent circular, as it will be the lexicon that ultimately determines the list of valid MSDs; in practice, the process of constructing the MSD list and lexicon therefore typically proceeds in a cyclic fashion. Optionally, the lexicon can contain also contain (4) a column, giving the frequencies of the lexical entries in a corpus – for this, a MSD tagged and lemmatised corpus of the language must of course be available. Figure 4 gives some example entries from the Slovene lexicon.

It is usually not the case that MULTEXT-East lexicons are produced from scratch but rather converted from some existing morphosyntactic lexica for a language. The MULTEXT-East lexica up to Version 3 were constructed according to different principles, but an ideal lexicon obeys the following principles:

```
alibi        =         Ncmsn
alibi        alibi     Ncmsa--n
alibija      alibi     Ncmda
alibija      alibi     Ncmdn
alibija      alibi     Ncmsg
alibije      alibi     Ncmpa
alibijem     alibi     Ncmpd
alibijem     alibi     Ncmsi
alibijema    alibi     Ncmdd
alibijema    alibi     Ncmdi
```

Fig. 4. Example of a MULTEXT-East morphosyntactic lexicons: the start of the paradigm for the Slovene masculine nominal lemma "alibi".

1. The lexicon should contain all the valid MSDs for the language, even if only single exemplars are provided for particular MSDs. This criterion is in fact more strict than it seems, as languages with a large number of MSDs (e.g., Slovene has almost 2,000) exhibit a Zipfian distribution, i.e., quite a large number of MSDs can be quite rare in practice.
2. The lexicon should, for the lemmas it contains, include their complete inflectional paradigms. This is not always possible, as certain languages (e.g., agglutinating ones) can have "paradigms" with over a million word-forms but is manageable for even highly inflecting languages. The advantage is including the complete paradigms is that this makes the lexicon a very good resource for machine learning of lemmatisers; additionally, it also makes it more likely to obey the condition 1) above.
3. The lexicons should be of reasonable size (most current MULTEXT-East have around 15,000 lemmas), and, of course, the larger, the better. Ideally, the lemmas appearing in the lexicon should be grounded in an annotated corpus of the language, and the entries accompanied by corpus frequencies.

We do not here attempt to tackle the difficult problem of conversion of existing lexica to MULTEXT-East ones, but it should be noted that the `mtems-expand.xsl` in its `check` mode can be of considerable help in validating the lexical MSDs.

4.2 Annotated corpus

A corpus, annotated with context disambiguated MSDs and lemmas, provides the final piece of the "morpho-syntactic triad", as it contextually validates the specifications and lexicon, and provides examples of actual usage of the MSDs and lexical items.

Corpora currently included in MULTEXT-East deliverables are all (translations of) the novel "1984" by G. Orwell. The complete novel has about 100.000 tokens, although this of course differs between the languages. The corpus is annotated with MSDs and lemmas, which makes it suitable for MSD tagging and lemmatisation experiments. Because it was the first such resource for many of the languages involved the annotation had to proceed mostly manually. The corpus is, in Version 3, encoded in XML, according to the Text Encoding Initiative Guidelines P4 [19], but it is planned to upgrade it to TEI P5 in Version 4. To exemplify the current structure, Figure 5 gives the start of the Slovene part of the corpus.

This parallel corpus also comes with separate alignment files, which contain, in V3, hand-validated pair-wise sentence alignments (not necessarily 1-1) between English and the translations. For V4 we also plan to provide pair-wise alignments between all the languages, which have been automatically induced from the alignments with English.

5 Conclusions

The paper presented the morphosyntactic specifications that will be part of the MULTEXT-East resources Version 4. The specifications currently cover 13 languages, and are encoded in TEI P5, with dedicated XSLT scripts to help with authoring the specifications for new languages, convert them into feature-structures or into a display HTML encoding. As the specifications cover a number of languages for which not many available and standardised resources exist, they can be a valuable reference point, and, together

```
<text id="Osl." lang="sl">
 <body>
  <div type="part" id="Osl.1">
   <div type="chapter" id="Osl.1.2">
    <p id="Osl.1.2.2">
     <s id="Osl.1.2.2.1">
      <w lemma="biti" ana="Vcps-sma">Bil</w>
      <w lemma="biti" ana="Vcip3s--n">je</w>
      <w lemma="jasen" ana="Afpmsnn">jasen</w>
      <c>,</c>
      <w lemma="mrzel" ana="Afpmsnn">mrzel</w>
      <w lemma="aprilski" ana="Aopmsn">aprilski</w>
      <w lemma="dan" ana="Ncmsn">dan</w>
      ...
```

Fig. 5. Example of the annotation of the MULTEXT-East "1984" corpus: the start of the Slovene text "*Bil je jasen, mrzel aprilski dan*" (*It was a bright cold day in April*).

with the accompanying lexica and corpora, can serve as a "gold standard" dataset for language technology research and development, as well as for comparative linguistic studies.

There are a number of possible directions for further work. The language particular parts of the specifications could be further formalised and operationalised, esp. the combinations sections, as this would help in validating the MSD set for new languages. The attributes and their values could also be linked to other related attempts at standardisation of morphosyntactic features, in particular the ontology for descriptive linguistics GOLD `http://linguistics-ontology.org/gold.html` and the ISOcat Data Category Registry `http://www.isocat.org/`. There is also work to do in further formalisation of the MSDs and their relation to feature-structures, e.g., in allowing MSDs to include the metasymbols '*' or '.', i.e., having underspecified features in the MSD string.

Of course, we also hope that further languages will be added to the specifications. An obvious extension in this direction would be to add the original MULTEXT languages. However, we would encounter several problems: the specifications are incompatible outside the "common" features, so a way would needed to resolve this inconsistency, and in a backward compatible manner. More importantly, the associated resources, namely the lexicon and annotated corpus would have to be produced as well, to give the specifications some grounding in data. This is a relatively lengthily process, and it is unlikely that it could be carried out without dedicated international funding.

The situation is somewhat different, and better, for other, non Western European languages, where national efforts are underway to produce components of Basic Linguistic Resource Toolkits or BLARKs [16]; these can easily take the well-travelled route of developing MULTEXT-East compatible resources. Hopefully such an expansion could take place in the MONDILEX project, to include further Slavic languages into the specifications.

Finally, the most important aspect of the resources should be further encouraged, namely their use. Developing linguistic resources is not an end to itself, and they are worth only as much as they are used. We have therefore tried to maintain their quality and standardise their structure, to ensure that they can be interchanged and re-used for various purposes.

References

[1] Bel, N., Calzolari, N., and Monachini (eds.), M. (1995). Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets. MULTEXT Deliverable D1.6.1B, ILC, Pisa.

[2] Calzolari, N. and Monachini (eds.), M. (1996). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG—CLWG—MORPHSYN/R, ILC, Pisa. `http://www.ilc.cnr.it/EAGLES96/morphsyn/`.

[3] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevič, V., and Tufiş, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada. ACL.

[4] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., and Žele, A. (2006). Towards a Slovene Dependency Treebank. In *Fifth International Conference on Language Resources and Evaluation, LREC'06*, Paris. ELRA.

[5] EAGLES (1996). Expert Advisory Group on Language Engineering Standards. `http://www.ilc.pi.cnr.it/EAGLES/home.html`.

[6] EAGLES (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG–TCWG–MAC/R, ILC, Pisa. `http://www.ilc.cnr.it/EAGLES96/annotate/`.

[7] Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, pages 1535 – 1538, Paris. ELRA. `http://nl.ijs.si/et/Bib/LREC04/`.

[8] Erjavec, T. (2006). MULTEXT-East Morphosyntactic Specifications and XML. In Slavcheva, M., Simov, K., and Angelova, G., editors, *Readings in multilinguality*, pages 41–48. Bulgarian Academy of Science, Sofia.

[9] Erjavec, T. and Džeroski, S. (2004). Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

[10] Erjavec, T. and Krek, S. (2008). The JOS morphosyntactically tagged corpus of Slovene. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

[11] Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., and Vitas, D. (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32. ACL.

[12] Ide, N. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463–470, Granada. ELRA. `http://www.cs.vassar.edu/CES/`.

[13] Ide, N., Erjavec, T., and Tufiş, D. (2002). Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia. ACL.

[14] Ide, N. and Véronis, J. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 90–96, Kyoto. ACL.

[15] Ivanovska, A., Zdravkova, K., Erjavec, T., and Džeroski, S. (2006). Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns, Adjectives and Verbs. In *Proceedings of 5th Slovenian and 1st international Language Technologies Conference*, Jožef Stefan Institute, Ljubljana.

[16] Maegaard, B., Krauwer, S., Choukri, K., and Jorgensen, L. D. (2006). The BLARK concept and BLARK for Arabic. In *Fifth International Conference on Language Resources and Evaluation, LREC'06*, Paris. ELRA.

[17] QasemiZadeh, B. and Rahimi, S. (2006). Persian in MULTEXT-East Framework. In *FinTAL 2006: 5th International Conference on Natural Language Processing*, pages 541–551, Turku, Finland.

[18] Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and evaluating a Russian tagset. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris. ELRA.

[19] Sperberg-McQueen, C. M. and Burnard, L., editors (2002). *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium.

[20] TEI Consortium, editor (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.

[21] Tufiş, D. (1999). Tiered Tagging and Combined Language Model Classifiers. In Jelinek, F. and Noth, E., editors, *Text, Speech and Dialogue*, number 1692 in Lecture Notes in Artificial Intelligence, pages 28–33, Berlin. Springer-Verlag.

[22] Vojnovski, V., Džeroski, S., and Erjavec, T. (2005). Learning PoS Tagging from a Tagged Macedonian Text Corpus. In *Proceedings of the 8th International Conference Information Society, IS 2005*, Jožef Stefan Institute, Ljubljana.

Design of a New Slovak-Czech Lexical Database*

Radovan Garabík[1] and Jana Špirudová[2]

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[2] Institute of the Czech Language, Academy of Sciences of the Czech Republic, Prague

Abstract. We present an electronic Slovak-Czech lexical database, being compiled with the help of the MoinMoin wiki system. The lexical entry microstructure is organised into a tabular form and special plugins have been written to support easy compiling and editing of entries. Streamlined, traditional-like dictionary entries are then created of the data entered, with the aim to obtain create a printed dictionary.

1 Introduction

Czech and Slovak belong to the West Slavic languages. They have a lot of common in their morphology, phonology, lexicon and syntax. The languages are generally considered to be mutually intelligible.

After the break-up of Czechoslovakia in 1993, sociolinguistic connections between the languages started to weaken, and with the loss of perceptive bilingualism (predominantly on the side of Czech speakers), the mutual intelligibility is no longer universal, however, it is still sufficient for general communication. The situation is highly asymmetrical: while in Slovakia, Czech (both spoken and written) is ubiquitous in the TV, books and other media, in the Czech Republic, presence of Slovak language is rather rare[8]. Slovak speakers have nearly 100% understanding of all the varieties of Czech, but the Czech speakers (especially the younger ones) have sometimes troubles coping with Slovak, in particular with lexical items which are considerably different in the two languages. Consequently, a pressing need for general purpose dictionaries helping the Czech speakers in reading and understanding Slovak texts has emerged.

Ideally, we would like one single dataset to be used to construct all the possible dictionaries, and even a database to be used in all sorts of NLP (e.g. machine translation). This puts additional, often conflicting requirements on the design and building process of the lexical database, and therefore some compromises need to be made.

The primary design goals of the dictionaries to be obtained are:

- to be primarily a passive readers' dictionaries
- to be general purpose, "traditional" middle sized (cca. 20–30 thousand entries) dictionaries, with good coverage of different expressions and false friends
- to contain information on levels of usage

From this it follows that the lexical database has to meet the following requirements:

- to be a web based database with queries performed not just by lemmata, but also by varying wordforms
- to include links into various entry related information (such as morphology paradigm)
- to enable easy, online updating and editing by multiple editors

The last two points can be easily met by a wiki based software. We decided to use the MoinMoin wiki engine, because it supports custom page parsers and plugins that can be tailored to the needs of an online lexical database. On the other hand, MoinMoin full-text search is not really scalable – it is a problem especially concerning the *Category* pages, which internally use the full-text search mechanism. Therefore we refrained from using category pages in the database design.

* 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. The lexical database project has received support from the National Scholarship Programme of the Slovak Republic for the Support of Mobility of Students, PhD. Students, University Teachers and Researchers.

2 Basic structure of the database

Basic building block of the database is an entry, which we call a *page*[3]. A page is used to cover information pertaining to strictly one word meaning, information about homonyms is delegated to the overlying database structure. Each page is uniquely identified by its name, which by convention corresponds to the lemma, or, in case of homonymy, the page name consists of a lemma and a disambiguation identifier (Roman or Arabic numeral).

3 Lexical entry microstructure

Each page (database entry) is kept in a tabular form, where each item (row) has a predefined form and/or content. As an aid for the editors, fields that contain primary linguistic information have a language flag that indicates the language of that field (i.e. either *sk* or *cs*).

3.1 Paradigm (sk)

The *paradigm* field contains an identification of lemma's inflectional paradigm, as used in the morphology database[6]. Since the morphology is also stored in a MoinMoin wiki, the identifier is formatted and displayed as an interwiki link, to allow easy one-click access to the complete word morphology. Since all the word forms are available, the entries do not contain any other inflectional information (traditionally, Czech and Slovak dictionaries contain genitive singular and nominative plural suffixes for nouns, or the 3rd person singular and plural indicative forms for verbs). Similarly, since the paradigm contains a complete morphosyntactic specification including a part of speech category, we do not need to indicate the part of speech separately in the database.

3.2 Translation (cs)

The *translation* field contains direct Czech translation of the Slovak word (or of its particular meaning). We choose the best Czech equivalent. In case there are two or more equally good possibilities, we introduce them all, separated by a semicolon (;). We also take into account etymological relation between the words, and use preferably etymologically related translation[4].

In case there is no direct or indirect Czech equivalent of the Slovak word (e.g., *pahreba*), this field should contain a description of the semantic content.

3.3 Number specification (sk)

This field contains the classification of typical or prevalent number or gender characteristics of the word (for nouns). Possible values are:

- usually plural
- usually masculine or feminine
- masculine or feminine
- feminine or neuter
- feminine, usually plural
- masculine, usually plural
- neuter, usually plural
- exclusively plural
- exclusively singular

[3] Using MoinMoin terminology.

[4] For example, we translate the Slovak word *jazykoveda* by the Czech *jazykověda*, even if we can also translate it by Czech *lingvistika*, and we translate the Slovak word *lingvistika* as *lingvistika*, even if the Czech *jazykověda* would be an equally good translation.

3.4 Qualifier (sk)

This field contains a terminological and/or style qualifier(s), or a special keyword denoting a phrase. The qualifiers are taken out of a fixed set of abbreviated words. When editing this field, the lexicographer is provided with a checkbox entry for each of the qualifiers.

3.5 Gloss 1 & 2

Gloss 1 narrows down the semantics – shade of meaning of the entry word or its semantic and functional equivalent. *Gloss 2* comments on the typical usage of the word.

3.6 Exemplification

The *exemplification* is not a single field, but consists of a variable number of Slovak-Czech exemplification pairs. The Slovak exemplification is primary, the Czech exemplification should be an appropriate translation of the Slovak one. The table displays all the non-empty exemplifications, plus an empty input field for the last Slovak one (to enable the editor to add another exemplification pairs).

3.7 Note

The *note* contains assorted notes for the dictionary user, relevant to the entry. By convention, we use a magic word *viz*[5] to denote a reference to another entry (such as a close synonym, an antonym, comments on significant style characteristics of the Czech equivalents or other related word).

3.8 False friends

This field contains a list of false friends, separated by a semicolon. We do not distinguish between variants of false friends (originating in Slovak or Czech, with a similar meaning, with a completely different meaning...)

3.9 Comment

This field is intended for any other comments by the editors – as such, it will not be displayed in the final entry form.

4 Sense disambiguation mesostructure

There is (intentionally) no place in the entry microstructure to be filled in with hints concerning homonymy disambiguation. We opted to encode this information into the overlaying database nomenclature of entries instead, following to some extent the usual lexicographic classification. At the lowest level, an entry is identified by its headword (MoinMoin page name), which – as its first function – directly encodes the lexeme's lemma. If there are two or more closely related, functionally and pragmatically identical word variants (e.g. spelling variations, such as *mliekar; mliekár*), a headword can contain more variants, separated by a semicolon (;) as a convenient shortcut. This should be thought of as a shorthand for database compilers, nothing more – functionally, such an entry is equivalent to describing both (or more) variants in full.

A headword can have a trailing uppercase Roman numeral, separated by a space. This is used to mark off major homonyms (or even homographs – such as part of speech homonymy, or a completely – even etymologically – unrelated meaning).

An entry can be created as a subpage of an already existing entry, by using MoinMoin's mechanism for subpages. A subpage *XX* of a page *YY* is an ordinary page, with a special name written as *YY/XX* (i.e. the

[5] Czech for *cf.*

subpage name follows the main page, separated by a slash). Subpages of a given page are logically clumped together, in the formatted entry output they are displayed nested with the primary page. We use subpages to connect diminutives, augmentatives and phrasal units to the principal word. Although MoinMoin allows for the whole hierarchy of subpages, we use only the first level subpages in our dataabse (with the exception of sense disambiguation, as outlined the following paragraph).

A headword can have a trailing slash and an Arabic numeral. While technically a subpage, this is used as a weaker variant of a Roman numeral disambiguation in cases, where the words are related and the meaning does not diverge that much. A Roman numeral major disambiguation can be combined with an Arabic numeral minor one (e.g. *čap I/1* – a pivot, journal (mechanical device), *čap I/2* – a hinge, *čap II/1* – a splash, *čap II/2* – a catch (act of catching)).

A headword can contain parenthesized reflexive pronouns *(sa)*, *(si)*[6]. This is used with those cases which are either very frequent, or where the reflexive form diverges in its meaning from the non-reflexive one.

Also, this is used with words which do not have straight one-to-one Czech equivalent, in case the presence of the reflexive does not change the basic meaning and usage of the word (e.g. *dopukat' (sa)* – to crack (about skin)).

5 Technical implementation

The dictionary has been pre-filled with a bilingual glossary of about 60 thousand word pairs[7] and with links into the morphology analyzer wiki, in order to ease the initial editing and to enhance the usefulness of the database by offering at least the first-guess translation and morphology paradigm of the words that would not get into the "core".

A page is internally stored as a flat plain text file (see Fig. 2), with each line corresponding to one table row, with the field name followed by a colon (:), followed by a field value (which can be empty). We have written a special MoinMoin formatter plugin that displays the table in a human-friendly way, together with a final, streamlined formatted entry (Fig. 1). We have also written a MoinMoin action that is used to edit just one specific table row. The action code has hardwired fields that can contain only a fixed set of values (number specification and qualifier) and provides the editor with checkboxes for all the possible values.

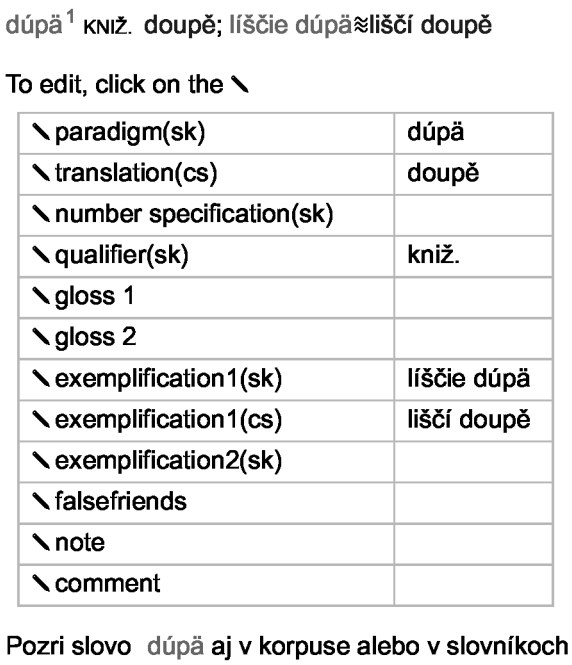
6 Formatted entry output

The tabular format of the dictionary entries displays the information in a clear and obvious way, however it is quite unsuitable for the intended published (paper) dictionary, and there is also the need to present the information in a more compact, concise form also for the internet-based version. Therefore the table is parsed and formatted into a traditionally looking entry.

7 Licensing issues

From the very beginning, we intended to publish the online dictionary entries under an open source/documentation license, in order to facilitate linguistic research and use of data in various NLP applications. The database is publicly accessible and editable under a triple license, GNU Free documentation license v. 1.2 [5] and Creative commons Contribution-Share alike (CC-BY-SA) license v. 3.0 [3] for the use in text document, and under Affero GNU Public license v. 3 [4] for use in computer programs (where by *linking* as specified in the license text we understand any use of the dictionary data by a computer program). This licensing concernes individual entries, while both our institutes keep special rights as a database compiler [1, 2] for the whole dictionary.

[6] Note that *sa* can be added to almost any transitive Slovak (and as *se* to a Czech) verb to express reflexivity, and *si* can be added to almost any verb.



dúpä [1] KNIŽ. doupě; líščie dúpä≈liščí doupě

To edit, click on the ↖

↖ paradigm(sk)	dúpä
↖ translation(cs)	doupě
↖ number specification(sk)	
↖ qualifier(sk)	knuž.
↖ gloss 1	
↖ gloss 2	
↖ exemplification1(sk)	líščie dúpä
↖ exemplification1(cs)	liščí doupě
↖ exemplification2(sk)	
↖ falsefriends	
↖ note	
↖ comment	

Pozri slovo dúpä aj v korpuse alebo v slovníkoch

Fig. 1. An example of a dictionary entry. Final, formatted output is displayed at the top.

```
paradigm (sk): dúpä
translation (cs): doupě
number specification (sk):
qualifier (sk): kniž.
gloss 1:
gloss 2:
exemplification1 (sk): líščie dúpä
exemplification1 (cs): liščí doupě
exemplification2 (sk):
exemplification2 (cs):
exemplification3 (sk):
exemplification3 (cs):
exemplification4 (sk):
exemplification4 (cs):
exemplification5 (sk):
exemplification5 (cs):
false friends:
note:
comment:
```

Fig. 2. Internal representation of a dictionary entry.

References

[1] §72 – §77, 618/2003 Z. z. *Zákon o autorskom práve a právach súvisiacich s autorským právom (autorský zákon), v znení neskorších predpisov*. Copyright Law of the Slovak Republic.

[2] §88 – §94, 121/2000 Sb. *Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů*. Copyright Law of the Czech Republic.

[3] Creative Commons (2009). Creative Commons Attribution-Share Alike 3.0 Unported. `http://creativecommons.org/licenses/by-sa/3.0/`. [Online; accessed 9 March 2009].

[4] Free Software Foundation (2007). GNU Affero General Public License. `http://www.gnu.org/licenses/agpl.html`. [Online; accessed 9 March 2009].

[5] Free Software Foundation (2008). GNU Free Documentation License. `http://www.gnu.org/licenses/gfdl.html`. [Online; accessed 9 March 2009].

[6] Garabík, R. (2008). Storing morphology information in a wiki. In *Lexicographic Tools and Techniques. MONDILEX First Open Workshhop. Proceedings*, pages 55–59, Moscow, Russia. IITP RAS.

[7] Hajič, J., Kuboň, V., and Hric, J. (2000). Machine Translation of Very Close Languages. In *6th ANLP Conference / 1st NAACL Meeting. Proceedings*, pages 7–12. Seattle, Washington.

[8] Nábělková, M. (2007). Closely Related Languages in Contact: Czech, Slovak, "Czechoslovak". *Small and Large Slavic Languages in Contact. International Journal of the Sociology of Language*, (183):53–73.

Experience with Building Slovak Electronic Lexical Database

Ján Genči

Technical university of Košice

Abstract. The paper presents author's experience in the design and development of Slovak electronic lexical database. It presents first attempts starting at the end of 80's, followed by processing of both on-line and printed data. Actual conceptual model of electronic database and recommendation regarding use of XML technology is presented also.

1 Introduction

Author's first touch with Slovak lexicon was accomplished in the late 80's during development of Slovak spellchecker (commercialized by Forma, s.r.o). Based on the acquired experience, our experiments with the "grammar checker algorithms" started in the middle of the 90's. In that time we understood that extended electronic version of Slovak lexicon database is necessary. Since then we carried out several attempts to achieve the goal. However, all of the attempts were based on students' semestral and/or diploma works without proper financial support. Moreover, it was clear from the beginning, that we require solution which would store every form of the word with corresponding morphological information.

2 Sources of data

The first source of data for our morphological database became our spellchecker database. Because of the lack of electronic data sources at the time of its development (end of 80's) it was based on the available edition of Príručka slovenského pravopisu pre školy [1]. All work was done without linguistic background and without any connection to linguists. To build the database, we proceeded with the following steps:

- typing of particular word;
- categorization of words by word classes;
- categorization of word classes according to defined paradigms (patterns);
- generating of all forms of words;
- building the database.

Having the most of related forms of the words we decided to develop application which for the given word would provide relevant word classes and their forms [2] (Fig. 1a and 1b). However, it was regarded as a proof of the concept only and never has been released for public use.

With the advent of internet the idea of using it as a source of words for building lexical database (including all forms) was explored. We discovered (what is clear to any linguist) that data in publicly available text is very noisy and frequencies of various forms of a word are very unequally distributed. Noisiness of the data we decided to reduce by narrowing our interest to newspaper and journal websites only.

Requirement to build a "clean" database leads us to the idea to use lexicographic sources (dictionaries). We explored it in two ways:

- using on-line dictionaries;
- using printed edition of dictionaries.

Using on-line dictionaries (i.e. [3]) is quite straightforward. Just download the corresponding webpage, parse the HTML code and use the result (see Fig. 2a, 2b). Parsed data can be used for generation of all word cases (Fig. 4)

.

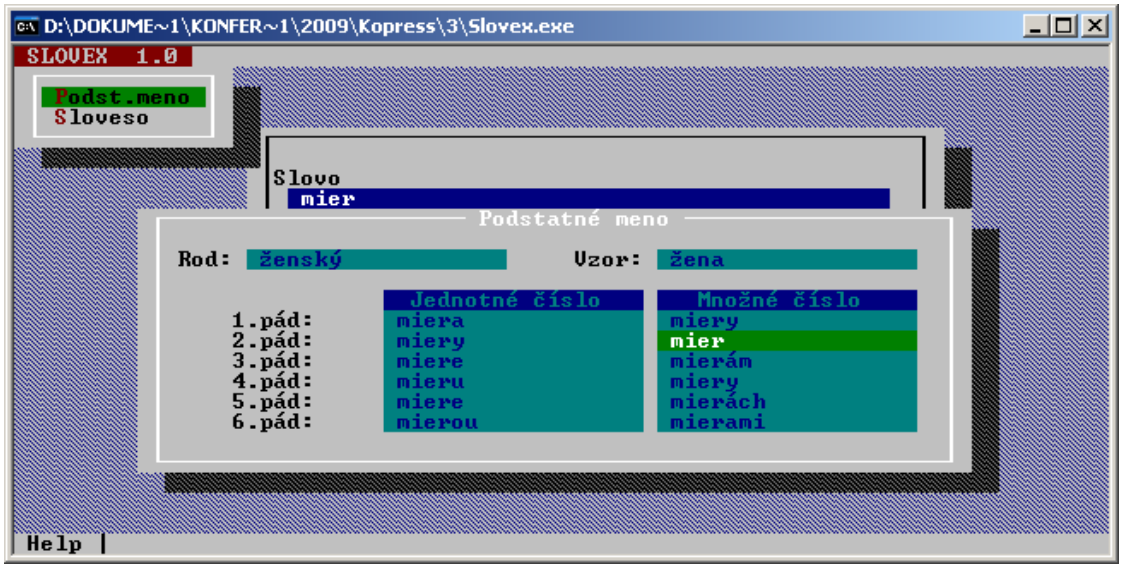


Fig. 1a – Word “mier” as the noun

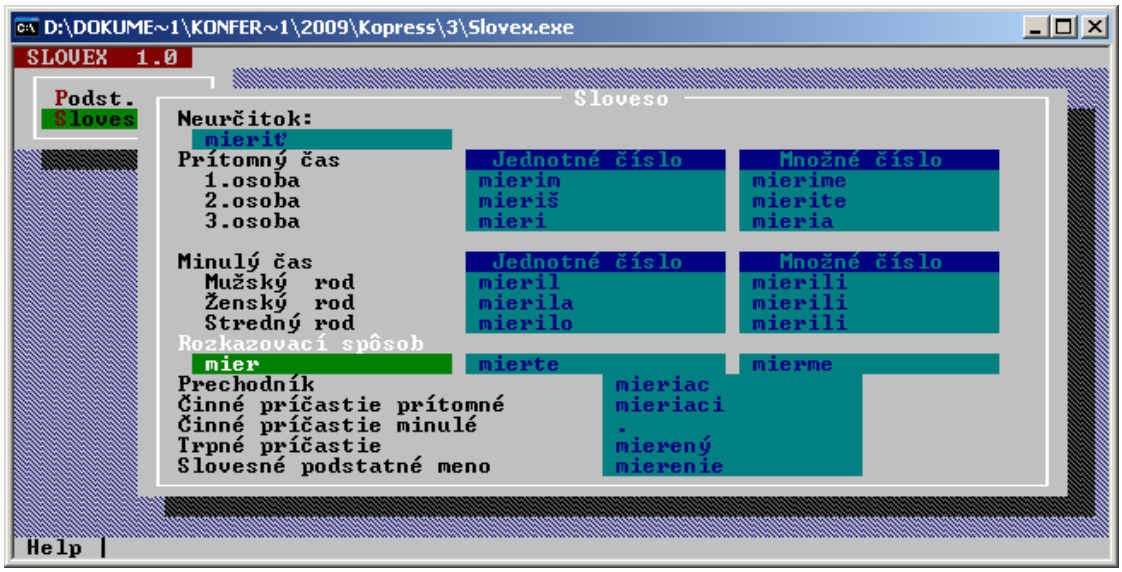


Fig. 1b – Word “mier” as the verb

Printed edition requires several more steps which often can be ambiguous. The first step is scanning of relevant pages, after that scanned pictures have to pass through OCR (Optical Character Recognition) process. The last step represents correction of recognition mistakes, which can be done manually (long lasting and error prone process) or automatically by computer program which tries to eliminate some common OCR mistakes (i.e. problem of recognition of “fi”,”fl” as two character strings, recognition errors which infringe alphabetical order of entries, changes of word root in the single entry etc.). After applying these steps we approximate the level of quality of data which can be compared to data acquired from on-line dictionaries.

However, to have (even excellent quality) data available, the further process is not always simple. The main issues are:

- data is produced by authors/editors without regard of further computer processing,
- limitations of printed edition have to be respected.

The first issue means that dictionary entry can contain hidden meanings which can be resolved only by human reader (i.e. based on data provided by Krátky slovník slovenského jazyka [4] (available at `http://slovnik.juls.savba.sk/`) it is impossible to determine algorithmically the pattern of the declination type "chlap" because there is no data allowing us to distinguish animate from inanimate objects – see Fig. 3. In some cases, this can be solved by exploring some additional resources.

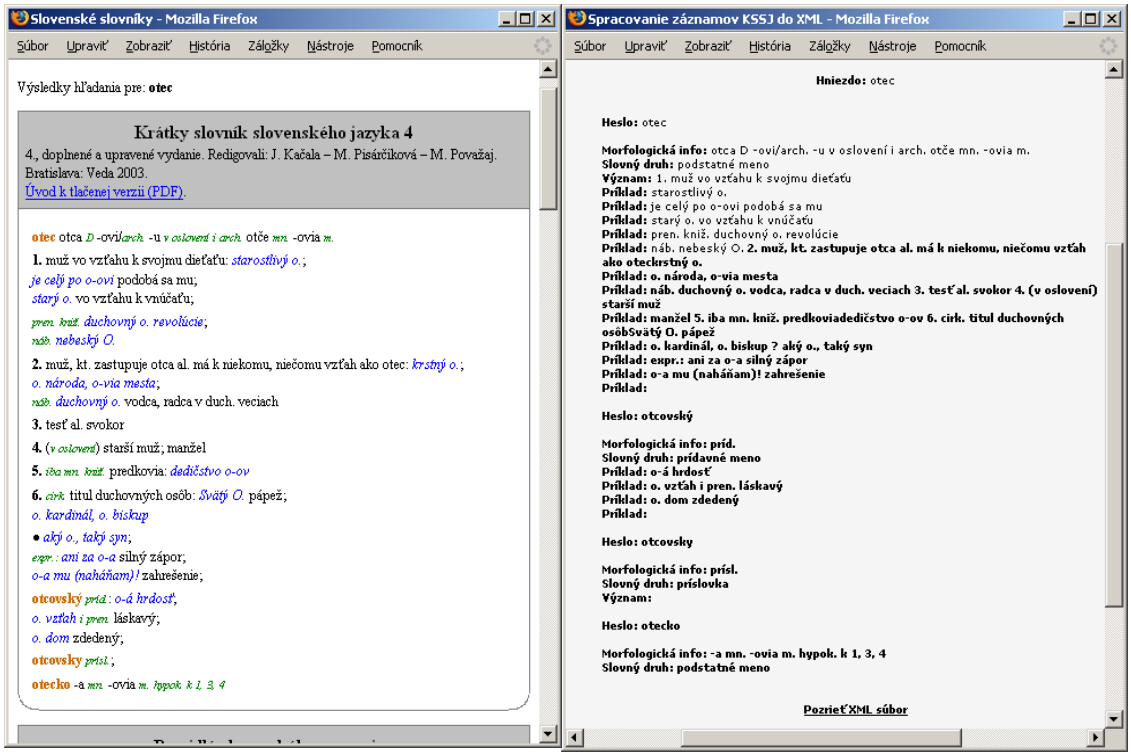


Fig. 2a – KSSJ entry for "otec"

Fig 2b – parsed data from KSSJ

Second issue means that some attributes are omitted intentionally because they either do not correspond to chosen dictionary entry or some relationships is not possible to present in two-dimensional structure of printed text.

Successfully resolving previous issues we are faced with last one – complex and irregular structure of dictionary entries. Today, computer programs use state of the art database technologies for fast access to data. It means - regular structure of data storage has to be designed and implemented. Just minor irregularity for a few entries (very often commented by words: "But it's just one entry") can require redesign of the whole structure of our database.

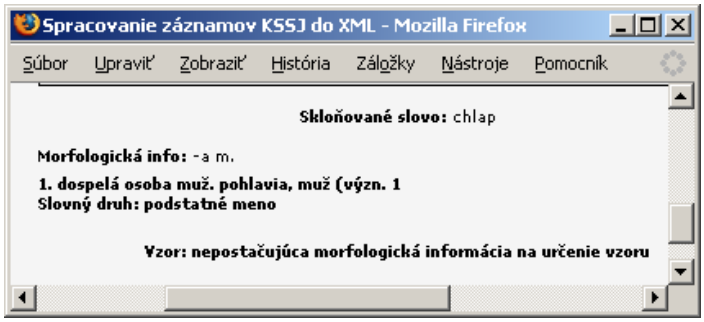


Fig. 3 – insufficient morphological data for word cases generation

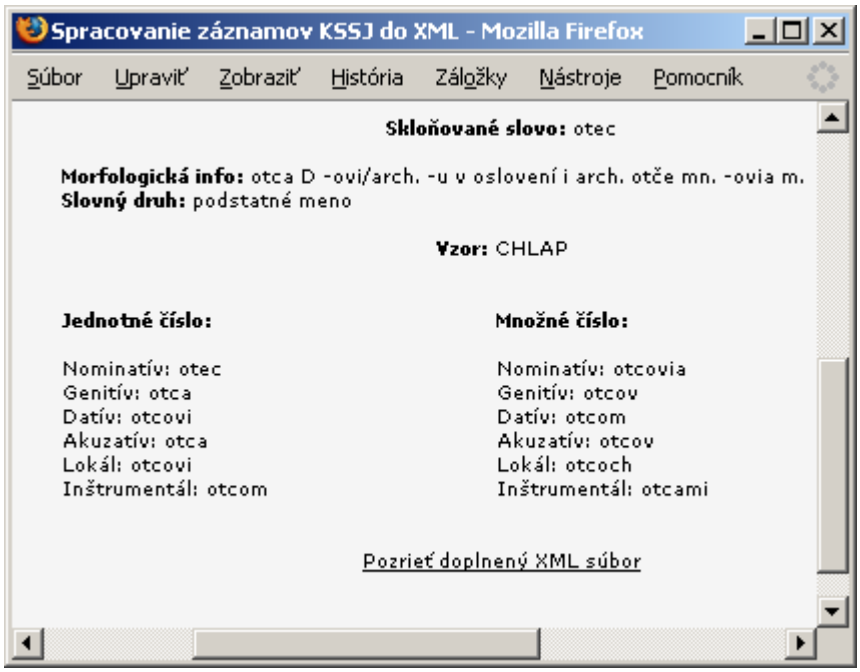


Fig. 4 – word cases generation based on morphological information

3 Conceptual model of morphological database

Conceptual model of morphological database is presented on the Fig. 6. All required data is centred around basic structure – lema, which represents basic form of the word. All forms of words are stored in the lexema table. Both lema and lexema has relationship to corresponding morphological information. We decided to store the data about meanings of the word represented by the set of synonyms (which we use for mining relevant foreign synonyms) and source(s) of each lema (URL or dictionary, where the word was/can be found).

To populate morphological database we use data generated by processes described in the previous section. We use XML as output of these processes (example is presented in the Fig. 5.)

```
<?xml version="1.0" encoding="WINDOWS-1250"?>
<hniezdo xmlns=""
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:SchemaLocation="http://lingua.cnl.tuke.sk/~soltysova/hniezdo.xsd">
            <nazov_hniezda>otec</nazov_hniezda>
            <heslo>
               <nazov_hesla>otec</nazov_hesla>
               <morf_info>
                        <slovny_druh>podstatné meno</slovny_druh>
                        <dalsie_info> otca D -ovi/arch. -u v oslovení i arch. otče mn. -ovia .</dalsie_info>
                        </morf_info>
                        <vyznam>
                             <vyklad> 1. muž vo vzťahu k svojmu dieťaťu</vyklad>
                            <priklad>starostlivý o.</priklad>
                            <priklad>je celý po o-ovi podobá sa mu</priklad>
               <priklad>starý o. vo vzťahu k vnúčaťu</priklad>
              <priklad>pren. knIž. duchovný o. revolúcie</priklad>
                            <priklad>náb. nebeský O.
</vyznam>
<vyznam>
<vyklad>2. muž, kt. zastupuje otca al. má k niekomu, niečomu vzťah ako otec</vyklad>
                <priklad>krstný o.</priklad>
                <priklad>o. národa, o-via mesta</priklad>
              <priklad>náb. duchovný o. vodca, radca v duch. veciach </priklad>
</vyznam>
<vyznam>
    <vyklad> 3. test' al. svokor </vyklad>
</vyznam>
<vyznam>
    <vyklad> 4. (v oslovení) starší muž </vyklad>
              <priklad> manžel </priklad>
</vyznam>
<vyznam>
     <vyklad> 5. iba mn. knIž. predkovia </vyklad>
     <priklad>dedičstvo o-ov</priklad>
</vyznam>
<vyznam>
    <vyklad> 6. cirk. titul duchovných osôb </vyklad>
    <priklad>Svätý O. pápež</priklad>
    <priklad>o. kardinál, o. biskup </priklad>
</vyznam>
            <sklonovacie_tvary>
            <singular>
                <nominativs>otec</nominativs>
                <genitivs>otca</genitivs>
                <dativs>otcovi</dativs>
                <akuzativs>otca</akuzativs>
                <lokals>otcovi</lokals>
                <instrumentals>otcom</instrumentals>
            </singular>
            <plural>
                <nominativp>otcovia</nominativp>
                <genitivp>otcov</genitivp>
                <dativp>otcom</dativp>
                <akuzativp>otcov</akuzativp>
                <lokalp>otcoch</lokalp>
                <instrumentalp>otcami</instrumentalp>
            </plural>
         </sklonovacie_tvary>
    </heslo>
```

Fig. 5 – XML output of processed dictionary entry

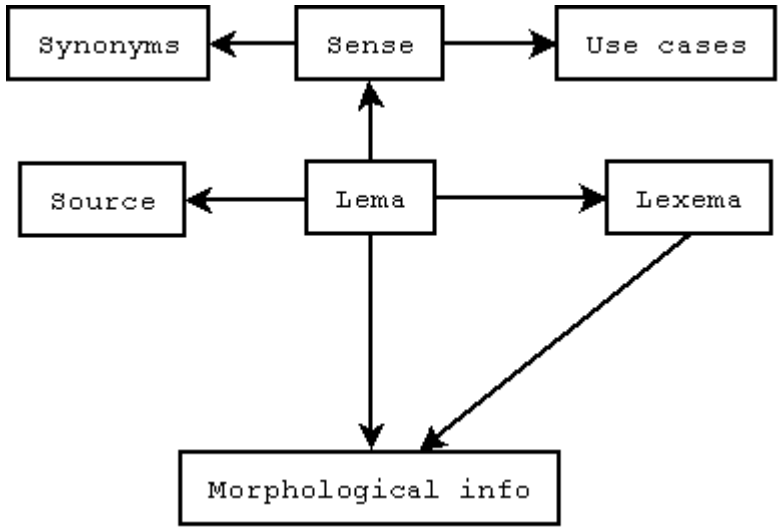


Fig. 6 – conceptual model of morphological database

4 Conclusion

The purpose of the paper is to illustrate processes regarding building electronic morphological database, highlights the drawbacks of them. We would like to persuade lexicographers to include IT specialists in the project teams and develop corresponding data structures (and respect it, of course). To be not bound by strict data structure provided by relational data model, we propose XML technology that could be used in these types of the projects. XML technology is mature enough to provide appropriate flexibility to cover needs in all dictionary entry structure variations. On the other side, other technologies provide tools for transformation of XML specification to the formats suitable for presentation of data on the WWW or in the printed form (PDF, text version).

References

[1] Oravec J., Laca V.: Príručka slovenského pravopisu pre školy. SPN, Bratislava, 1978.
[2] Soltysova A.: Spracovanie záznamov on-line verzie Krátkeho slovníka slovenského jazyka. Diplomová práca. KPI FEI technická univerzita v Košiciach. 2006
[3] Slovenské slovníky. Jazykovedný ústav Ľ. Štúra SAV. http://slovniky.juls.savba.sk/
[4] Krátky slovník slovenského jazyka. Bratislava: Veda 2003, available on-line at http://slovniky.juls.savba.sk/

Development of a Russian Tagged Corpus with Lexical and Functional Annotation*

Igor Boguslavsky, Leonid Iomdin, Tatyana Frolova, Svetlana Timoshenko

Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow
bogus@iitp.ru, iomdin@iitp.ru, frolova@iitp.ru, nyrestein@gmail.com

Abstract. A project aimed at creating a deeply tagged corpus of Russian texts with morphological, syntactic, lexical semantic and lexical functional annotation is presented.

1. Introductory Remarks

Tagged corpora are primarily intended for providing the basis for linguistic research in all fields of the vocabulary and the grammar (including changes occurring in the language throughout its history). There are two significantly different areas of such research. On the one hand, there are traditional linguistic studies for which mass material of texts is needed: such demand is much easier met if good and deeply tagged corpora are available. On the other hand, modern computational linguistics itself becomes an eager and interested user of such corpora as these are used on an increasing scale as training sets in machine learning. As a result of such learning, computer programs enhance their capability for extracting sophisticated types of data, which are contained in training text sets, from new texts. Generally speaking, the deeper the level of corpus annotation, the more advanced types of information could be learned from the corpus.

Recently, a new project has been started by the Laboratory of Computational Linguistics (LCL) of the Institute of Information Transmission Problems in Moscow, aimed at supplying SYNTAGRUS, the morphologically and syntactically tagged corpus of Russian texts, with lexical semantic and lexical functional annotation. The enhanced corpus will serve both areas of linguistic research: traditional and computational.

2. SYNTAGRUS Treebank

The Russian dependency treebank, SYNTAGRUS, developed and maintained by the LCL (Boguslavsky *et al*. 2002, Apresjan *et al.* 2005), currently contains about 40,000 sentences (roughly 520,000 words) belonging to texts from a variety of genres (contemporary fiction, popular science, newspaper, magazine and journal articles dated between 1960 and 2008, texts of online news, etc.) and is steadily growing. It is an integral but fully autonomous part of the Russian National Corpus developed in a nationwide research project and can be freely consulted on the Web[1].

Since Russian, as other Slavic languages, has a relatively free word order, SYNTAGRUS adopted a dependency-based annotation scheme, in many respects parallel to the Prague Dependency Treebank (Hajič *et al*., 2001).

So far, SYNTAGRUS is the only corpus of Russian supplied with comprehensive morphological and syntactic annotation. The latter is presented in the form of a full dependency tree provided for every sentence. In the dependency tree, nodes represent words annotated with parts of speech and morphological features, while arcs are labeled with syntactic dependency types. There are over 65 distinct dependency labels in the treebank, half of which are taken from Igor Mel'čuk's Meaning ⇔ Text Theory (see e.g. Mel'čuk, 1988).

* The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX. This study has also received partial funding from the Russian Foundation of Basic Research (grant No. 08-06-00373), which is gratefully acknowledged.

[1] See http://www.ruscorpora.ru/syntax-search.html.

Fig.1 below is a sample dependency structure for the sentence

Наибольшее	*возмущение*	*участников*	*митинга*	*вызвал*
$Most_{NEUT,SG,ACC}$	$indignation_{SG,ACC}$	$participant_{PL,GEN}$	$meeting_{SG,GEN}$	$cause_{PAST,PERF,,SG,MASC,GEN}$
продолжающийся	рост	цен	на	бензин,
$continue_{PART,PRES,,IMPERF,,SG,MASC,NOM}$	$growth_{SG,NOM}$	$price_{PL,GEN}$	on_{PREP}	$petrol_{SG,ACC}$
устанавливаемых	*нефтяными*	*компаниями*		
$set_{PART,PRES,,IMPERF,,PASS,PL,GEN}$	$oil\text{-}Adj_{PL,INSTR}$	$company_{PL,INSTR}$		

'It was the continuing growth of petrol prices set by oil companies that caused the greatest indignation of the participants of the meeting'.

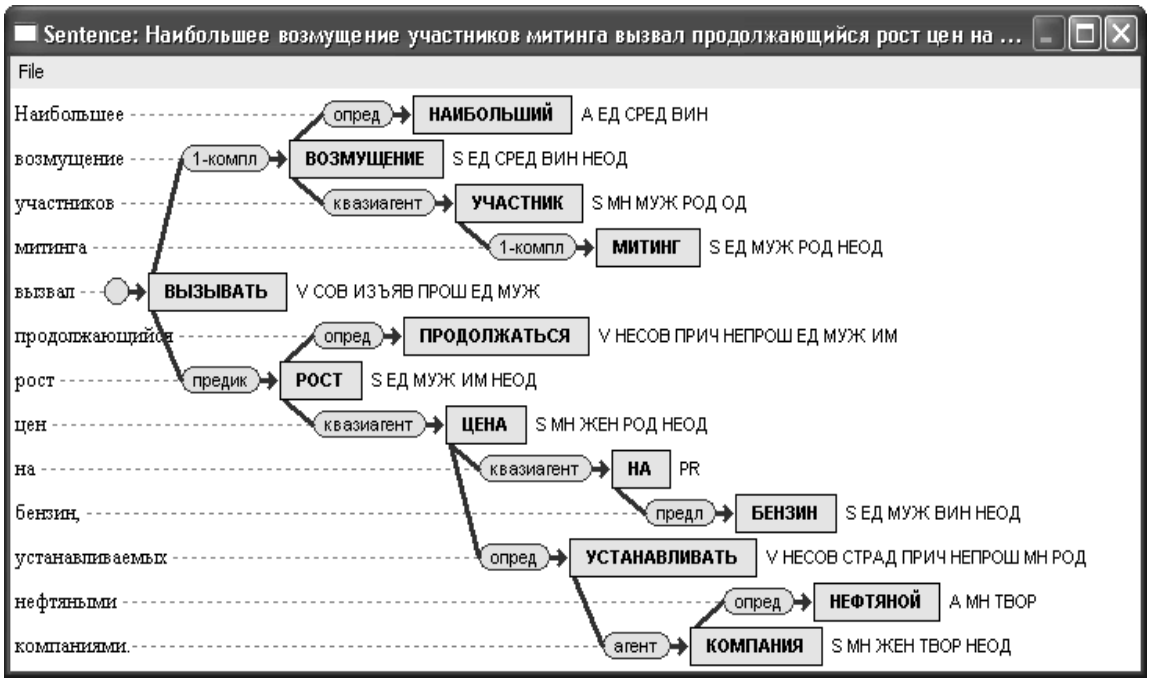


Fig.1. A syntactically tagged sentence.

Dependency types used in Fig. 1 include:

1. предик (predicative), which, prototypically, represents the relation between the verbal predicate as head and its subject as dependent;
2. 1-компл (first complement), which denotes the relation between a predicate word as head and its direct complement as dependent;
3. агент (agentive), which introduces the relation between a predicate word (verbal noun or verb in the passive voice) as head and its agent in the instrumental case as dependent;
4. квазиагент (quasi-agentive), which relates any predicate noun as head with the word implementing its first syntactic valency as dependent, if such a word is not eligible for being qualified as the noun's agent;
5. опред (modifier), which connects a noun head with an adjective/participle dependent if the latter serves as an adjectival modifier to the noun;
6. предл (prepositional), which accounts for the relation between a preposition as head and a noun as dependent.

Dependency trees in SYNTAGRUS may contain non-projective dependencies.
Normally, one token of the sentence (roughly, a word taken from space to space) corresponds to one node in the dependency tree. There are however a noticeable number of exceptions, the most important of which are the following:

1. compound words like *пятидесятиэтажный* 'fifty-storied', *стопятидесятипятимиллиметровый* 'one hundred fifty five millimeter wide', where one token corresponds to two or more nodes;

2. so-called phantom nodes for the representation of hard cases of ellipsis, which do not correspond to any particular token in the sentence; for example, *я купил рубашку, а он галстук* 'I bought a shirt and he a tie', which is expanded into *я купил рубашку, а он купил*$_{\text{PHANTOM}}$ *галстук* 'I bought a shirt and he bought$_{\text{PHANTOM}}$ a tie';

3. multiword expressions like во что бы то ни стало 'whatever happened', where several tokens correspond to one node.

Morphological and syntactic annotation for SynTagRus is performed semi-automatically: each sentence of the corpus is first processed by the rule-based Russian parser of an advanced multipurpose NLP system, ETAP-3 (Apresjan *et al*., 2003) and then edited manually by linguists, who correct errors made by the parser and handle cases of ambiguity that cannot be reliably resolved without extralinguistic knowledge.

Morphological annotation in SynTagRus is based on a comprehensive morphological dictionary of Russian that counts about 130,000 entries (over 4 million word forms). The ETAP-3 morphological analyzer uses the dictionary to produce morphological annotation of words belonging to the corpus, including the lemma, the part-of-speech tag and additional morphological features dependent on the part of speech: e.g. values of such features as 1) animacy, gender, number, case, degree of comparison, short form – for adjectives and participles, 2) representation (with values of finiteness, infinitive, participle, or gerund), aspect, tense, mood, person, voice – for verbs, etc. The morphological analyzer operates in a context-free manner, offering almost no morphological disambiguation for the sentence.

The syntactic parser processes morphologically analyzed sentences using a sophisticated set of syntactic rules, or syntagms, that produce one binary syntactic link each. Unlike many similar parsers, ETAP-3 uses no statistics-based prior part-of-speech tagging module.

When editing SynTagRus annotation, the developers use a powerful software tool, Structure Editor, which enables them to easily access all sorts of data necessary for efficient work (ETAP-3 dictionaries and rules) and handle even the hardest cases in a smooth and consistent way.

SynTagRus has already been used for a number of linguistic research and application tasks. In particular, it has been used as benchmark in regression tests designed to ensure stable performance of the ETAP-3 Russian parser in the course of its development (see e.g. Boguslavsky *et al*. 2008) and as a source for the creation, by machine learning methods, of a successful statistical parser for Russian (Nivre *et al*., 2008).

3. Lexical Semantic Annotation

Lexical semantic annotation means that, for all cases of word sense ambiguity of the corpus, the concrete lexical meaning should be identified and explicitly marked. In its present state, SynTagRus does not provide exact lexical meanings, showing only the lemmas of the words occurring in texts. This means that lexical ambiguity is only resolved in the corpus if ambiguous words happen to have different lemmas and/or different part of speech tags. Accordingly, SynTagRus distinguishes between *печь* as a verb ('bake') and *печь* as a noun ('oven') or the pronominal adjectives *сам* 'oneself" and *самый* 'very', so that

ambiguous sentences containing ambiguous word forms like *Она любит печь* 'She likes to bake' vs. 'She likes the oven', or *Я знаю самого главного инженера* 'I know the chief engineer himself' vs. 'I know the most important engineer' can be distinguished in the corpus. Conversely, if ambiguous lexemes (no matter whether they belong to one polysemic vocable or are lexical homonyms) have the same lemmas, they are not distinguished. For this reason, even very different words are underrepresented if they happen to have the same lemmas.

In the new corpus, all ambiguous lemmas will be supplied with concrete word senses as they are specified in the combinatorial dictionary of Russian. This dictionary is a vital component of the ETAP-3 linguistic processor that counts almost 100,000 words. Thanks to this annotation, corpus users will be able to search for lexical meanings of words and study lexical ambiguity in broad linear and syntactic context. Among other things, we expect that such data will contribute to the development of a statistically driven module of automatic word sense disambiguation for Russian.

Benefits that accrue from lexical semantic annotation of the corpus can be illustrated by the ambiguous Russian verb *толковать*. This verb has (at least) three manifestly different lexical meanings: *толковать 1* 'interpret', "define' (in a dictionary, law etc.), as in *Русские словари толкуют честолюбие как негативную черту характера* 'Russian dictionaries interpret ambition as a negative character trait', *толковать 2* 'explain insistently', 'try to convince' *Он толковал мне, почему я ошибаюсь* 'He was explaining to me why I am wrong, and *толковать 3* 'converse', 'discuss', 'reason', as in *Они долго толковали о чем-то* 'They conversed long about something'. Importantly, these lexical units have very different linguistic properties. These properties, fully documented in the dictionaries of ETAP-3, include

(1) valency structures (*толковать 1* has a subcategorization frame close to that of *интерпретировать* 'interpret': *толковать что-л. как что-л.* 'define smth. as smth' or *толковать что-л. через что-л.* 'define smth. through <with, by way of> smth.; the subcategorization frame of *толковать 2* resembles (but is not identical to!) that of *объяснять* 'explain': *толковать о чем-л. кому-л.* 'explain smth. to smb', whilst *толковать 3* approaches the behavior of the symmetrical verb *беседовать* 'talk': *толковать о чем-л. с кем-л.* 'speak about smth. with smb';

(2) derivation (*толковать 1* has a deverbal noun *толкование* 'act of interpretation' or 'lexicographic definition', while *толковать 2* and *толковать 3* have no derivatives), and even

(3) morphological peculiarities '*толковать 1* is a transitive verb which has passive forms but it has no perfective aspect; *толковать 2* is, formally, a transitive verb (even though its direct object can only be realized by certain pronouns in the accusative case, like *толковали что-нибудь, <такое, свое>* 'explained something <something of this kind, their own thing>') but has neither passive forms or perfective aspect, whereas *толковать 3* is an intransitive verb that has no passive forms but it has the perfective aspect *потолковать*).

In a lexically underspecified corpus, it is impossible to sort out sentences that contain *толковать* in one particular sense, so it would be hard to establish, validate or rectify the information on specific lexical units, which could otherwise be used in many actual tasks (including those requiring machine learning).

It should be emphasized that, since SynTagRus is compiled semi-automatically, in many cases the linguist expert that edits the results of automatic parsing corrects the resulting structure containing particular words even it is not corroborated by the existing dictionary or grammatical data (which may be incomplete or not very accurate), without actually updating such data – the natural reason being that the expert may lack expertise, authority, or simply time. As a result, the deeply tagged corpus – not only SynTagRus but any corpus built on similar principles – acts, in many respects, as a source of invaluable data for linguists.

To continue with the example of *толковать*, a corpus that distinguishes word senses will enable us to see that e.g.

(a) the sentence *Ресторанные словари толкуют о каком-то соусе и каштанах* ‘restaurant dictionaries [whatever these are!] talk about some sort of sauce and chestnuts’ contains the word *толковать 2,* rather than *толковать 1,* which could be anticipated in the context of the word *словари* ‘dictionaries’;

(b) for subtle semantic reasons (*writing techniques* hardly needs interpretation, and interpretation hardly requires an addressee) the sentence *Помню, как он улыбался, толкуя мне китайскую письменность* ‘I remember how he was smiling, explaining to me Chinese writing’ also contains the word *толковать 2* despite the fact that the verb in this sense hardly accepts a non-pronominal direct object, unlike *толковать 1* – and the respective piece of information on this verb sense should be added to the dictionary;

(c) the sentence *Боюсь, что она это превратно понимает и толкует, как будто я забыл ее и не хочу ее видеть* ‘I am afraid that she misapprehends it and interprets (it) as though I have forgotten her and do not wish to see her’ contains *толковать 1* even though its third valency (of content) is presented in a highly non-canonical way – by a subordinate clause introduced with the conjunction *как будто.*

As follows from these examples, it is not at all easy to provide quality lexical semantic annotation of the corpus: this endeavour requires much time – and intellectual labour – of experienced annotators. The amount of work to be done can be properly assessed if we take into account the number of ambiguous words in 100,000-strong ETAP-3 dictionary (ca. 3,300 vocables whose lexemes share the same lemma names, representing about 6,700 word senses). We strongly believe, however, that the resulting corpus will be well worth this effort.

Fig. 2 below presents the structure of a corpus sentence from (c) with ambiguous words marked for concrete senses (here, words *что 1, толковать 1*, *как будто 1, и 1* and *не 1* specify such senses), while Fig. 3 summarizes the information on one of the respective lexemes – *толковать 1*.

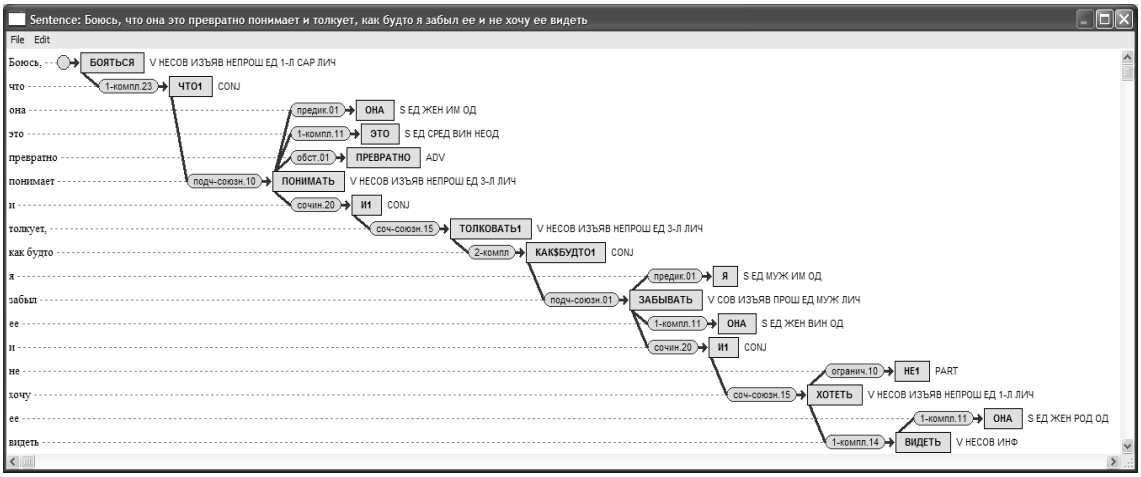


Fig.2. A sentence tagged for syntax and lexical semantics

Most of the boxes of the view presented by Fig. 3 are self-evident. KS name is that of the the word’s entry in the combinatorial dictionary of ETAP-3: it is clear that the corpus essentially relies on this particular dictionary, so that future researchers working with this corpus may require access to it.

So far, the number of SynTagRus sentences fully tagged for word senses is over 6,000, and it is constantly growing.

Apparently, lexical semantic annotation adds predictive power to the corpus and makes it a much more valuable linguistic resource.

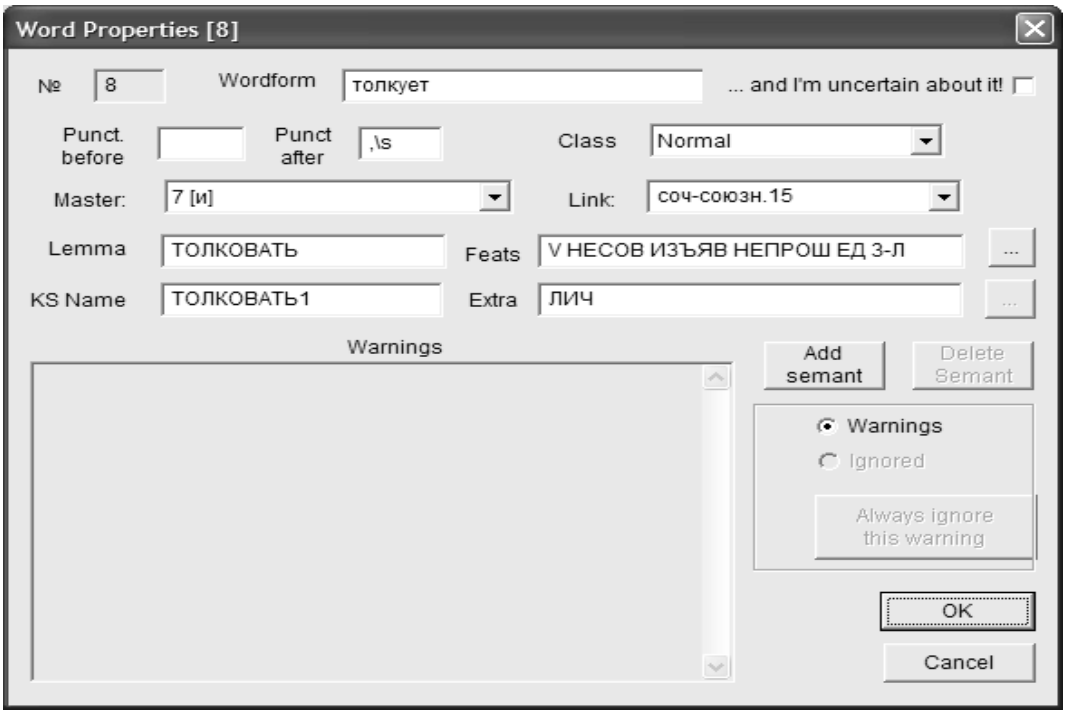


Fig.3. Properties of a specific word from the corpus sentence.

4. Lexical Functional Annotation

Lexical functional annotation consists in detecting lexical functions (LF) and their values in texts and tagging them for this type of data. Specifically, we plan to reveal occurrences of collocate type LFs in the sentences of SynTagRus and record them as part of sentence annotation. As a result, ample LF material will be available to researchers. So far, very few dictionary resources have provided such data. If collocates are marked in the text, direct observation and research of contexts in which lexical functions are realized will be possible. These data are of immense value for natural language processing systems.

The notion of lexical function was first proposed by the author of the Meaning ⇔ Text linguistic theory, Igor Mel'čuk, in 1970s and has been extensively studied and developed by the Moscow Linguistic School, in particular, by the Laboratory of Institute of Information Transmission Problems with active participation of Juri Apresjan. We have developed a number of NLP applications using LFs, including machine translation, where LFs are used to resolve lexical and syntactic ambiguity and achieve idiomatic translation of collocations, and an experimental system of synonymous paraphrasing for Russian.

A prototypical LF is a triple of elements {R, X, Y}, where R is a certain general semantic relation obtaining between the argument lexeme X (the keyword) and some other lexeme Y which is the value of R with regard to X (by a lexeme in this context we mean either a word in one of its lexical meanings or some other lexical unit, such as a set expression). Y is often represented by a set of synonymous lexemes Y_1, Y_2, …., Y_n, all of them being the values of the given LF R with regard to X. To give a simple example, MAGN is a LF for which the semantic relation is 'high degree'. Respectively for English,

MAGN (*desire*) = *strong / keen / intense / fervent / ardent / overwhelming*,

and for Russian,

MAGN (*желание*) = *сильный /упорный/ настойчивый / горячий / страстный / неудержимый / неутолимый / большой.*

Two types of LFs are distinguished: – paradigmatic LFs (substitutes) and syntagmatic LFs (collocates, or parameters).

Substitute LFs replace the keyword in the given utterance without substantially changing its meaning or changing it in a strictly predictable way. Examples are synonyms, antonyms, converse terms. A special subclass of substitute LFs is represented by various types of derivatives of X (*nomina actionis*, as in *to encourage – encouragement*, typical agents, as in *to build – builder* or *to judge – judge*, typical patients, as in *to nominate – nominee*, *to teach – student* and the like). All of them play an important role in paraphrasing sentences. Cf., for example: *She bought a computer for 500 dollars from a retail dealer – A retail dealer sold her a computer for 500 dollars – She paid 500 dollars to the retail dealer for a computer – The retail dealer got 500 dollars from her for a computer*

Collocate LFs appear in an utterance together with the keyword. Typically, such LFs either dominate the keyword syntactically or are dominated by it, although more elaborate syntactic configurations between the keyword and an LF value are not infrequent. Typical examples of collocate LFs are adjectival LFs, such as MAGN, or support verbs of the OPER / FUNC family.

This family of LFs can be exemplified by OPER 1 – a semantically empty verb such that the **first** actant of a certain situation functions as the subject of this verb and the name of the situation itself is the verb's first object: In Russian, OPER 1 (*контроль*) = *осуществлять* (cf. *to exercise control*).

In much the same way, OPER 2 is a semantically empty verb such that the **second** actant of a certain situation functions as the subject of this verb and the name of the situation itself is the verb's first object: OPER2 (*контроль*) = *подвергаться* (*контролю*), *находиться под* (*контролем*), *быть под* (*контролем*). (cf. *be under control*),

Collocate LFs play a leading role in the paraphrasing system of ETAP-3, providing paraphrases like *He respects his teachers – He has respect for his teachers – He treats his teachers with respect – His teachers enjoy his respect*", or *The United Nations ordered Iraq a report on chemical weapons – the United Nations gave Iraq an order to write a report on chemical weapons – Iraq was ordered by the United Nations to write a report on chemical weapons – Iraq received an order from the United Nations to write a report on chemical weapons*.

We are planning to mark a substantial part of our corpus with lexical functional annotation, too. As with syntactic and lexical semantic annotation, this work will be done semi-automatically. Since the ETAP-3 parser has a set of special post-syntactic rules that identify arguments and values of most collocate LFs (primarily if they appear in prototypical syntactic positions), the results will be used as raw material for manual correction and tagging by the annotator.

To give an example, for the sentence *Лил проливной дождь* 'A heavy rain was pouring' the parser will provide the following information on lexical functions:

MAGN(ДОЖДЬ) = ПРОЛИВНОЙ
FUNC0(ДОЖДЬ) ЛИТЬ1

These data will supplement the syntactic and lexical semantic tagging of SynTagRus. By the end of the year 2009, we expect to have at least 1,000 sentences of the corpus marked for lexical functions.

References

[1] Apresjan et al., 2003: Apresjan, Ju., I. Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V, and Tsinman, L. ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT. // Proceedings of the First International Conference on Meaning-Text Theory, 279–288.

[2] Apresjan et al., 2005: Apresjan, Ju.D., Boguslavsky, I.M., Iomdin, L.L., Iomdin, B.L., Sannikov, A.V., Sannikov, V.Z., Sizov, V.G., Tsinman, L.L. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. (Syntactically and Semantically Annotated Corpus of Russian: State-of-the-Art and Prospects) // National Corpus of Russian 2003–2005 (Results and Prospects). M: Indrik, 2005. P.193-214. (In Russian.)

[3] Boguslavsky et al., 2002: Boguslavsky, I.M., Iomdin, L.L. Chardin, I.S., Kreidlin, L.G. et al. Development of a Dependency Treebank for Russian and its Possible Applications in NLP // Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). Paris: European Language Resources Association, 2002. Vol. III. P. 852-856.

[4] Boguslavsky et al., 2008: Boguslavsky, I.M., Iomdin, L.L. Valeev, D.R., and Sizov, V.G. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов (ETAP Parser and its Evaluation Using a Deeply Annotated Corpus of Russian Texts) // Corpus Linguistics 2008. St. Petersburg, St. Petersburg State University, 2008. ISBN 978-5-288-04769-5. P. 56–74.

[5] Hajič et al., 2001: Hajič, J., Vidova-Hladká, B., Panevová, J. Hajičová, E., Sgall, P., and Pajas, P.. 2001. Prague Dependency Treebank 1.0. LDC, 2001T10.

[6] Mel'čuk, 1988: Mel'čuk, Igor. Dependency Syntax: Theory and Practice. State University of New York Press, 1988.

[7] Nivre et al., 2008. Nivre, Ioakim; Boguslavsky, Igor; Iomdin, Leonid. Parsing the SYNTAGRUS Treebank of Russian // Coling 2008. 22nd International Conference on Computational Linguistics. Proceedings of the Conference. ISBN: 978-1-905593-47-7. Vol. 2. pp. 641–648.

Slovak Medical Terminology
Is a World Wide Interoperability in Medicine Possible?

Oskár Kadlec

Civic Association For Slovak Medical Terminology
Smetanova 19, 811 03 Bratislava
o.kadlec@chello.sk

Abstract. We have identified the key role of the importance of national medical terminologies and the key role of syntactic tools in the creation of electronic health record. We propose actions to achieve full semantic interoperability across not only European but global worldwide health systems. The health is not a privilege of English speaking people.
Keywords: Electronic Health Record, Medical Terminology, Semantic Interoperability, eHealth

1. Introduction

The specific natural communication tool of the man is the language. The elements of the languages are organized in vocabularies, dictionaries, lexicons, thesauri, encyclopaedias etc. [28, 35, 38, 55, 61]. There is no science or human activity for which the communication is so important as it is for medicine and health care [29, 30]. Therefore, medicine has always paid an extraordinary attention to ontology and terminology [22, 24, 27, 38, 44, 47, 49, 53, 60, 66]. There is, however, a considerable disunity in the field of medical terminology on the meaning of several concepts, terms, names including the names of drugs [20, 33, 36, 50, 69].

In the era of information communication technology the **computer science** is man's powerful tool is [11, 23, 30, 31, 34, 37, 43, 45]. It enables to carry out the healthcare more effectively than ever before. In medicine and healthcare, during the implementation we encounter, however, with many barriers that make the understanding between the **man – machine – man** difficult if not impossible. The machine does not understand neither concepts nor terms without coding. The man, other hand, does not understand encoded concepts or terms in a strange language. Therefore, there must be a way of translating concepts into a digital form [encoding] and even a way of translating encoded concepts from a source language into a target one. There are, in our opinion, only one way how to try to solve this problem – each participant of the Unified Medical Language System must posses their own national coded terminology – e. g. translation of SNOMED CT. Thus can be obtained a national dictionary – a system of encoded concepts or terms with their semantic meanings. Their aim is to generate machine readable representations of medical concepts. This would facilitate its adoption as the standard for medical knowledge representation in biomedical informatics [40]. For more than 15 years, the European Commission has recognized the importance of terminologies and interoperability by funding research in this fields. Therefore, in its Semantic HEALTH roadmap various challenges in respective domains have been pointed out requiring to take actions on the path to semantic interoperability in order to support European health services. A policy of incremental steps and a focused, modest approach to terminology development in an open, collaborative environment is the ultimate recommendation resulting from the project's work [3, 4, 23, 58].

The issues of technical standardization are no longer the most prominent ones in realizing the interoperability. The most challenging part still to achieve is semantic interoperability of Electronic Health Record systems. It plays a prominent role in the recently published Recommendation on Interoperability of Electronic Health Record Systems (COM(2008)3282). It calls not only for interoperability at regional and national level but also at EU level – a goal which realistically might take another **20 years** to be fully achieved [58].

As far as we know, however, there are no studies on interoperability between medical terminologies in the countries of the Slav community. It is, therefore, an important challenge to form prerequisites – organizational, institutional as well as personal ones - for the development of semantic and syntactic tools enabling the interoperability between the languages of member states of EC and Slav languages.

2. What does semantic interoperability mean?

According to the Recommendation of EC semantic interoperability [58] means ensuring that precise meaning of exchanged information is understandable by any other system or application not initially developed for this purpose, whereas computer interoperability of electronic health record means the ability of two or more electronic health record systems to exchange both computer interpretable data and human interpretable information and knowledge [58].

There are four levels of interoperability, two of them relating to semantic interoperability (SIO). To explain and distinguish the 4 different levels, consider the following scenario [58]: 50-year old patient recently moved from Slovakia to Ireland to take up his new job. A few weeks after arrival, he falls ill, consults his local (Irish) general practitioner (GP) and is transferred to the next hospital for further tests. Depending on the level of established SIO the hospital has to initiate the following steps:

Level 0 – no interoperability at all: The patient has to undergo a full set of lengthy investigations for the doctor to find out the cause of his severe pain. Unfortunately, results from the local GP as well as from his Slovak GP are not available at the point of care within the hospital due to the missing technical equipment.

Level 1 – technical and syntactical interoperability: patient's doctor in the hospital is able to receive electronic documents that were released from the Slovak GP as well as his local GP upon request. Widely available applications supporting syntactical interoperability (such as web browsers and email clients) allow the download of patient data and provide immediate access. Unfortunately, none of the available doctors in the hospital is able to translate the Slovak document and only human intervention allows interpretation ofthe information submitted by the local GP to be added_into the hospitals information system.

Level 2 – partial semantic interoperability: The Irish hospital doctor is able to securely access, via the Internet, parts of patient's Electronic Health Record released by his Slovak GP as well as by the local GP that he had visited just hours earlier. Although both documents contain mostly free text, fragments of high importance (such as demographics, allergies, diagnoses, and parts of medical history) are encoded using international coding schemes, which the hospital information system can automatically detect, interpret and meaningfully present to the attending physician.

Level 3 – full semantic interoperability, co-operability: In this ideal situation and after a thorough authentication took place, the Irish hospital information system is able to automatically access, interpret and present all necessary medical information about the patient to the physician at the point of care. Neither language nor technological differences prevent the system to seamlessly integrate the received information into the local record and provide a complete picture of the patient's health as if it would have been collected locally. Further, the anonymous data feeds directly into the tools of public health authorities and researchers.

It must be kept in mind that SIO implementation also depends on social, cultural and human factors within respective organisation, region and country, system and time period.

3. Classification, Nomenclatures and Thesauri

Statistically reliable data based on qualified **classifications** are essential for an efficiently regulated Health Care System [35, 38, 55, 61]. Appropriate classifications help to unite various medical terms. [5, 42, 51, 52, 68], There are many classification systems in medicine and Health Care Systems, as follows

ICD [10] – International Classification of Diseases released by the World Health Organization [WHO] serves globally as a diagnosis related classification and is the basis for internationally comparable mortality. However, many countries have issued their own version of ICD. For example Deutsches Institut für medizinische Dokumentation und Information [DIMDI], in GB ICD-9 is still in use and so on.

There are also versions of ICD-10, such as **ICD-O-3** – a special adaptation for the documentation of tumours, **ICF** – International classification of Functioning, Disability and Health. It is a result of medical progress and the rising life expectancy age, chronic illnesses and the treatment of persons with permanent defects. The concept of "disease" itself is no longer sufficient to describe the population's state of health etc. [67].

MeSH – the Medical Subject Headings, a medical thesaurus published and annually updated by the US National Library of Medicine (NLM) in Bethesda (Maryland, USA). It is used for cataloguing library holdings and indexing databases that are produced by the NLM (e. g. MEDLINE). Since a comparable thesaurus is missing, the MeSH has been translated into many languages including Slovak and among others also German [46].

UMLS – United Medical Language System that includes medical terms and semantic relations between them. The terms originate from about 100 heterogeneous conceptual order systems and medical nomenclatures of many languages. DIMDI for example supplies extensive German-language vocabularies to the UMLS annually and, in the meantime, has made German second most frequent language in the Metathesaurus [64].

SNOMED CT® (Systemized Nomenclature of Medicine Clinical Terms) & **IHTSDO** (International Health Terminology Standards Development Organization in Copenhagen [Denmark] was formed in 1991 by USA's SNOMED RT and UK's CTV3 (Read codes). SNOMED CT owned by the College of American Pathologists [Northfield, DC] [32,52, 56, 57].

SNOMED CT® is a comprehensive clinical terminology that provides clinical content and expressivity for **clinical documentation and reporting**. It can be used to code, retrieve, and analyse clinical data. The terminology comprises concepts, terms and relationships with the objective to precisely represent clinical information across the scope of health care. Content coverage is divided into 19 hierarchies (e. g. clinical finding, procedure, observable entity etc.).

SNOMED CT provides a **standard for clinical information**. Software application can use concepts, hierarchies, and relationship as a common reference point for data analysis. SNOMED CT serve as a foundation upon which health care organizations can develop effective analysis applications to conduct outcomes research, evaluate the quality and cost of care, and design effective treatment guidelines.

Standardized terminology can provide benefits to clinicians, patients, administrators, software developers and payers. Clinical terminology can offer the health care providers accessible and complete information pertaining to the health care process more easily(medical history, illnesses, treatment, laboratory results, etc.) and thus can result in improved patient outcomes. A clinical terminology can allow a health care provider to identify patients based on certain coded information in their records, and thereby facilitate follow-up and treatment.

We would like to inform you about **some problematic issues** with which we are often encountered in the creation of Slovak medical terminology and translation of SNOMED CT®.

The vocabulary used to describe terminologies, ontologies, and classification systems has always been a source of confusion, since different authors used the same words differently.[58]

4. Unified medical languages and communication barriers

Most considerations about eHealth are based on a false assumption that there exists a **unique international terminology** (Latin-Greek or English) and it is only a question of time, when all countries will accept and employ it. The history, however, teaches us that there is no nation that renounces its mother tongue on its own free will, even if it is not for it its sake [29, 30].

Unfortunately, the health care administrators and health care providers are not aware of all real requirements in computerization of medicine and health care system. There are many obstacles that hinder the employment of computers and the implementation of information systems in practice.

The **communication barriers** are various, as follows:

- **linguistic regional barriers** – there are about 3000 thousand languages in the world [without dialects]` the question is: should all the nations have their own medical terminology?
- **interpersonal barriers** – doctor/patient, doctor/other health professionals (it is difficult for the layman to understand many professional terms: should be the medical terms for the patient's sake expressed in colloquial language?)
- **interdisciplinary barriers** – each science has its own language as one of its main characteristics, has its own tools and rules; the language has a function as the organizer of the knowledge etc. (the are more than 100 medical disciplines or branches with their own terminologies; e. g. Terminologia anatomica, Nomina histologica and Nomina embryologica, which act as standards in their fields [1,8, 10, 13 – 18, 21, 22, 39, 48, 54, 59, 62, 70, 71]. These terminologies are available only in Latin and English and their worldwide adoption is subject to the addition of terms from other languages; on the other hand, Nomina anatomica, the previous standard, has been widely translated)
- **legislative barriers** – there are many conventional nomenclatures, classifications and other systems reached by mutual achievement between professional or scientific associations, e. g. example Système International of Units and Quantities – SI, International Union of Pure and Applied – IUPAC International Federation of Clinical Chemistry etc.
- **Alphabetical differences** – Cyrillic, Chinese etc. scripts
- However, the main problem that could be most easy solved is the discrepancy between **US and European terminologies and standards**

A medical terminology enables the employment of information and communication technology in making the health care system more effective and economically favourable.

Based on the SNOMED CT® every language can formulate its own medical terminology, i. e. its own **extension** of the core. A number of incorrect and misleading terms are to be replaced. Each term must have a unique code number and must be supplied with a national equivalent. The use of eponyms is discouraged, but a list of well known ones can be appended to facilitate accessibility of older literature. Relevant suggestions about amendments are eagerly awaited and a broad basis of endorsement among the medical scientific world is hoped for.

The nomenclature is presented either per system or organ or according to the main domains of the medical science and health care. An alphabetic index follows medical terminology as well as English and Latin medical terminology list. These translation products should be edited in form of national terminological dictionaries [41, 42].

The creation of coded national medical terminology is, however, only one part of the problems. If we have a dictionary, it does not mean, that we are able to form sentences, statement, judgements and so on. Each interested party or the system as a whole must have available **syntactic tools** for the creation of electronic health records and similar products.

We consider the work on the creation of **national coded medical terminology** as a starting point for any further activities associated with the computerization of the healthcare system.

The **National eHealth Strategy** included in the implementation priorities for eHealth development in Slovakia a possibility of the existence of the **national terminology** as a natural prerequisite, but which in fact does not exist in a consistent and coded form.

The most important issues of the National eHealth Strategy comprise these tasks:

- development of the National Healthcare Information System
- healthcare related national portal for both, professionals and public
- upgrading the network of national healthcare providers with provisions for domestic and international interoperability
- citizen and professional electronic health identification cards
- ePrescription/e-Medication
- active participation in development of electronic health record in close cooperation with EuroRec and ProRec Center Slovakia
- telemedicine and independent living
- ICT supported home – health and social – care systems [65]
- knowledge based advisory and decision support (expert) systems for general practitioners, clinicians, and management
- introduction of the surveillance systems with regard to clinical practices, patient, safety, and quality of care certification of clinical guidelines
- application of ICT and healthcare related standards (from CEN TC_{251} and ISO_{215}, SNOMED CT, HISA, DICOM, …)

5. Summary and propositions

As the most important tasks in the field of the computerization of eHealth we can consider:

1. **Unification of International systems** of terminology, nomenclature and classification (SNOMED CT, MeSH, ICD, SI etc.) and their worldwide acceptance. Unfortunately, disunity of expression of names of units and quantities still persists mainly in physics, chemistry and biochemistry, e. g. of the names of measures, weights, lengths etc.

2. Creation of a system of **coded unified and certificated national medical terminology** in general and subsequent creation of particular domains terminology (biology and genetics, anatomy, histology, embryology, individual disciplines of clinical medicine and paramedical sciences and so on); elaboration of a database of preferred medical terms and of their synonyms and eponyms.

3. Inclusion of the medical terminology in the national thesauruses (Corpus) and coordination of terms from other related discipline ("exact" sciences, as biophysics, biochemistry and molecular biology, "metatheoretical" sciences, as biomathematics, biostatistics, etc., psychology, sociology, ethics etc.).

4. Establishment of an **Expert Committee for settlement of a Consensus between Slave nations** in the field of coded medical terminology that will enable interoperability between them in termsof worldwide medicine without frontiers.

5. Putting a section (column) in the web site of JÚĽŠ with editorial board as an informal body devoted to the **international questions of medical terminology.**

References

[1] Anatomické názvoslovie. Vydavateľstvo SAV, Bratislava, 576 p (1962).

[2] Becher, I., Lindner, A., Schulze, P.: Lateinisch-griechischer Wortschatz in der Medizin. Berlin, Volk und Gesundheit 251 p. (1986).

[3] Centre for Health Informatics & Miltiprofessional Education (CHIME), "The Good European Health Record", available from: http://www.chime.ucl.ac.uk/work-areas

[4] Commission on the European Communities: e-Health – making healthcare better for European citizens: An action plan for a European e-Health Area. 2004 Available from: http://europa.eu.int/information societyx/doc/qualif/health/COM 2004 0356 F EN ACTE.pdf (accessed 2006-11-21).

[5] Dorland's Illustrated Medical Dictionary. 29th ed., W. B. Saunders Company, Philadelphia, 2987 p. (2000)

[6] Duden. Das Wörterbuch medizinischer Fachausdrücke. 4. vyd. Mannheim, Bibliographisches Institut, Stuttgart , Thieme Verlag, 744 p. (1985).

[7] Ďuriš, I.: K problému elektronického chorobopisu. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 22 – 25 (2008).

[8] Dvořák, J.: Srovnávací slovník anatomických nomenklatur. Praha, SZN, 293 p. (1960).

[9] Encyklopedičeskij slovar medicinskich terminov. 1–3. Moskva. Sovetskaja encyklopedija, 463 p., 447 p., 512 p. (1982 – 1984).

[10] Feneis, H.: Anatomický obrazový slovník. GRADA Publishing, Praha (1996).

[11] Fitoš, E.: e-Medicatiopn – rakúsky príklad elektronickej komunikácie v systéme zdravotníckej starostlivosti. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 55 (2008).

[12] Health Level Seven, Inc., "HL7 Version 3 Standards" 2005, homepage on the Internet, available from: http://www.hl17.cz/.

[13] Holomáňová, A., Brucknerová, I.: Míľniky v histórii anatomického názvoslovia. In M. Weis (ed.): Medicína a informačnokomunikačné technológie.Bratislava, Asklepios, p. 14 – 18 (2008).

[14] Holomáňová, A., Brucknerová, I.: Srdcocievna sústava. Anatonické názvy. Latinsko-anglicko-slovenský slovník. Vydavateľstvo Elán, Bratislava, 69 p. (2000).

[15] Holomáňová A., Ivanová A. Brucknerová. I., Beňuška, J.: Andreas Vesalius – the reformer of anatomy. Bratisl. Lek. Listy, 102, č. 1, p. 48 – 54 (2001).

[16] Holomáňová, A., Brucknerová, I.: Anatomické názvy III. Latinsko-anglicko-slovenský slovník. Vydavateľstvo Elán, Bratislava, 154 p. (2003)

[17] Holomáňová A., Ivanová, A., Brucknerová I.: Mondino – anatomicus clarus. Abstracta 5. medzinárodní sympozium k dějinám medicíny, farmácie a veterinární medicíny. Hradec Králové VJA JEP, p. 49. ISBN 80-85109-34-4 (2001).

[18] Holomáňová, A., Ivanová, A., Brucknerová I.: Fragment of Albrecht von Haller's life, Trendy a perspektívy současné morfologie. Plzeň: Ústav histologie a embryologie LF UK, p. 5. ISBN 80-223-8500-6 (2002).

[19] Horecký, J.: Základy slovenskej terminológie. Vydavateľstvo SAV, Bratislava, 146 p. (1956).

[20] Hügel, H.: Kurze Einführung in die pharmazeutische und medizinische Terminologie. Stuttgart, Deutscher Apotheker-Verlag, 80 p. (1973)

[21] Hyrtl, J.: Onomatologia anatomica. New York, Georg Olms Verlag, Hildesheim, 626 p. (1970).

[22] Ivanová A.: Anatomická nomenklatúra z historického hľadiska. Bratisl. Lek. Listy 91, n. 2, p. 146 – 173 (1990).

[23] Jung, B., Grimson, J.: "Synapses/SynEx goes XML", Studies in Health Technology and Informatics. Vol. 68, p. 906 – 911 (1999).

[24] Kábrt J., Chlumská E.: Lékařská terminologie. Avicenum, 2. vyd. Praha, 1980, 324/16 p (1980).

[25] Kábrt, J., Kábrt, J., Jr.: Lexiconum medicorum. Druhé, doplněné a přepracované vydáni. Galén, Praha, 1136 p. (2001).
[26] Kábrt, J., Valach, V.: Latina pro mediky s přihlédnutím k řečtine. Druhé vydanie, SZdN, Praha (1962).
[27] Kábrt, J., Valach, V.: Stručný lékařský slovník. Tretie vydanie SZdN, Praha (1965).
[28] Kadlec, O.: Encyklopédia medicíny. Asklepios, Bratislava 5600 p. (1993 – 2005).
[29] Kadlec, O.: Jednotná terminológia v elektronickej medicíne. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 7 – 13 (2008).
[30] Kadlec, O.: Význam terminológie pre elektronizáciu zdravotníctva. In M. Weis (ed.): Medicína a informačnokomunikačné technológie. Bratislava, Asklepios p. 23 –32 (2008).
[31] Kadlec, O., jun.: Skúsenosti s informačnokomunikačnými technológiami vo svete. In M. Weis (ed.): Medicína a informačnokomunikačné technológie. Bratislava, Asklepios, p. 54 – 59 (2008).
[32] Kadlec, O., jun.: Základná charakteristika SNOMED CT. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios 2008, p. 14 – 25 (2008).
[33] Kazačenok, T.G.: Posobije po izučeniju farmakologičeskoj terminologii. Minsk. Vyššaja škola, 287 p. (1974).
[34] Kovárová, M.: Prínosy implementácie SNOMED CT. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 48 – 50 (2008).
[35] Krátky slovník slovenského jazyka. Veda, Vydavateľstvo SAV, Bratislava, 1997, 592 p. (1997)
[36] Krchňák, Š.: ePreskripcia pohľadom lekárnika. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 44 – 47 (2008).
[37] Krištúfek, P.: Míľové kroky elektronickej medicíny. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 26 – 29 (2008).
[38] Langová, T.: Anglicko-slovenský slovník medicíny. Veda, Vydavateľstvo SAV, Bratislava, 578 p. (1996).
[39] Ledényi, J.: Nomina anatomica. Slovenské telovedné názvoslovie. Matica Slovenská, Turčiansky Svätý Martin, 242 p. (1935).
[40] Lipka, J., Mukenšnábl, Z., Horáček, F., Bureš, V.: „Současný komunikačný standard českého zdravotnictví DASTA“, In: Zvárová, J., Přečková, P. (eds): Informační technologie v péči o zdraví, EuroMISE s. r. o., Praha, p. 52 – 59 (2004).
[41] Manuila, A. (ed.): Progress in Medical Terminology. Basel, S. Karger, 116 p. (1981).
[42] Manuila, L., Manuila M. Nicole, Lambert H.: Dictionnaire français de médicine et de biologie. 1 – 4. Ed. Paris, Masson a Cie 1970 – 1975, 865 s., 923 s., 1193 s, 561 p.
[43] Marko, P., Janíčková E.: Využitie informačných technológií v praxi všeobecného lekára pre dospelých. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios 2008, p. 36 – 43.
[44] Martinková, K.: K dejinám slovenského názvoslovia. In M. Weis (ed.): Medicína a informačno-komunikačné technológie.Bratislava, Asklepios, p. 7 – 13 (2008).
[45] Miro International Pty Ltd®, “Ocean Informatics”, 2000–2004, homepage on the Internet, available from: http://oceaninformatics.biz/CMS/index.php. (2000-- 2004).
[46] National Library of Medicine, “Medical Subject Headings”, homepage on the Internet, available from: http://www.nlm.nih.gov/MBrowser.html.
[47] Nociar, A.: Význam jazyka v medicíne a psychológii. In M. Weis (ed.): Medicína a informačno-komunikačné technológie.Bratislava, Asklepios, p. 19 – 22 (2008).
[48] Nomina anatomica, 6 ed., Nomina Histologica, 3 ed., Nomina Embryologica, 3 ed. Churchill Livingstone, Edinburgh (1989).
[49] Porep, P., Steudel, W. J.: Medizinische Terminologie. Stuttgart, Beorg Thieme Verlag, 330 p., (1974).
[50] Praescriptiones magistraliter. 3. vyd. Praha, Avicenum 1974, 431 p. (1974).

[51] Pschyrembel, W.: Klinische Worterbuch mit klinischen Syndromen. 256. Aufl. Berlin-New York, W. de Gruyter, 1876 p. (1990).
[52] Regenstrief, Inc. "Logical Observation Identifiers Names and Codes – LONC®", homepage pn the Internet, 2005, available from: hhtp://www.cog.ufl.edu/publ/apps/icdo/ (2005).
[53] Sedláček, S.: Medicínská terminologie. Učební texty vys. Škol, SPN, Praha (1967).
[54] Simpson, I.: Anatomie človeka, REBO, Praha, 144 p. (1994).
[55] Slovník súčasného slovenského jazyka. A –G. Veda, Bratislava, 1134 p. (2006).
[56] SNOMED International®, "Systemized Nomenclature of Medicine", homepage on the Internet, available from: http://www.snomed.org/. (2004).
[57] SNOMED International, "Systemized Nomenclatur of Medicine – Clinical Terms", homepage on the Internet, available from: http://www.snomed.org/snomedct/.
[58] Stroetman, V. N. (ed.): Semantic Interoperability for better Health and Safer Healthcare. European Commission (available at www.semantichealth.org) (2009).
[59] Šimo, J., Finka, M.: Sémantická integrácia pomocou virtuálneho modelu ľudského tela – The Medical Information Hub. In O. Kadlec (ed.): Elektronická medicína. Bratislava, Asklepios, p. 51 – 54 (2008).
[60] Šimon, F.: Latinská lekárska terminológia. Osveta, Martin, 1990, p. (1977).
[61] Špaňár, J., Hrabovský, J.: Latinsko-slovenský a slovensko-latinský slovník. Pedagogické nakladateľstvo, Bratislava, 1222 p. (1988).
[62] Terminologia Anatomica. International Anatomical Terminology. FCAT Thieme, Stuttgart. 292 p. (1998).
[63] Tomečková, M.: „Minimální datový model kardiologického pacienta – výběr dat", In Cor et Vasa, vol. 44, č. 4, suppl-. ISSN: 0010-8650, p. 123 (2002).
[64] United States National Library of Medicine, National Institute of Health, „Unified Medical Language System", homepage on the Internet, available from: http://ww.nlm.nih.gov/research/umls/.
[65] Vörösová, G. a kol.: Klasifikačné systémy a štandardizácia terminológie v ošetrovateľstve. Martin, Osveta 113 p. (2007).
[66] Wiese, I.: Fachsprache der Medizin. Leipzig, VEB Verlag Enzyklopädie 144 p. (1984)
[67] World Health Organization®, "International Classification of Diseases", homepage on the Internet, avalaible from: http:/www.who.int/classifications/icd/en/ (2005).
[68] Woxbridge Solutions Ltd®, "General Practical Notebook – a UK Medical Encyclopaedia on the World Wide Web", http://www.gpnotebook.co.uk/simplepage.cfm? ID=1134166031 (2005).
[69] Zábojníková, M., Kovalčík, V.: Receptúrna propedeutika. 2. vyd. Martin, Osveta 1976, 184 p. (1976)
[70] Zrzavý, J.: Anatomie pro výtvarníky. Avicenum, Praha, 1977, 400 p. (1977).
[71] Zrzavý, J.: Latinsko-české anatomické názvosloví. Univerzita Palackého, Olomouc (1985).

Russian Dictionary Base – First Steps

Karel Pala[1], Adam Rambousek[1], Maria Khokhlova[2], and Victor Zakharov[2]

[1] Center for Natural Language Processing, Faculty of Informatics, Masaryk University
[2] Faculty of Philology and Arts, St. Petersburg State University

Abstract. The present paper deals with digitization of the Russian explanatory dictionaries. The aim of this paper is to present the main ideas of the digitization project for explanatory dictionaries of Russian and to describe the current state of the data sources. This project is intended to be realized in cooperation of the Faculty of Philology and Arts, St. Petersburg State University, Russia and the Faculty of Informatics, Masaryk University, Brno (FI MU), Czech Republic. The ultimate aim is to provide lexicographic software tools for developing explanatory dictionaries of Russian.

1 Introduction

The information society has become very quickly a computerized one. Constantly, new technologies come to new spheres of human activity. The arrival of corpus linguistics and corpora have become a relevant point in this respect. The corpora stimulated a considerable progress that has been gained in the field of automatization of lexicographic work. This has its own reason. There is no integrated software that enables to work both with traditional dictionaries and new electronic sources of lexical data.
The present paper deals with the explanatory Russian dictionaries.

The first explanatory dictionaries of Russian date as back as to the beginning of the XIXth century. Among dictionaries of contemporary Russian we can name Ushakov's Dictionary (1920-1930s) [1], Ozhegov's Dictionary (the first edition was published in 1949) [2], the Dictionary of the Russian Language in 17 volumes (also known as BAS – "Bol'shoj akademicheskij slovar' russkogo jazyka", 1948-1965) [3], the Dictionary of the Russian Language in 4 volumes [4](also known as MAS – "Malyj slovar' russkogo jazyka", 1957-1961), the Complex Normative Dictionary of the Modern Russian Language ("Komplexsnyj normativnyj slovar' sovremennogo russkogo jazyka") [5], and the Dictionary of the Russian Language in 25 volumes (also known as the "new" BAS – "Bol'shoj akademicheskij slovar' russkogo jazyka", since 2005) [6].

The intention is to collect resources of these dictionaries within the one framework. All these data will be converted into the well-structured format (e.g., XML format) and concentrated in a unified database. Such a database will be prepared for all kinds of linguistic research.

The idea has been existing for several years and was inspired by several similar projects abroad, as the Celex database [7], and the Czech lexical database [8, 9].

2 Entry Structures

As can be observed, the lexicographers follow several general but rather pragmatic principles in building the dictionary definitions. In other words, the techniques applied in building dictionary definitions are based on the selected general principles but we can hardly say that they form a consistent and complete theory. Though lexicographers use well-established techniques, many objections can be raised with regard to the consistency of the dictionary definitions, both from the formal and from the semantic point of view. Most of dictionaries of the same type have different structure of entries. A considerable number of the dictionary definitions are expressed just by examples. It is useful to have a look at the types of the definitions (meaning descriptions) that can be found in dictionaries:

- definitions using **genus proximum** (GP) and the distinguishers (differentia specifica); these are mostly typical for nouns: e.g. *poodle = a dog with thick curling hair;*
- definitions using **semantic components** or **features** (primitives), quite often with verbs: e.g. *kill = cause to die;*

- definitions based on the **relation of troponymy** are typical for verbs: e.g. *talk* = *whisper*, *cry* = *sob;*
- definitions using **synonymical explanations** or just one word synonyms, typical for adjectives, e.g. *clever* = *bright*, *beautiful* = *nice, pretty;*
- definitions based on **collocational determination** of the sense of entry, typical for adjectives, e.g. *good student*, *good mile*;
- definitions exploiting various kinds of **ad hoc descriptions** or explanations (these can occur with any part of speech);
- definitions based on the **descriptions of events** or **situations** (see e.g. the following definition: *if you ask for a* ***table*** *in a restaurant, you want to have a meal there).*

Emotional, expressive and stylistic connotations are indicated by special labels (for instance, *«неодобр.»* – with disapproval, *«презр.»* – derogatory, *«шутл.»* – humorous, *«ирон.»* – ironic, *«книжн.»* – bookish, *«разг.»* – colloquial etc).

Depending on dictionary size some meanings can be illustrated by examples – typical phrases that have the given word as its part or in case of large dictionaries by citations. As a rule explanatory dictionaries give also grammatical information, indicating by labels part of speech, gender, aspect etc., sometimes other word forms. Also entries sometimes include phonetic characteristics as stress or pronunciation.

All this produces a complex structure that in print is realized by the collection of labels and fonts.

Below you can see a dictionary entry for the word “goods” (Russian “tovar”) from the Dictionary of the Russian Language in 4 volumes [4].

ТОВА́Р, -а, *м.*

1. *Экон.*Продукт труда, произведенный для продажи. *Товар есть, во-1-х, вещь, удовлетворяющая какой-либо потребности человека, во-2-х, вещь, обмениваемая на другую вещь.*Ленин, Карл Маркс.

2. (*ед. ч. может употребляться и в знач. мн. ч.*). Предмет торговли. *Товары широкого потребления. Отпуск товара. ◊ В лавке у Караваева были собраны товары со всей страны – табаки из Феодосии, грузинские вина, астраханская икра, вологодские кружева, стеклянная мальцевская посуда, сарептская горчица и сарпинка из Иваново-Вознесенска.*Паустовский, Далекие годы.

3. *только ед. ч.*В сапожном деле: выделанная готовая кожа. *Одна за другой падали светлые капли на заскорузлые, черные, пропитанные варом руки Епишки, на сверкавшее острием шило, на дратву, на пахнущий товар сапога, зажатого между коленами.*Серафимович, Епишка.

◊ **Живой товар** *см.* живой.

Показать товар лицом *см.* лицо.

Let's analyze several fields of the entry.

ТОВА́Р, – entry word (with stress);

-а – morphological/grammatical information/zone: indication of word’s inflexion, typically in Genitive case as it’s usually difficult to reconstruct this word form;

м. – grammatical field: gender, eg. masculine;

1, 2, 3 – meaning number (in case of polysemy);

Эконом. – stylistic zone: indication the field of word usage, eg. economics

Продукт труда, произведенный для продажи – definition of the first sense/meaning;

Товар есть, во-1-х, вещь, удовлетворяющая какой-либо потребности человека, во-2-х, вещь, обмениваемая на другую вещь – illustrative zone: citation is used as an example;

Ленин, Карл Маркс, Паустовский – illustrative field: author whose citations about "tovar" were used in the entry;

◊ - phraseological field; this label is used to indicate: 1) lexical collocability, collocations, phrases or terminological units; 2) syntactic collocability; 3) word's typical usage, e.g. degrees of comparison; 4) expressive word usage (various connotations), as ironic or jocular;

Живой товар, Показать товар лицом – illustrative field: collocation is used as an example;

(см. живой) – reference field: links between entries, the label refers to another entry (eg., "живой") that defines the given collocation (eg., "живой товар").

If we gave examples of the entries of different explanatory dictionaries we could see a plenty of both common and distinct characteristics. This raises a question about a single structure of dictionary entries in electronic form and also about software and linguistic mechanisms that allow to represent existing dictionaries within this framework.

3 Electronic Dictionaries of Russian

Nowadays many dictionaries of the Russian language (including explanatory ones) exist in an electronic form. But usually these are scanned texts in either graphical or text formats. Lack of structuring makes it difficult to search in them.

Several Russian explanatory dictionaries are available on-line (through Feb-web: Fundamental Electronic Library[1]): Ushakov's Dictionary, the Dictionary of the Russian Language in 4 volumes, and the Dictionary of the Russian Language of the XVIIIth century [10].

But there is an option to look up only in one dictionary at the same time and browse in it but not to use it as a database. Because entries of different dictionaries have various structures that makes it hard to work with the data.

This raises the question of one integrated structure of Russian explanatory dictionaries and their conversion to this structure. Moreover, this also leads to the question of developing one tool that could be used both as browser and editor.

As data for our work we have chosen two dictionaries of Russian. They are the "Complex Normative Dictionary of the Modern Russian Language" ("Komplexsnyj normativnyj slovar' sovremennogo russkogo yazyka") [5] and the above mentioned Dictionary of the Russian Language in 4 volumes [4]. Below we will discuss the former one.

The "Complex Normative Dictionary of the Modern Russian Language" is being compiled at the Laboratory of Computational Lexicography of the Faculty of Philology and Arts (St. Petersburg State University, Russia) under the guidance of Prof. G.N. Sklyarevskaya. It is intended for users to provide them with information on correct word usage of latest and newest terms and concepts of modern Russia. The dictionary includes active vocabulary that isn't chosen on statistical principle but on its semantic, grammatical, orthoepic or other difficulty for language users. The usage of these words has to be normalized. The data is being actively revised and supplemented on the basis of corpus examples, Internet data, various terminological or explanatory dictionaries, and linguistic studies. Dictionary word list is compiled on the data of the Fund of Modern Russian (cca. 17 ml. tokens).

1 http://feb-web.ru

For the project implementation we have chosen the platform DEB that was developed at the Center of Natural Language Processing FI MU.

Let's illustrate the entry tree structure of the "Complex Normative Dictionary of the Modern Russian Language". For noun entry it has the following shape:

- headword zone зона заголовочного слова
  - headword заголовочное слово
    - morphology-inflection словоизменение
      - morphology словоизменение
      - free_textтекст комментария
      - mark помета
        - free_text_before текст комментария
        - mark_proper собственно помета
        - free_text_after текст комментария
      - morphology_variant вариант словоизменения
    - additional_morphology дополнительное словоизменение
    - …
  - headword variant вариант заголовочного слова
    - …
  - syntax управление
    - syntax управление
    - mark помета
    - syntax_example пример управления
  - pronunciation произношение
  - etymology происхождение
  - …
- data zone зона данных
  - meaning значение
    - meaning number номер значения
    - …
    - meaning толкование
    - example речение
      - example речение
      - example explanation подтолкование речения
    - phrase устойчивое сочетание
      - phrase устойчивое сочетание
      - free_textтекст комментария
      - phrase_variant вариант устойчивого сочетания
      - phrase_pronunciation произношение устойчивого сочетания
      - …
      - phrase_meaning толкование устойчивого сочетания
      - …
- additional data zone зона дополнительных данных
  - encyclopaedia энциклопедическая информация
    - encyclopaedia_mark помета энциклопедической информации
    - encyclopaedia энциклопедическая информация
  - error зона ошибок
    - error_mark помета при ошибке
    - error ошибка

4 Platform DEB

In cooperation between St. Petersburg State University and Center of Natural Language Processing (Masaryk University, Brno, Czech Republic) we propose to represent a number of Russian explanatory dictionaries by means of tools based on the given platform.
The Dictionary Editor and Browser (DEB) platform has been developed as a general framework for fast development of wide range of dictionary writing applications [11]. The DEB platform provides several very important features that are common to most of the intended dictionary systems. These basic features include:

- a strict separation of the client and server parts in the application design. The server part provides all the necessary data manipulation functions like data storage and retrieval, data indexing and querying, but also various kinds of data presentations using templates. In DEB, the dictionary entries are stored using a common XML format, which allows to design and implement dictionaries and lexicons of all types. The client part of the application concentrates on the user interaction with the server part, it does not produce any complicated data manipulation. The client and server parts communicate by means of the standard HTTP protocol;
- a common administrative interface that allows to manage user accounts including user access rights to particular dictionaries and services, dictionary schema definitions, entry locking administration or entry templates definitions;
- XML database backend for the actual dictionary data storage. Currently, we are working with the Oracle Berkeley DB XML database, which provides a flexible XML database with standard XPath and XQuery interfaces. We use two approaches to client part of the applications, depending on the complexity of the dictionary and user requirements.

Mozilla Development Platform. The Mozilla platform provides a complete set of tools for software development. Firefox web browser is one of the many applications created using this platform.

The platform is used to create rich applications with the grafical user interface. Applications are installed as an add-on to Firefox browser and thus works in every operating system supported by Mozilla Firefox.

Standard HTML webpage, enhanced with the Javascript functions. Main advantage is that the webpage can be accessed from any web browser. On the other hand, it can't provide all the features of the Mozilla Platform.

Even though webpages are generated by the server, they act as a client application and use the same HTTP API interface to communicate with the server part. Web browser access is used for editing dictionaries with less complicated entry structure and are produced by transforming XML data with XSLT templates.

XML is very flexible markup language and XML databases support its extensibility. It is possible to store any valid XML document in the XML database, even mix documents with different XML structure in one database. Of course, the application has to "know" the structure of the documents (DTD) to provide search, browsing and editing functions. The DEB platform core offers browsing and entry searching without the need to modify the application.

Below the short example of the XML structure for the dictionary entry "ТОВАР" is shown:

```
<entry>
 <hw>ТОВАР</hw>
 <morph>-а</morph>
 <gram>м.</gram>
 <style>Эконом.</style>
 <sense n="1">
  <def>Продукт труда, произведенный для продажи</def>
  <exm>Товар есть, во-1-х, вещь, удовлетворяющая какой-либо потребности человека, во-2-х,
       вещь, обмениваемая на другую вещь</exm>
  <col>Живой товар, Показать товар лицом</col>
 </sense>
</entry>
```

Conclusion

In the paper we have presented the outline of the project which allows to create a complex database of a number of Russian dictionaries. The data contained in it could serve for different purposes: for presentation of dictionaries as a whole via browsers, for facilitating of lexicographers' work and as a source for different applications in the field of Natural Language Processing.

References

[1] Tolkovyj slovar' russkogo jazyka v 4 tomakh. Ed. by D.N. Ushakov. (1935-1940).
[2] Ozhegov, S.I. (1949). Slovar' russkogo jazyka. Ed. by S.P. Obnorskij, Moscow.
[3] Slovar' sovremennogo russkogo literaturnogo jazyka v 17 tomakh. Ed. by A.M. Babkin, S.G. Barkhudarov, P.P. Philin. (1948-1965). Moscow, Leningrad, toma 1-17. (Widely used abbreviation BAS).
[4] Slovar' russkogo jazyka v 4 tomakh. Ed. by A.P. Jevgen'jeva. (1957-1961). Moscow, toma 1–4. (Widely used abbreviation MAS).
[5] Komplexsnyj normativnyj slovar' sovremennogo russkogo jazyka. – In print.
[6] Bol'shoj Akademicheskyj slovar' russkogo jazyka Rossijskoj akademii nauk. Ed. By A.S. Gerd. (2005-2008). St. Petersburg, toma 1-10.
[7] Celex. (2004). CELEX Lexical Database `http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html`
[8] Klímová, J., Oliva, K., Pala, K. (2005). Czech Lexical Database – First Stage. In *Papers in Computational Lexicography, Complex'05 organized by Research Institute of Linguistics, Hungarian Academy of Sciences* Budapest 2005, 142-151.
[9] Pala, K., Horák, A., Rambousek, A., Rangélova, A. (2007). Nové nástroje pro českou lexikografii – DEB2. In Šticha, F., Šimandl, J. (ed.), Gramatika a korpus / Grammar & Corpora 2005. Praha: ÚJČ AV ČR, 190-196.
[10] Slovar' russkogo jazyka XVIII veka. Ed. by Ju.S. Sorokin. (Since 1984) Leningrad-St. Peterburg.
[11] Horak, A., Pala, K., Rambousek, A., Rychly, P. (2006). New clients for dictionary writing on the DEB platform. In DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems. Torino, Italy: Lexical Computing Ltd., U.K., 17–23.

Form, Its Meaning, and Dictionary Entries*

Violetta Koseska-Toszewa

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw

Abstract. As we know, a language form is a unit which plays a specific form in the language, e.g. a semantic or syntactical one. We establish the function of a form based on its use (occurrence), i.e. its relation with the meanings of other forms in speech or in a text. The meaning of a form is the value of its function. In the traditional grammar, form is opposed to its meaning. However, various grammar schools have big problems with distinguishing between a form and its function. For example, the present tense form has a number of basic temporal meanings in Bulgarian as well as in Polish and Russian, and in none of those languages this is only the present time, (see past, future and habituality expressed using the present tense form). It is a big mistake not to distinguish between the meanings of article in article languages. For example, in Bulgarian the same form of article can express both uniqueness and universality (or, respectively: definiteness and indefiniteness). In the quoted book (Koseska-Toszewa 1982), I put forward a hypothesis on the development of the meaning of Bulgarian article. In my opinion, initially the article expressed uniqueness of an element (object), and then started to express also uniqueness of a set, which later, due to equalling two completely different semantically-logical structures, i.e. structures with universal and unique quantification, lead to a homonymy and to the article expressing also universality, i.e. indefiniteness. Similarly in English, French, Rumanian or Albanian, where the same form of article can express either uniqueness or universality This proves that the above homonymy is of a general rather than typological (e.g. Balkan) character. Naturally, in the above languages the definite article form can also express uniqueness of an object or a set, so it also expresses definiteness. Ambiguity of the definite article form is a phenomenon exceeding the area of Balkan languages, and the only Balkanism is the position of the article – speaking more precisely, its postpositiveness (postpositive position). However, that position gives us no right to treat it differently than the English or French article. In Bulgarian, Rumanian and Albanian the postpositive article is written together with the name its concerns, but it is neither a unit belonging to the root of the word nor the ending of the word.
The above observations, based first of all on the semantically-logical aspects of the definiteness category, have been confirmed by the language material from the Suprasl Code, where Bulgarian article does not occur in universally quantified nominal structures, but in uniquely quantified nominal expressions, denoting satisfaction of the predicate either by one element of the sentence or by the whole set treated as the only one.
It is worth stressing that distinguishing between the form and its meaning in comparing the material 6 languages belonging to three different groups of Slavic languages (as is the case in the MONDILEX Project) will allow us to avoid numeorus substantiva mistakes and erroneous conclusions. Hence dictionary entries should be verified and made uniform in that respect before they are "digitalized"... Distinction between the form and its meaning in a dictionary entry is fully possible, as shown by works of Z. Saloni (Saloni 2002) and A. Przepiórkowski (Przepiórkowski 2008).

Introduction

Linguistics is a broad and already well-developed theoretically knowledge area. To elaborate the system of some language according to the contemporary linguistic knowledge, it is not enough to know that language. Hence in what follows I will deal with examples which show the pitfalls leading to errors in descriptions of language structures – in order to help avoid them.

1 Language form. Function. Value of a function. Meaning of a form.

As we know, language form is a language unit which plays a specific form in the language, e.g. a semantic or syntactical one. We establish the function of a form based on its use (occurrence), i.e. its relations with

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

meanings of other forms in speech or in a text. The meaning of a form is the value of its function. In the traditional grammar, form is opposed to its meaning. However, various grammar schools have big problems with distinguishing between a form and its function.

According to my experience, it is the grammar schools in southern Slavic countries, and – more broadly – grammar schools in the Balkans that have the most troubles with distinguishing between a language form and its meaning. The grammars which have found themselves under the influence of structuralism in language studies fare much better. Without coming into much detail, let me quote here works by J. Bauddouine de Courtenay, known already in the 19th century, those by J. Kuryłowicz, dating from early 20th century, works of the famous Prague school of structuralism, R. Jacobson's works from the 1960s, and many others.

Let me begin with examples from the traditional academic grammar of Bulgarian concerning aspect, tense and the definiteness/indefiniteness category.

2 Aspect of a verb

I will consider aspect and the problem of its classification as a specific language category in connection with analysis of temporal issues in Bulgarian. The following deliberations are of a fragmentary character. In the literature on that subject discussing the issue of aspect of Slavic verbs, there is no unique answer to the question: What is the aspect? In his fundamental work on aspect in Bulgarian, Masłow makes the reservation that he is not considering aspect as a "lexically-grammatical or word formation category, but as a solely grammatical category" (Maslov 1963). The notion of a "grammatical category" itself is adopted in different ways in linguistics, so there is no unequivocal answer either to the question whether aspect is a grammatical category or not. (Piernikarski 1989: 10). Some Czech and Slovak linguists treat aspect as a "grammatical category" as well (I. Poldauf 1964). A similar approach is adopted by W. Śmiech, according to whom aspect is a grammatical category which consists in the fact that each verb is either perfective or imperfective in all its mode and tense variants (Śmiech 1971: 5,6). In turn, A. Isachenko assumes that aspect is a lower morphological category (Isachenko 1966: 26). Further, a Polish scientist Z. Stieber is of the opinion that the aspect category can hardly be considered as an inflected category (Stieber 1973: 9). The opposition between perfective and imperfective verbs is, according to him, expressed both in the pra-Slavic language and in present-day Slavic languages with word formation means rather than inflected means. A. Heinz, Z. Gołąb and K. Polański define aspect as a morphologically-inflected category of a verb which expresses the semantic opposition between perfectiveness and imperfectiveness (Z. Gołąb, A. Heinz, K. Polański 1968). J. Kuryłowicz states the semantic character of the aspect category, which in his opinion has been built on the previousness category (Kuryłowicz 1972: 93-98). It is the semantic category of previousness which is the feature of all languages, while verbal aspect is only known to some of them. We know that it is a property of Slavic languages, which is opposed to other Indo-European languages, e.g. Latin and Greek, where perfectiveness and imperfectiveness are expressed as an opposition based on inflection (Safarewicz 1947: 198). In a work of exceptional importance for aspect-related issues in Bulgarian, S. Ivanchev brings up all problems concerning aspect in literary Bulgarian against the background of other Slavic languages, arguing that aspect of a Bulgarian verb is a live category (Ivanchev 1971: 3-246) In the author's opinion, aspect exhibits complicated morpho-semantic relationships in the contemporary literary language. In the context of those problems, Ivanchev develops a proposal for a new temporal model for the system of Bulgarian, rejecting the theory of absolute and non-absolute (relative) tenses adopted in the literature on temporal meanings of the verb. Up to that time, this was the way tenses were treated in the academic grammar of Bulgarian, see (Penchev 1967: 134), (Koseska 1972: 233-245)

The classification of tenses into absolute and non-absolute (relative) was most probably tailoured to languages which possess the previousness category but do not possess a formalized aspect category (see e.g. French imperfait = present dans le passé (Stankov 1969). Such a classification is underlain by the semantic category of previousness. However, in Slavic languages, where the aspect category is a grammatically developed one, classification of tenses into absolute and relative ones fails to explain temporal relations in a satisfactory way, and in fact makes them more complicated. This is also the case with the theory of action types (Aktionsart) taken from German, where, in opposition to Slavic languages, there is no aspect category, so its transfer to any Slavic language is unjustified. In Aktionsart, the division of verbs into action

types is not disjoint, and the individual verb types often overlap. Since from the semantic viewpoint aspect is also a type of action, then what is the difference between aspect and other kinds of action types? While in German, and maybe in Germanic languages, this kind of theory has some application in classification of verb meanings, in Slavic languages, where aspect is a live, developing category, the above theory has no proper application. In the Contrastive Bulgarian-Polish Grammar, aspect is treated as a semantic category, and in order not to confuse the form of aspect with its contents, we write there about the "semantic category of aspect", see (Karolak 2008).

3 Aspect and tense

Regardless of whether aspect is a grammatical, morphological or semantic category, it cannot be disregarded during the analysis of temporal relations, especially in Bulgarian. This fact is an argument in the discussion between Bulgarian linguists representing the so-called temporal school with representatives of the so-called aspect school. The temporal school is exemplified by works of L. Andrejchin, V. Stankov, M. Dejanov, and the aspect school – by those of J. Maslov, E. Demina, S. Ivanchev. Since we know that in the languages with aspect there are few tenses, like in north-Slavic languages, while languages devoid of aspect have a higher number of them (like Latin or French), we could expect that in southern Slavic languages there are two tendencies: one going towards reducing the number of tenses (as in Serbian and Croatian), and a second one, connected with gradual disappearance (or underdevelopment) of aspect, and maintaining a large number of tenses. This tendency has been searched for e.g. in Bulgarian. However, the aspect category still exists in the eastern group of southern Slavic languages, and yet the number of tenses in those languages does not decrease. Southern Slavic languages, and especially their eastern group, from the typological viewpoint represent the transitional stage between Greek and Latin on the one hand (large number of tenses, absence of the aspect category) and northern Slavic languages (aspect category, small number of tenses) on the other hand. This is why the problems of temporal relations in southern Slavic lands are especially important both for explaining the Slavic aspect category and for the semantics of tenses in Slavic languages.

Consequently, we should recall the thesis of S. Ivanchev (Ivanchev, op. cit.: 129), who claims that there is a genetic connection between imperfectiveness and imperfectum. He considers the aorist : imperfectum relation not as a temporal or aspectual one, but as a joint temporally-aspectual relation.

In Serbian, the imperfectum form could only be built for imperfective verbs and had a clearly aspectual character, in opposition to the Serbian aorist form, which could be not only perfective, but also imperfective (though very rarely) (Vuković 1967: 276-313).

The language facts from old Bulgarian sources confirm that the ratio of imperfectum forms of perfective verbs to imperfectum forms of imperfective verbs was 1:99 (Dostál 1954). Based on this, some scholars consider the bi-aspectual character of the aorist and imperfectum forms as an archaic state of things (E. Kosechemieder 1963: 19). However, in southern Slavic languages, and especially in the Bulgarian-Macedonian area, this state is a live one, and it is not transient at the given stage of language development.

4 Semantic category of time

The connections between aspect and temporality in southern Slavic languages (except for Slovenian) confirm Kuryłowicz's thesis about the semantic character of aspect (K. Feleszko, V. Koseska-Toszewa, I. Sawicka 1974: 183-187). In turn, Gołąb, Heinz and Polański when considering the notions of aspect and its strict connection with the category of time propose a diagram which fully explains the differences in meaning between both categories. This reduces to the fact that exponents of time position a given action with respect to the speech state (the so-called moment of speaking), while a exponents of aspect position the same action with respect to the point which represents the moment of ending the action, regardless of the speech state, see (Z. Gołąb, A. Heinz, K. Polański, op. cit.), (Koseska-Toszewa 1974: 213-226).

By the semantic category of time I mean a category that orders states and events with respect to the speech state by using the previousness-successiveness relation (Koseska 2007). For the basic notions of time – states and events as elements of temporality, see A. Mazurkiewicz 1986). For example, the praesens form (present tense form) has a number of basic temporal meanings in Bulgarian as well as in Polish and

Russian, and in none of those languages this is only the present time, see (Grochowski 1972), (Koseska 1977). In those languages, the present tense form denotes:

1. present time:
Bulg. Анета спи в моята стая.
Pol. Aneta śpi w moim pokoju.
Russ. Анета спит в моей комнате.

2. future time
Bulg. Утре идвам в два, а не в три часа.
Pol. Jutro przychodzę o drugiej, nie o trzeciej.
Russ. Я завтра прихожу в два, а не в три часа.

3. past time
Bulg. И чак тогава той разбира своите грешки.
Pol. I dopiero wtedy on rozumie swoje błędy
Russ. И едва тогда он понимает свои ошибки.

4. habituality
Bulg. Той всеки ден са ражожда поне един час.
Pol. On codziennie spaceruje przynajmniej jedną godzinę.
Rus. Каждый день он гуляет хотя один час.

Sentences (1) are expressed in the present tense; they are indicative, and hence they have either true or false value. In this respect, sentences (1) differ from e.g. sentences (2) in the future tense, which do not have either true or false value, and hence are not indicative. Instead, they have a third value – possibility, which is a modal value. Do the sentences: *Jan ponoć teraz jest na spacerze. Ян бил сега на ражодка. / Ян уж е сега на ражодка.* refer to the present time, or are they just sentences with the present tense form? Certainly, they do not have either true or false value, and hence they cannot be sentences expressing the present time. This is evidenced by e.g. Bulgarian, where the *бил* form signals the imperceptive modality rather than the present time, see *Той сега е на ражодка*, where present tense occurs. Sentences with various types of the possibility modality, not only the imperceptive one like above, often occur with the praesens form, but can also have a third value – possibility, so during the speech state we do not know whether the described state or combination of state and events exist or not. In such a case, we cannot speak of the present time, but only of a present tense form, see e.g.:

(1) On jakoby jest złodziejem. / Той май е крадец. / Той бил крадец.

The interpretation of the above sentences as ones with the present time is a good example of a failure to distinguish between a verbal form and its temporal function. Defining the present time more precisely, it is worth stressing that the present, i.e. what is happening now according to the bearer of the speech state, should be understood as a state coexistent concurrent with the speech state. Very roughly, it can also be understood solely as a state coexistent with the speech state.

However, Bulgarian grammars commonly use statements of the type: “this is a metaphorical meaning of the present time”, though the present time is the meaning of a present tense form (Stankov 1969). Such statements often lead to speaking of another meaning of some meaning, i.e. to a tautology. Similarly, Serbian, Croatian and Slovenian grammars still distinguish between the so-called absolute and relative tenses, and do not always distinguish between a form and its meaning, see (Josip Silić, Ivo Pranjković 2005), (Toporišič 1976).

5 Semantic category of definiteness/indefiniteness

Research on the definiteness/indefiniteness category has usually reduced to describing its morphological exponents first of all in the so-called article languages. The researchers have also searched for lexical analogues corresponding to the contents of article in article-free languages. In consequence, the definiteness/indefiniteness category has been treated solely as a nominal phrase category. For many years,

this fact influenced the descriptions of the category we are interested in, which in article-free languages were often reduced solely to analysis of the meanings of pronouns. Studies have shown that the definiteness/indefiniteness category as a semantic category is expressed with various language means: lexical and morphological ones, also at the level of the verbal phrase rather than only the nominal phrase, as used to be the prevailing belief in the literature on that subject, and that this is a category of the sentence rather than of the nominal phrase (Koseska 1982).

The use of the term "definiteness" in the cases when the so-called "definite article" expressed indefiniteness, i.e. universality, was an obvious mistake, and followed just from not distinguishing between the form and its meaning. In our works, the definiteness/indefiniteness category was defined as a category with the semantic opposition: uniqueness: non-uniqueness, whereby by definiteness we mean only uniqueness of an element of a set (satisfying the predicate), and by indefiniteness – non-uniqueness (both in the sense of existentiality and of universality) (Koseska 1982), (Koseska, Gargov 1990).

In Bulgarian, the most typical morphological means for expressing uniqueness and universality in the nomen group is deemed to be the article. Its absence, i.e. morphological 0, is meaningful – it is an exponent of either existentiality or pure predication. The ambiguity of Bulgarian article is a good illustration of the difficulties encountered by a scholar studying that category during classification, here quantificational classification of natural language expressions. As I have already mentioned, in Bulgarian the same form of article can express both uniqueness and universality (or, respectively: definiteness and indefiniteness). In the already quoted book (Koseska-Toszewa 1982), I put forward a hypothesis on the development of the meaning of Bulgarian article. In my opinion, initially the article expressed uniqueness of an element (object), and then started to express also uniqueness of a set, which later, due to equalling two completely different semantically-logical structures, i.e. structures with universal and unique quantification, lead to a homonymy and to the article expressing also universality.

See:

(1) *Човек-ът е от нашето село.* / *Ten człowiek jest z naszej wsi*, where the article *-ът* expresses uniqueness of an element of a set of people.
(2) *Човек-ът е мислещо и разумно същество.* / *Każdy człowiek i tylko on jest istotą myślącą i rozsądną*, where the article *-ът* expresses uniqueness of a set. (Only the set of people satisfies the predicate: *x is a thinking and rational being*).
(3) *Човек-ът е смъртен. Człowiek jest śmiertelny*, where the article *-ът* expresses universality.

Not only this form of Bulgarian article, but also its other forms can express both uniqueness and universality, i.e. definiteness and indefiniteness. Similarly in English, French, Rumanian or Albanian, where the same form of article can express either uniqueness or universality. This proves that the above homonymy is of a general rather than typological (e.g. Balkan) character. For details on that subject, see (Koseska 1982), (Koseska-Toszewa 1986: 25–37). Examples in which the English definite article expresses indefiniteness are discussed by Reichenbach (Reichenbach 1967: 101), who writes about the fact that the English "the" can express "universality" rather than definiteness!

Examples:

Eng. The lion is a ferocious animal 'The lion is a dangerous, wild animal'
French: Le lion est un animal feroce 'The lion is a dangerous, wild animal'
Rum. Omul este muntor 'Each man is mortal'
Alb. Qeni është mik i nijeriu 'The dog is a friend of the man'
Bulg. Човек-ът е смъртен. 'Each man is mortal'

Naturally, in the above languages the definite article form can also express uniqueness of an object or a set, so it can also expresses definiteness.

Examples:

Eng. The man closed the door
French: L'homme a ferme la porte
Rum. Omul a intrat in camera

Alb. Libri është mbi tryeze ‘(The) book is on the table’
Bulg. Човекът затвори вратата / Книгата лежи на масата

From the above examples it is evident that ambiguity of the definite article form is a phenomenon exceeding the area of Balkan languages, and the only Balkanism there is the position of the article – speaking more precisely, its postpositiveness (postpositive position). However, that position gives us no right to treat it differently than the English or French article. In Bulgarian, Rumanian and Albanian the postpositive article is written together with the name its concerns, but it is neither a unit belonging to the root of the word nor the ending of the word.

The above observations, based first of all on the semantically-logical aspects of the definiteness category, have been confirmed by the language material from the Suprasl Code, where Bulgarian article does not occur in universally quantified nominal structures, but in uniquely quantified nominal expressions, denoting satisfaction of the predicate either by one element of the sentence or by the whole set treated as the only one (Zaimov 1982: 5–9), (Koseska-Toszewa 1987).

It is worth stressing that without distinguishing between the form and its meaning, a comparison of material taken from 6 languages belonging to three different groups of Slavic languages may involve numerous substantive errors, and lead to erroneous conclusions. Hence dictionary entries should be verified and made uniform in that respect before they are “digitalized”... Distinguishing between the form and its meaning in a dictionary entry is fully possible, as shown by works of Z. Saloni (Saloni 2002) and A. Przepiórkowski (Przepiórkowski 2008)

A dictionary entry should obligatorily distinguish between a language form and its meaning. A further stage is to determine what we understand by the meaning of a given language form. This is discussed in more detail in the article by V. Koseska and A. Mazurkiewicz in this volume.

References

[1] Dostál. 1954: Dostál, *Studie o vidovém systému v staroslověnštině*, Praha 1954.
[2] Feleszko, Koseska-Toszewa, Sawicka. 1974: K. Feleszko, V. Koseska-Toszewa, I. Sawicka, *Związki aspektu z temporalnością w językach południowosłowiańskich*, SFPS XIV, 1974, 183–187.
[3] Grochowski 1972: M. Grochowski, *Znaczenia polskiego czasownika: aktualne, potencjalne, uniwersalne, w świetle kategorialnego znaczenia form „czasu teraźniejszego“*, Studia semiotyczne, t. III, Wrocław,
[4] Heinz, Z. Gołąb, K. Polański 1968: Z. Gołąb, A. Heinz, K. Polański, *Słownik terminologii językoznawczej*, Kraków 1968.
[5] Isachenko 1966: A. Isachenko, *Grammaticheskij stroj russkogo jazyka v sopostavlenii s slovackim*, Morfologija, t. I, Bratislava l966, 26.
[6] Ivanchev 1971: S. . Ivanchev, *Problemi na aspektualnostta v slavjanskite ezici*, 1971, Sofija.
[7] Karolak 2008, S. Karolak, *Semantyczna kategoria aspektu, GKBP*, t. 8, Warszawa 2008.
[8] Koseska-Toszewa 1974: V. Koseska-Toszewa, *Z problematyki temporalno-aspektowej w języku bułgarskim (relacja imperfectum : aoryst)*, SFPS XIV, 1974, 213–226.
[9] Koseska 1977: V. Koseska, *System temporalny północno-zachodniej gwary bułgarskiej na tle języka literackiego*, SFPS XII, 1972, 233–245.
[10] Koseska-Toszewa, 1982: V. Koseska, *Semantyczne aspekty kategorii określoności / nieokreśloności (na materiale języka bułgarskiego, polskiego i rosyjskiego)*, Wrocław, Ossolineum, 1982
[11] Koseska 2007: V. Koseska, *Semantyczna kategoria czasu, GKBP*, SOW, Warszawa, 2007
[12] Kosechemieder 1967: E. Kosechemieder, *Aspekt und Zeit*, Opera Slavica IV, Göttingen, 1963, 19.
[13] Kuryłowicz 1972: J. Kuryłowicz, *Miejsce aspektu w systemie koniugacyjnym. In: Symbolae polonicae in honorem Stanislai Jodlowski*; Wrocław 1972, 93–98.
[14] Maslov 1963, Maslov: *Morfologija glagol’nogo vida v sovremennom bolgarskom jazyke*, Moskva-Leningrad 1963, 3.
[15] Penchev 1972: J. Penchev, *Kăm văprosa na vremenata v săvremennija bălgarski ezik*, Bălgarski ezik XVII, 1967, 134.
[16] Piernikarski 1989: C. Piernikarski, *Typy opozycji aspektowych czasownika polskiego na tle słowiańskim*, Wrocław 1989, 10.

[17] Pokdauf 1964: I. Poldauf, *Podíl mluvnice a nauky o slovníku na problematice slovesného vidu* In: Studie a práce linguistické. In: Sborník k šedesátým narozeninám akademika B. Havránka, Praha 1964, 204–205.
[18] Przepiórkowski 2008: A. Przepiórkowski, *Powierzchniowe przetwarzanie języka polskiego*, Warszawa 2008.
[19] Safarewicz 1947: J. Safarewicz, *O wyrażaniu dokonaności i niedokonaności w języku łacińskim*, Eos 41, 1940–1946, z. I,2(1947),198.
[20] Saloni 2002: Z. Saloni, *Czasownik polski*, Wiedza Powszechna, Warszawa 2002.
[21] Silić, Pranjković 2005: Josip Silić, Ivo Pranjković, *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*, Školska Knjiga, Zagreb 2005.
[22] Stankov 1969 : V. Stankov, *Bălgarskite glagolni vremena*, Sofija 1969.
[23] Stieber 1973: Z. Stieber, *Zarys gramatyki porównawczej języków słowiańskich, Fleksja werbalna*, Warszawa 1973, 9.
[24] Śmiech 1971: W. Śmiech, *Funkcje aspektów czasownikowych we współczesnym języku ogólnopolskim*, Łódź 1971, 5, 6.
[25] Toporishich 1976: Jože Toporišič, *Slovenska slovnica*, Maribor 1976.
[26] Vuković 1967: J. Vuković, *Sintaksa glagola*, Sarajevo 1967, 276–313.
[27] Zaimov 1982: J. Zaimov, Uvod i komentar na starobălgarskija tekst. Suprasălski ili Retkov sbornik, Sofia 1982.

On the Meaning of Verbal Forms and Its Net Representation*

Violetta Koseska-Toszewa[1] and Antoni Mazurkiewicz[2]

[1] Institute of Slavic Studies, Polish Academy of Sciences, Warsaw
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw

Abstract. In the present paper we propose to construct a catalogue of temporal situations that are used in different languages by means of different linguistic formalisms. Entries to such an catalogue are thought to be (parameterized) names of temporal situations, and values corresponding to them should be descriptions of temporal situations, described as formally and precisely as possible. In the paper temporal situations are presented by the formalism of Petri nets, although any other formalism can be used for this purpose as well. Starting from the meaning of temporal situations rather than from grammatical forms makes possible to compare a wide bunch of languages with different types of temporality formalism.

1 Formalized situation description

The main difficulty of explanation or comparison of different verbal forms is the necessity of defining the situation expressed by the described forms. In this paper we propose to define a number of so-called *situation functions* that maps chosen verbs into situations corresponding to the used verbal form. There can be a number of various methods of situation describing; according to our previous papers we use the Petri net formalism describing situations in many aspects, both temporal as modal. In general, the syntax of situation function is:

$$\textit{Function\_name}(x_1, x_2, \ldots, x_n;\ p_1, p_2, \ldots, p_k) \ = \ Situation$$

where *Function_name* is the name of a verbal form, $x_1, x_2, \ldots, x_n$ are verb arguments, $p_1, p_2, \ldots, p_k$ are some auxiliary information, if necessary (as e.g. point of reference, passive or active voice indications, or other subjects of verbs), and *Situation* is the situation, to which the verbs $x_1, x_2, \ldots, x_n$ and data $p_1, p_2, \ldots, p_k$ are referring to. This reference is made by the verbal form specific for the chosen function. Schemes of actions, corresponding to verbs of languages, can consist of a number of states and/or events mutually connected.

It is worthwhile to make clear the intention for introducing situation descriptions. Such descriptions are not thought as a material for machine processing, but as a mean for understanding the meaning of sentences referring to chosen situations. To process sentences (not situations) there is a need of formal and precise meaning conveyed by them. Introducing a catalogue of situations, one can assign chosen entries of such catalogue to some (parts of) sentences subjected for processing and then create a formal basis for comparison them in different languages. It should be stressed that the sentences are subjected to processing, not positions in such catalogue. In order to make a progress in machine translation there is no escape of dealing with the meaning of sentences. The intention of this paper (and preceding ones) is to offer (at least partial) formal means to cope with this issue.

There are several possibilities of defining meaning of temporal properties of sentences. Here, we chose net description, since nets can grasp (a) difference between events and states; (b) the temporal sequencing, not only linear but also partial; (c) coexistence or exclusion of some parts of situations; (d) choice of different possibilities, accomplished or not; (e) some aspects of modality; (f) language independency. We are aware of existence of other possibilities of situation description and of shortcomings or incompleteness of our approach; however, we are convinced that our proposal is a step in proper direction. Clearly, the

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

introduced formalism can be subjected to further completions and improvements; for the time being, we limit ourselves to Petri nets formalism with some net elements marked, if necessary.

2 Situation functions

Situation descriptions by nets consist of net schemes (using circles for representing states, boxes for representing events, and arrows for representing sequencing). The state of speech is marked with a dot. Some net elements can be marked with symbols of variables that are provided for representing actual actions, states, or events while the function is used. A number of net elements can be marked with the same variable, if this variable refers to all of them; on the other hand, some net elements can be left unmarked, if they serve for a proper sequencing and the scheme building only. In what follows some examples of situation functions usage is presented, for situations that are used most frequently.

3 Present tense

A simple example of a situation function is function *Pr*(x) corresponding to the present tense. This function takes verb x and returns the situation given in Fig. 1. The only verb variable occurring in the scheme is x; one can substitute for it different concrete verbs. The scheme described the situation with action determined by x is being performed when the speaker is telling about it. Moreover, the beginning and ending of the speaker statement occur while x is holding. It means that during the whole act of utterance the action x (or a state described by it) is holding.

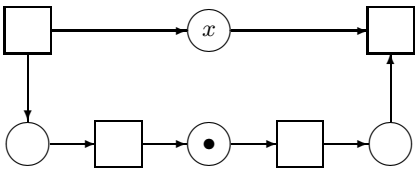


Fig. 1. *Pr*(x)

Linguistic examples of *Pr(x)* for x = ‘to read’ are:

English *He is reading a book (now)*
Bulgarian *Той (точно сега) чете книга*
Polish *On (teraz) czyta książkę*
Russian *Он (именно сейчас) читает книгу*

4 Past Perfective tense

The value of *Pp*(x) function (corresponding to Past perfective tense) is the situation where x expresses an activity completed before the state of utterance. In Bulgarian this situation is described by the aorist form of perfective verbs, in Polish and Russian by the praeteritum form of perfective verbs. The situation function *Pp*(x) is presented in Fig. 2.

Linguistic examples of *Pp*(x) for x = ‘to open’ are:

Bulgarian *Мария вчера отвори вратата*
English *Mary opened the door yesterday*
Polish *Maria otworzyła wczoraj te drzwi*
Russian *Мария открыла вчера эту дверь*

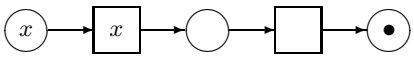


Fig. 2. *Pp*(x)

The result of action x may hold or may not hold at the state of utterance. Observe also that the speaker refers to x together with its termination, i.e. to the perfective version of action x.

5 Past Perfective Resultative tense

Similarly to the Past Perfective tense, the Past Perfective Resultative tense expresses an action terminated before the state of utterance, but now, in contrast to the above mentioned tense, with a result coexistent with the utterance state. In Bulgarian this tense is expressed by the perfectum form of perfective verbs, in Polish and Russian by the praeteritum form of perfective verbs. This tense is corresponding to situation function *Rpp*(x, y) defined in Figure 3 below.

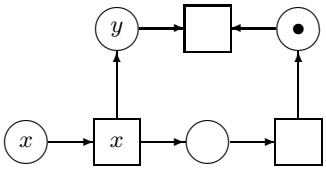


Fig. 3. *Rpp*(x, y)

Variable x is used for the verb defining the action in question, y represents its effect. Observe that the state y and the state of utterance are coexistent, as terminated by a common (anonymous) event. Linguistic examples of *Rpp*(x, y) for x = 'to open' and y = 'is open' are:

Bulgarian *Мария вече отвори вратата (вратата е отворена)*
English *Mary already opened the door (the door is open)*
Polish *Maria już otworzyła te drzwi (drzwi są otwarte)*
Russian *Мария уже открыла эту дверь (дверь открыта)*

6 Past Imperfective tense

This tense is used to describe situations similar to those expressed by Past Perfective, but without reference to the moment of the action termination; it may happen that before the state of utterance such a moment will never occur, or at least the speaker is not aware about that. The corresponding situation is the value of function $PImp(x)$ presented in Figure 4.

Linguistic examples of such situations are:

Bulgarian *Мария отваря́ тази врата*
English *Mary was opening the door*
Polish *Maria otwierała te drzwi*
Russian *Мария открывала эту дверь*

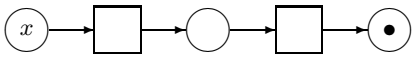


Fig. 4. $PImp(x)$

7 Past Imperfective Resultative tense

The value of *Irp*(x, y) for verbs x and y (corresponding to Imperfective Resultative Past tense) is the situation where the action x takes place before the state of utterance, but the state y resulting in effect of action x is coexistent with the state of utterance (Figure 5). The speaker does not refer to the completion of action x but, instead, to the result y of this action. In bulgarian this situation is expressed by form *Perfectum* of imperfective verbs, in Polish and Russian by form *Praeteritum* of imperfective verbs.

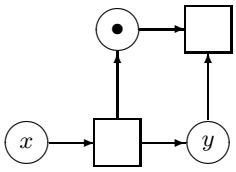


Fig. 5. *Irp*(x, y)

Linguistic examples of *Irp*(x, y) for x = 'to be have influenza' ('to write poems') and for y = 'to cough' ('possible to be read') are:

Bulgarian	*Той е боледувал от грип (и сега кашля)*
	На младини Мария е писала стихове (можеш да ги прочетеш)
English	*He had influenza (and he is coughing now)*
	Mary was writing poems in her youth (you can read them now)
Polish	*On chorował na grypę (i teraz kaszle)*
	W młodości Maria pisała wiersze (możesz je przeczytać)
Russian	*Он болел гриппом (у него теперь кашель)*
	В молодости Мария писала стихи (можешь прочитать их)

8 Conclusions

In the present paper we argue for (1) creating a catalogue of temporal situations that can be useful for comparison, analyzing, processing, or translating phrases in different languages containing temporal dependencies; (2) distinguishing verbal forms from temporal meaning in different languages. The first aim results from a need of proper understanding temporal statements in various languages; without understanding their proper meaning one is not able to compare them or to create a reliable correspondence between them. The second objective follows from the fact that the same or similar verbal forms in different languages may describe different temporal situations. Therefore we should rely on meaning rather than form while comparison phrases in different languages or trying to make their faithful translation. Some examples of different verbal forms with a similar functionality are given through the paper. In Table 1 a comparison of temporal meanings and corresponding to them verbal forms, discussed in the paper, are given. In Table 2 we list some situation functions together with their situation values.

In the present paper we limit ourselves to discuss only small part of temporal tenses used in natural languages, namely to present tense and some types of the past tenses. We hope they offer an opportunity of grasping the idea of situation functions that base on formal methods of situation description. In the future we plan to extend the domain of situation functions as well as to enrich their expressive power by introducing new information parameters and by improving their formalism.

Temporal meaning	**Verbal form**
Present	Present tense form (Eng., Bul., Pol., Rus.)
Past Perfective	Past Pefective form (Eng.) Aorist perfective form (Bulg.) Praeteritum of perfective verbs (Pol., Rus.)
Past Perfective Resultative	Past Perfective form (Eng.) Perfectum form of perfective verbs (Bulg.) Praeteritum form of perfective verbs (Pol., Rus.)
Past Imperfective	Past continuous (Eng.) Aorist form of imperfective verbs (Bulg.) Praeteritum form of imperfective verbs (Pol., Rus.)
Past Imperfective resultative	Perfective Continuous (Eng.) Perfectum of imperfective verbs (Bulg.) Praeteritum of imperfective verbs (Pol., Rus.)

Table 1. Comparison of temporal meanings and corresponding verbal forms

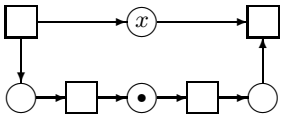
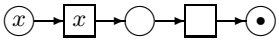
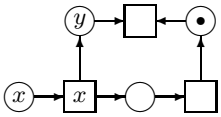
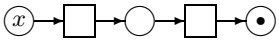
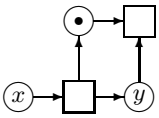
Entry	Situation	Meaning
$Pr(x)$		*Present*
$Pp(x)$		*Past Perfective*
$Rpp(x,y)$		*Past Perfective Resultative*
$PImp(x)$		*Past Imperfective*
$Irp(x,y)$		*Past Imperfective Resultative*

Table 2. Sample of situation function entries

References

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

General Architecture and Lexical Entry Structure of the Polish-Ukrainian Electronic Dictionary*

Natalia Kotsyba[1] and Igor Shevchenko[2]

[1] Institute of Slavic Studies, Polish Academy of Sciences
[2] Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine

Abstract. The paper describes the process of digitalization and further processing of a Polish-Ukrainian electronic dictionary, its technical and linguistic preparation for future lexicographic works, mainly: post-OCR problems and ways of their automatic correction, conversion of the dictionary file into a database; defining the core set of lexical entries with the help of frequency lists; lexical entry parsing procedure, automatic dictionary direction reversal. The approach presented here aims at producing an updated dictionary as well as a lexicographic editing environment and a tool set for further expansion and modification of the bilingual dictionary.

1 Introduction

Polish-Ukrainian lexicography, both paper and electronic, is represented nowadays by numerous small- or average-size dictionaries created on the basis of earlier paper editions with the addition of the most frequently used, essential new terminology covering the spheres of business, economy and tourism. An extensive review of existing Polish-Ukrainian lexicographic resources with their quality analysis – the macrostructure (choice of entries) and microstructure (entry content and design) – is presented in [1]. During the four years since the appearance of that publication, several new sources that deserve our attention became available. ABBYY Lingvo included a Polish↔Ukrainian dictionary in its version x.3 (2008) [5]. It is based on a modern paper edition and counts ca. 42000 words.[1] Trident Software Electronic Dictionary and Translator [3] includes the Polish↔Ukrainian language pair. Unfortunately no information about the sources and size of the dictionary is provided, and the project is commercial. Considerable progress, as compared to its state in 2005, can be seen in the development of the Multilingual Dictionary by Valentyn Solomko (updated in 2008), which is generated automatically from bitexts [6]. Dictionaries for each language pair in the MS Excel file format are available for download under GNU General Public License. The Polish-Ukrainian file contains 65000 words or word combinations with one-to-one correspondence of translation equivalents. This dictionary can be helpful for machine processing, but it is not particularly human-friendly. Summing up, as far as the size and the quality of entry description is concerned, there is still a need for a large modern electronic and freely available Polish↔Ukrainian dictionary suitable for both public use and linguistic research.

2 From paper to digital version, preparing dictionary background

A large electronic Polish-Ukrainian dictionary was developed by a joint group of linguists of the Institute of Slavic Studies of the Polish Academy of Sciences and the Ukrainian Linguistic-Informational Foundation of the National Academy of Sciences of Ukraine during 2005–2009. The basic core of the existing version of the Polish-Ukrainian electronic dictionary comes from the paper Polish-Ukrainian dictionary in two (three physical) volumes edited by Lukiya Humetska and published in Kyiv in 1958.

* The study and preparation of these results have received partial funding from the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

[1] Information about the size comes from ABBYY developers and concerns the electronic version of the dictionary.

This is the most comprehensive existing bilingual dictionary of very high lexicographic quality for Polish and Ukrainian. It contains about 100000 headwords. Since it was created half a century ago, its entry list and, sometimes, entry content are considerably outdated and do not fully reflect the modern state of both languages. Some domains (computers, finance) are not represented at all, while others (e.g., agriculture) are described in excessive detail. The dictionary is too biased ideologically, which is not surprising taking into the consideration the time and political circumstances of its appearance. Nevertheless, it is a good ground for further lexicographic works.

2.1 Technical editing

The paper dictionary was scanned and processed through the FineReader optical text recognition program in order to receive a text out of the scanned images. The resulting text was saved in the MS Word format. Its quality left much to be desired. The first edition of the dictionary file was the most tedious one and included correction of errors generated by the poor physical quality of the original paper edition and failures of the optical character recognition (OCR) proper. Some mistakes were systematic, which allowed us to apply multiple automatic replacement both in content and formatting. OCR mistakes were more numerous than in ordinary text due to the bilingual character of the dictionary using two different alphabets – Latin and Cyrillic – with several similar-looking letters; omnipresent stylistic and grammatical mark-up in an abbreviated form that is not found in standard OCR dictionaries; shortened forms with the common part replaced by the special character ~ (tilde), etc.

Grammatical and stylistic mark-up is crucial in the digitalizing process as it helps define the structure of the dictionary (see Sections 4 and 6). It is also important to preserve its original formatting (italic or boldface), as it is crucial for successful parsing. It is often impossible to visually determine whether a letter belongs to the Cyrillic or Latin alphabet, cf. “c” and “с”, “k” and “к”, “p” and “р”, as well as “a, e, i, o, y”, or Cyrillic „т” that looks like Latin „m” (*m*) in italic. Therefore, a series of heuristics was used to unify chains of letters delimited by a space to a single alphabet. For one- and two-letter abbreviations, the automatic replacement function of MS Word was used to check the consistency of alphabets and formatting. Some misreadings had a regular character and were corrected automatically as well, either in a supervised (one after another) or unsupervised way (all at once).

Examples of typical automatic substitutions (taking into account adjacent spaces as well):

v) → 1) (number of meaning)

om. → orn. (stylistic label “ornithology”)

Spelling errors were also detected by preparing a frequency list of space-delimited chains and checking the ones that contain up to five symbols and have the lowest frequency.[2] According to Zipf’s law, these are candidates for misspellings. Even though such automatization facilitated the editing work considerably, much labour remained to be done by hand.

2.2 Preliminary edition of the content

While editing the technical side of the dictionary it was impossible to ignore its content either. The two peculiarities of this dictionary are that it was overloaded with Soviet ideology and contained an unforgivable number of Russisms (Polonisms were met more rarely). These were removed from the file and replaced with more neutral and literary correspondents respectively. All the changes were recorded into a separate file. Below are some examples of ideologically biased entries.

[2] Another option, suggested by Janusz Bień, could be the use of the programme Kolokacje („Collocations”) by Aleksander Buczyński that can help detect unusual word combinations and in this way find words with wrong spelling. We did not experiment with it, though.

"Party" words[3]:

partyjny ("belonging to the party"). It is supplied with excessive examples of use and the party is understood as the Communist Party of the USSR in all usages: *aktyw* ~ партійний актив, -ву (партактйв); *grupa ~na* партійна група (партгрупа); *komitet* ~ партійний комітет, -ту (парт-ком, парткомітет); *konferencja ~na* партійна конференція (партконференція); *l є g i t y m a-c i a ~na* партійний квиток, (партквиток); партійний працівник, -ка (парт-працівнйк); *praca ~na* партійна робота (партробота); *staż* ~ партійний стаж, -жу (партстаж); *szkolą ~na* партійна школа (партшкола); *zebranir л:е парт,пні* .. к>ри, -рів (партзбори); *zjazd* ~ партійний з'їзд, -ду (партз'їзд): ("activists, group, committee, conference, membership card, worker, work, experience, school, meeting, congress").

The derivation for *partia* ("party") in its political sense is also overrepresented: *partyjność* ("the state of belonging to the Party"), *POP (Partyjna Organizacja Podstawowa) skr.* первинна партійна організація ("primary party organization"), etc.

"Anti" words:

przeciwsocjalistyczny антисоціалістичний ("antisocialistic"); *przeciwreligijny* антирелігійний ("antireligious"); *przeciwrepublikański* антиреспубліканський ("antirepublican"); *przeciwżydowski* антиєврейський ("anti-Jewish"); *przedkołchozowy* доколгоспний ("pre-kolkhoz"); *okres ~ od socjalizmu do komunizmu* перехідний період від соціалізму до комунізму ("the transferring period from socialism to communism"); ~ *rewolucji burżuazyj-no-demokratycznej w socjalistyczną* перероставання буржуазно-демократйчної революції в соціалістичну ("transformation of the bourgeois-democratic revolution into the socialistic"); *~dy burżuazyjne* буржуазні передсуди, -дів ("bourgeois prejudicies"); etc.

Russisms were used not only as translation equivalents, there were many of them in additional explanations of use, etc. Below are examples in the following format: *Russism → literary_Ukrainian_word (Russian_literary_equivalents) "English_translation".

*нуждатися → мати потребу/потребувати (нуждаться) "have a need"; *могучість → могутність/міць (могущество) "power"; *вірьовка → мотузка/шнур (веревка) "rope"; *лагер → табір (лагерь) "camp"; міліцейський *участок → дільниця (участок) "police station; lot"; *похожий → подібний (похожий) "similar"; *сахарний → цукровий (сахарный) "sugar, adj"; *жарке → печеня (жаркое) "stowed meat"; *гравіровка *печатей → гравірування печаток (гравировка печатей) "engraving seals"; *скучний → нудний (скучный) "boring"; *плеск → плескіт (плеск) "splashing"; *покрасити → пофарбувати (покрасить) "paint, v"; *командировочні → добові/відрядні (командировочные) "travel allowance"; *полуботинок → півчобіток (полуботинок) "(kind of) shoes"; *флажок → прапорець (флажок) "flag"; *пересахарити → перецукрувати (пересахарить) "put too much sugar"; *прощитатися → прорахуватися (просчитаться) "miscalculate"; *передаточний → передавальний (передаточный) "transformational"; *снотворний → снодійний (снотворный) "soporific"; *напиток → напій (напиток) "drink, n"; *приємного апетиту! → Смачного! (приятного аппетита) "Bon appétit!"; *італьянське → італійське (итальянское) "Italian"; *ізумруд → смарагд (изумруд) "emerald"; *шокірувати → шокувати (шокировать) "shock, v"; *готовитися → готуватися (готовиться) "prepare".

[3] We also leave here the original after-OCR format to give the idea what the dictionary text looked like after scanning and text recognition.

3 Conversion to a database format

Working with the dictionary text in a text editor such as MS Word is very inconvenient, as it is impossible to directly access particular structural units of word entries, and the pace of processing large text files is very slow. This is why the dictionary was converted into a database where its structure is reflected in separate tables and their columns and rows. This was done in several steps. First, dictionary text was split into entries with the most primitive structure: the headword and the rest. This format enabled relatively convenient check and further edition of the dictionary, already as a database. After the second edition the larger part of the dictionary entry was further parsed and recorded into a more complex database (see Section 6 for details).

4 Automated detection of structural elements boundaries of the dictionary

Information about the entry word limits, defined in the original by bold font and restored in the post-OCR MS Word file, made it possible to mark the border between the headword and its explanation in the database by placing them in separate columns. The borders between lexical entries were marked by line breaks. The grammatical and stylistic information, highlighted by italics within the dictionary entry, was marked up accordingly but retained in the same column for easier edition before the final, most detailed, parsing.

To mark the boundaries of structural elements in a semi-automatic mode we used a variety of complex context-dependent substitutions which took into account punctuation, the alphabet used (Latin or Cyrillic), text formatting: regular, italic or boldface font, and the content of the word entry. In cases where the context and the printing style were insufficient to clearly identify an element, the correction was made manually.

Upon analysing the word entry structure and formal signs of structural elements, we can see the following general picture:

Left-hand part
Headword (bold, new line)
* opt. homonym ([(I, II, III, IV)], [space])
* optional (additional forms, e.g., perfect aspect forms of verbs, phonetic variations, etc.)
grammatical forms ((* opt. [hyphen], [form], [comma]), * opt. hyphen [form], space)
mark grammatical categories [sort of] for declensions ((italic, [form], * opt. (dot, comma)), italic, [form], * opt. dot)
tags of style
tags of topics and terminology
* opt. valency frame ([(], ((* opt. prepositions), forms) [)], space)
clarification / definition (__italic__: [(], [content] [)], [space])
interpretation: the basic form (Cyrillic, * opt. [[(], option ,[)], [space]], END :{[,], [;], [.]}, space)
* opt. phrases (bold: [1st part], [space], [2nd part] (* opt. [space], [3rd part]) sign [:])
* opt. verbal form "się" ([;], [space], [/ / ~ się], [space], [right side], [.])

Right-hand part
* opt. meaning number (integer, symbol [)], space)
tag style / theme and terms (italics, * opt. [* opt. (point, point)], [dot] [space])
* opt. option value ([Cyrillic: (a, b, in)] [)], [space])
opt. valency frame ([(], ((opt. prepositions), forms) [)], space)

clarification / definition (__italic__: [(], [content] [)], [space])
interpretation: the basic form (Cyrillic, * opt. [[(], option,[)], [space]], END :{[,], [;], [.]}, space)
opt. grammatical forms ((opt. [hyphen], [form], [comma]), * opt. hyphen [form], comma)
*opt. collocation examples ([;], * opt .[~], [variable part], [space], * opt. [the rest of the collocation], [space], [construction], {[;], [.]})
opt. phraseological ([;], [space], [<>], [space], * opt. [tag style])
[newline]

Here are examples of contextual replacements to identify structural elements of the word entry.

CONTEXT	REPLACEMENT PATTERN
[new line] [Latin, bold]	[new line] <Рес> [Latin, bold]
[Latin, bold], *opt.[,] space, [non-bold]	[Latin, bold] </Рес>, *opt.[,] space, [non-bold]
space, [integer], [closed bracket], space	space, <НЗн> [integer], [closed bracket], </НЗн> space
[Latin, bold], space {[I], [II], [III], [IV]} space	[Latin, bold], space <Ом>{[I], [II], [III], [IV]} </Ом> space
</Рес> space, [Latin, italic]	</Рес> space <ГрП> [Latin, italic]
{</Рес>, </Нзн>}, space, [Cyrillic]	{</Рес>,</Нзн>},space,<Екв>[Cyrillic]
</Рес>[,] space [-] [Latin bold]	</Рес> [,] space <ПСз> [-] [Latin bold]
[Cyrillic], space, [-][Cyrillic]	[Cyrillic],</Екв> <Усз>-[Cyrillic]
</НЗн> space, [(][Cyrillic italic]	</НЗн>space,<Уточ>[(][Cyrillic italic]
[Cyrillic italic], [)], space, [Cyrillic]	[Cyrillic italic], [)], space, </Уточ> <Екв> [Cyrillic regular]
</НЗн> space, [(],[Latin italic]	</НЗн>space,<ПКер>,[(],[Latin italic]
space, [див.] space, [Latin bold]	space,<Пос>[див.]</Пос>space,<Адр> [Latin bold]

Tab. 1. Examples of context replacements in the dictionary text for identification of structural elements

During the conversion some data were lost; in cases where entries were split between columns or pages this was systematic, although not too frequent. During the second edition the loose ends were added manually and further errors resulting from oversight during the first edition and parsing errors were corrected.

5 Defining the core vocabulary

Already in this simple format, the dictionary database has more functions than a simple text file, namely, we can work with the entry list of the dictionary. As the actual database resulting from the paper edition appeared too large for experimenting with lexicographic methods and producing preliminary ready-for-use results, it was decided to select a core vocabulary of ca. 30 thousand lexical entries for the pilot version of the dictionary. This selection is also the first part of the dictionary that is intended for public release for use through a web interface. The frequency parameter was chosen as the criterion of selection. A frequency list was generated from the IPI PAS corpus of the Polish language[4] with the help of the program Poliqarp 1.2, which allows for statistic reports on corpora. Since Poliqarp has restrictions on the length of query reports, a query for each part-of-speech (or a flexeme in IPIPAN Corpus tagset presentation) was run, which gave the additional advantage of supplying the frequency list with part-of-speech (POS) information.

[4] Available at `http://korpus.pl`.

In order to avoid proper names, or rather to separate them from common nouns, adjectives and nouns starting with a capital letter were excluded from the search. A typical query looks as follows: [orth="[qwertyuiopasdfghjklzxcvbnmżźćńłóęąś].*" & pos="subst"] group by base sort by freq count all.

The table below shows the distribution of types generated for a given flexeme.

Flexeme	Tag	Types
Adjective (starting with lowercase letters only)	adj	7157
Adjective (including those starting with a capital letter)	adj	7283
Adverb	adv	2762
Conjunction	conj	67
Punctuation	interp	43
Predicative	pred	19
Preposition	prep	66
Particle	qub	448
Substantive (including those starting with a capital letter)	subst	19957
Substantive (starting with lowercase letters only)	subst	16798
Verb	verb	12411
Verb (together with gerunds)	verb	12546
Sum (without proper name candidates and gerunds)		39771

Tab. 2. Distribution of flexeme types

Gerunds, or so-called *-nie* forms, are treated in the IPI PAS corpus in a special way. They are included to both 'verb' and 'noun' categories, and their lemma is identical with the infinitive of the corresponding verb. Polish gerunds are an important part of the vocabulary; they are used more widely than their formal Ukrainian correspondents. However, their formation is not completely regular: they are often homonymous with abstract nouns. Their list was extracted from the corpus on the basis of the ending *nie. This list had to be manually cleaned afterwards.

In general, the procedure of extracting the lexicon basing on the frequency criterion gave us the following advantages: singling out words of low frequency that were included into the original dictionary version; receiving a list of words of high frequency that was not included into the original dictionary version. This information gives valuable information for further manipulation with the lexicon. For example, Polish words that were not found in the IPI PAS corpus at all (or received a minimal frequency rank) but whose Ukrainian equivalents receive high frequency rank in the Ukrainian corpus call for revision as suspects for archaisms. This is the case with Polish *obuwać*, *obuć*[5], *rozzuwać się*, *prześpiewanie*, *zakipieć*, etc.

Inter-POS homonymy was accounted for due to POS limitation of the search, while intra-POS homonymy had to be ignored—the same frequency value was assigned for all homonyms within the same part of speech.

[5] There are 21 uses of forms lemmatized *obuć* "put on shoes" in the IPI PAS corpus, 19 of them are participles form *obuty*, still in wide use, and only two are finite past verb forms *obuł*, both from a novel written in 1985. No occurrence of its aspectual counterpart *obuwać* has been found at all.

6 Parsing the lexical entry and recording it in a lexicographic database

The next step of the work is a proper lexical entry parsing that enables creating a lexicographic editing tool. The selection of the structural elements of the dictionary is carried out according to the original lexical entry design. Polygraphic formatting peculiarities can be used for automatic identification of text structure. In order to convert the primitive table into a lexicographic database, special labels are defined to mark the beginning and the end of entries' structural parts. The following formal boundaries of structural elements have been detected from the analysis of text entries.

STRUCTURAL ELEMENTS	LABEL
Polish register unit (word or phrase)	Рее
Grammatical and semantic properties of a word equivalent	ГрПа
Homonym number	Ом
Meaning number	НЗн
Ukrainian equivalent word	Екв
Polish inflectional element	ПСз
Ukrainian inflectional element	Усз
Grammatical and semantic properties of a word equivalent	ГЕк
Phrase (collocation)	Кол
Polish prepositional agreement element	Пкер
Ukrainian prepositional agreement element	Укер
Phraseology label	Фрз
Reference label	Пос
Comparison label	Пор
Reference address	Адр
Specification of meaning	Уточ
Additional form (phonetic variant or verb aspect match elements)	Дод

Tab. 3. Structural elements of words, and their labels.

In comparison with monolingual dictionaries, the bilingual dictionary has more a complex and specific structure. The main difference is that the explanatory dictionary in its left-hand part describes formal elements of the lexical unit and in its right-hand part deals with the content, its semantic elements. Therefore the left-hand and right-hand parts of the word entry are clearly separated one from another in (almost) all cases. The bilingual dictionary is characterised by a slightly different situation: the left-hand side of the word entry describes grammatical characteristics and semantic features of the source-language units, while the right-hand one describes the content represented by equivalents of words and phrases in another language (in our case Ukrainian). Moreover, elements of the left-hand and right-hand parts are given in a mixed order, creating a complex, intertwined structure.

6.1 Parsing steps

Let us consider a relatively simple bilingual dictionary entry:

dobry 1) добрий; **~re** słowo добре (ласкаве) слово; ludzie **~rej** woli люди доброї волі; z **~rej** woli з доброї волі, добровільно; 2) (do czego) підхожий (для чого); **~** do tej roboty підхожий для цієї роботи; 3) (na co) придатний (на що); materia **~ra** na płaszcz матерія придатна на плащ; ◊ *розм.* a to **~re!** от тобі й маєш! от тобі й на! *розм.* **~ra** nasza! наша бере!

We can see in the entry the Polish headword „dobry". Its three meanings are rendered by different Ukrainian equivalents: „добрий" („good"), „підхожий" („suitable"), „придатний" („fit"). Further we have Polish phrases (collocations) as examples of word usage, and their Ukrainian equivalents. We can notice Polish words in a truncated form in the entry, where the initial part of the word is marked with a tilde. When used independently (space- or punctuation-separated mode) the tilde indicates the register word as a whole. Besides, in the above example there are tags for prepositional agreement with appropriate values, both of the Polish entry word and its Ukrainian equivalents, phraseological label ◊ , stylistic tags like *розм.* and so on.

Having replaced polygraphic formatting marks with explicit labels – HTML tags for boldface and/or italic fonts – we can get the entry to look as shown below. The dictionary text that was marked up in this way became the ground for further automatic entry parsing and additional tagging of the structural elements:

<B>dobry</B> 1) добрий; <B>~</B>re słowo добре (ласкаве) слово; ludzie <B>~rej</B> woli люди доброї волі; z <B>~rej</B> woli з доброї волі, добровільно; 2) (do czego) підхожий (для чого); <B>~</B> do tej roboty підхожий для цієї роботи; 3) (na co) придатний (на що); materia <B>~ra</B> na płaszcz матерія придатна на плащ; ◊ <I>розм.</I> a to <B>~re!</B> от тобі й маєш! от тобі й на! <I>розм.</I> <B>~ra</B> nasza! наша бере!

After the rearrangement of the labels by means of complex contextual replacements we receive the following structural elements in a linear form with explicit marking of the limits (beginning and end) of all structural elements of the entry:

<Реє><B>dobry</B></Реє> <НЗн>1)</НЗн> <Екв>добрий</Екв>; <Кол><B>~</B>re słowo</Кол> <Екв>добре (ласкаве) слово</Екв>; <Кол>ludzie <B>~rej</B> woli</Кол> <Екв>люди доброї волі</Екв>; <Кол>z <B>~rej</B> woli</Кол> <Екв>з доброї волі, добровільно</Екв>; <НЗн>2)</НЗн> <ПКер>(do czego) </ПКер> <Екв>підхожий</Екв> (для чого); <Кол><B>~</B> do tej roboty</Кол> <Екв>підхожий для цієї роботи</Екв>; <НЗн>3) </НЗн> <ПКер> (na co) </ПКер> <Екв>придатний</Екв> <УКер> (на що) </УКер>; materia <B>~ra</B> na płaszcz <Екв>матерія придатна на плащ</Екв>; <Фрз>◊ </Фрз> <ГрП><I>розм.</I></ГрП> <Кол>a to <B>~re!</B></Кол> <Екв>от тобі й маєш! от тобі й на!</Екв> <ГрП><I>розм.</I></ГрП> <Кол><B>~ra</B> nasza!</Кол> <Екв>наша бере!</Екв>

The linear format can be further split into a hierarchical tree on the basis of links between entry elements. The figure below shows that the first meaning of the Polish headword corresponds to one Ukrainian equivalent. Additionally, three examples of collocations with the headword are given together with their Ukrainian equivalents. The phraseology zone includes two Polish phrases marked as colloquialisms, the former corresponding to two Ukrainian equivalents, and the latter only to one.

6.2 Tree-structured entry record

<<Реє><В>dobry</В></Реє>
<НЗн>1)</НЗн>
<Екв>добрий</Екв>;
<Кол><В>~</В>re słowo</Кол>
<Екв>добре (ласкаве) слово</Екв>;
<Кол>ludzie <В>~rej</В> woli</Кол>
<Екв>люди доброї волі</Екв>;
<Кол>z <В>~rej</В> woli</Кол>
<Екв>з доброї волі, добровільно</Екв>;
<НЗн>2)</НЗн>
<ПКер>(do czego) </ПКер>
<Екв>підхожий</Екв> (для чого);
<Кол><В>~</В> do tej roboty</Кол>
<Екв>підхожий для цієї роботи</Екв>;
<НЗн>3) </НЗн>
<ПКер> (na co) </ПКер>
<Екв>придатний</Екв>
<УКер> (на що) </УКер>;
<Кол>materia <В>~ra</В> na płaszcz</Кол>
<Екв>матерія придатна на плащ</Екв>;
<Фрз>◊ </Фрз>
<ГрП><І>розм.</І></ГрП>
<Кол>a to <В>~re!</В></Кол>
<Екв>от тобі й маєш! от тобі й на!</Екв>
<ГрП><І>розм.</І></ГрП>
<Кол><В>~ra</В> nasza!</Кол>
<Екв>наша бере!</Екв>

Fig. 1. The entry „dobry” as a tree structure.

Another example of a word entry with more structural elements:

ale 1) але; та *(рідше)*; 2) *(після заперечної частини речення)* a; nie tutaj, **~** tam не тут, а там; ◊ **~**i tak певна річ, звичайно; *прик.* nikt nie jest bez **~** немає людини без вади.

We can see here, *inter alia*, a clarification of the meaning, in this case through providing the context of usage: *після заперечної частини речення* “after the negative part of a sentence”; additional information about the frequency of use for one of the equivalents: *рідше* “more rarely”; not fully synonymous equivalents separated with a semicolon; a mark indicating a set expression, *прик.* “saying”.

Upon the replacement of the formatting tags with explicit labels this entry looks as follows:

<В>ale</В> 1) але; та <І>(рідше)</І>; 2) <І>(після заперечної частини речення)</І> a; nie tutaj, <В>~</В> tam не тут, а там; ◊ <В>~</В>i tak певна річ, звичайно; <І>прик.</І> nikt nie jest bez <В>~</В> немає людини без вади.

Upon contextual replacements inserting structural labels:

<Реє><В>ale</В></Реє> <НЗн>1)</НЗн> <Екв>але; та</Екв> <І>(рідше)</І>; <НЗн>2)</НЗн> <Уточ><І>(після заперечної частини речення)</І></Уточ> <Екв>а</Екв>; <Кол>nie tutaj, <В>~</В> tam</Кол> не тут, а там; <Фрз>◊ </Фрз> <В>~</В>i tak <Екв>певна річ, звичайно</Екв>; <Прк><І>прик.</І></Прк> <Кол>nikt nie jest bez <В>~</В></Кол> <Екв>немає людини без вади</Екв>.

6.3 Generalized structure of the word entry

Thus, a generalized structure of the word entry for the Polish-Ukrainian dictionary can be presented with certain simplification in the following way. Elements of the right-hand side of the dictionary, i.e. Ukrainian equivalents with their appropriate labels, are in italics.

Headword
Homonym number
  Inflectional elements (can recur)
    Variants or parallel forms (recurring)
  Headword variant (phonetic variant or verb aspect counterpart)
    Inflectional elements (recurring)
      Variants or parallel forms (recurring)
  Linguistic characteristics (labels for grammatical categories, style, terminology)
    Inflectional elements (recurring)
  Labels of style and/or terminology (recurring)
  Number of meaning
  Linguistic characteristics (labels of grammatical categories, style, terminology)
    Valency frame (agreement labels)
    Specification
    Word equivalent
    Inflectional elements (recurring)
    Specification of meaning
    Variants or parallel forms
      Inflectional elements (recurring)
        Specification of meaning
  Phrase (recurring)
  Phrase equivalents(recurring)
      Grammatical parameters (stylistic labels)
  Set expression (recurring
  Set expression (recurring)
      Grammatical parameters (stylistic labels)
  Verbal forms with reflexive **się**
    Inflectional elements (recurring)
      Variants or parallel forms (recurring)
    Headword variant (phonetic variant or verb aspect counterpart)
      Inflectional elements (recurring)
        Variants or parallel forms (recurring)
    Linguistic characteristics (labels of grammatical categories, style, terminology)
      Inflectional elements (recurring)
      Labels of style, terminology (recurring)

Number of meaning
Linguistic characteristics (labels of grammatical categories, style, terminology)
    Valency frame (labels)
    Specification
    Inflectional elements (recurring)
    Variants or parallel forms
        Inflectional elements
Set expression (recurring)
Phrase equivalents(recurring)
        Linguistic parameters (stylistic labels)
Set expression (recurring)
Set expression (recurring)
        Linguistic parameters (stylistic labels)

Fig. 2. Generalized tree structure of the word entry in the Polish-Ukrainian dictionary.

7 Reversing the language direction in a bilingual dictionary

It is desirable in a bilingual lexicographic system to be able to access this system not only through the source-language entry list (the left-hand part of a bilingual dictionary) but from the target-language units (the right-hand part) as well. Thus, the reversal of the bilingual dictionary so that the left-hand and the right-hand parts of the entries change places becomes another important task. The objective actually is to transform the $Language_1 \rightarrow Language_2$ dictionary into a $Language_2 \rightarrow Language_1$one. This task is far from being trivial because, as we can see, the information about the correspondence between words and word combinations of the two languages is recorded according to lexicographical tradition in a laconic, compressed form, most economic and convenient for the user. This problem is solved through "unfolding" the word entry into a set of basic equivalents, i.e., separating rows of original words or phrases and their respective equivalents in the other language, along with the corresponding grammatical, stylistic and thematic information.

The conversion of a word entry of the initial dictionary into a set of elementary equivalents requires several operations. First of all, abbreviated words with tildes are to be replaced with their full versions, i.e., „~ra", „~re", „~rej" are restored to „dobra", „dobre", „dobrej". This is done automatically by searching the first letter (after the tilde) of the shortened word in the full-form word; the search is carried out from right to left. The part from the entry word on the left of this letter gives us the string to be inserted instead of the tilde. The next step is to detect the limits of the equivalents together with their source-language counterparts. The boundary is defined due to obligatory occurrence of the equivalent expression from the target language after any source-language word or phrase. One word is often translated as several words and/or phrases. Equivalents are often presented by short synonymic rows, where synonyms are separated by commas.
A comma inside an equivalent expression often, although not always, means a limit between synonymous equivalents. Therefore, it can be used for dividing an entry into basic sets of equivalents automatically.
Here is a fragment of our sample entry **dobry**:

z **~rej** woli з доброї волі, добровільно;
with the first step it turns into the line:
z **dobrej** woli з доброї волі, добровільно,
with the second step the line is split into two more basic sets of equivalents:
z **dobrej** woli з доброї волі 1;
z **dobrej** woli добровільно 2.

The equivalent rank, taken from the order of the equivalent expression in the entry, is assigned automatically. It usually indicates a kind of priority, a higher frequency or higher standard of the translation equivalent of the entry in question. This information can be useful for further stages of work with the reverse dictionary. In our example we receive information about the priority of the translation equivalent „з доброї волі" (lit. „of one's free will") for the Polish phrase „z dobrej woli", although in general another translation equivalent, „добровільно" „voluntarily", is equally common.

Sometimes a comma inside the equivalent zone is not a sign to separate two different (synonymous) values, but is a part of an equivalent phrase, as in:

~ (ten), który to powiedział „той, який (що) це сказав"

In this case, „той, який (що) це сказав" (lit. „the person who (that) said this") is an integral equivalent. At the same time, brackets are another indicator of variability of the translation equivalent and point to a compressed translation. Thus, we have two elementary equivalents here:

ten, który to powiedział „той, який це сказав"

ten, który to powiedział „той, що це сказав"

Apart from a pair of equivalent words or phrases with the same meaning, an elementary equivalent set, as we define it, should also include various labels available for this pair. For this particular dictionary these are: grammatical category, peculiarities of morphological forms, stylistic and terminological tags, as well as an extended valency frame that also includes information about prepositional agreement. Although prepositional agreement is also a kind of valency information, a significant difference in rendering information about proper valency frames is that the former ones are given in italics, and the latter ones in regular type and, normally, in brackets. Clearly all phraseology, proverbs, etc., found in the original dictionary, preserve their status in the reverse dictionary as well.

dobry 1) добрий;
dobre słowo добре слово 1;
dobre słowo ласкаве слово 2;
ludzie **dob**rej woli люди доброї волі;
z **dobrej** woli з доброї волі 1;
z **dobrej** woli добровільно 2;
dobry 2) (do czego) підхожий (для чого);
dobry do tej roboty підхожий для цієї роботи;
dobry 3) (na co) придатний (на що);
materia **dobra** na płaszcz матерія придатна на плащ;
◊ *розм.* a to **dobre!** от тобі й маєш! 1
◊ *розм.* a to **dobre!** от тобі й на! 2
◊ *розм.* **dobra** nasza! наша бере!

The next step is to swap the elementary equivalents, which is a trivial operation of replacement of the left-hand side of the line with the respective right-hand side:

добрий; **dobry** 1)
добре слово 1; **dobre** słowo
ласкаве слово 2; **dobre** słowo
люди доброї волі; ludzie **dob**rej woli
з доброї волі 1; z **dobrej** woli

добровільно 2; z **dobrej** woli
підхожий (для чого); **dobry** 2) (do czego)
підхожий для цієї роботи; **dobry** do tej roboty
3) (na co) придатний (на що); **dobry**
materia **dobra** na płaszcz матерія придатна на плащ;
◊ от тобі й маєш! 1 *розм.* a to **dobre!**
◊ от тобі й на! 2 *розм.* a to **dobre!**
◊ наша бере! *розм.* **dobra** nasza!

However, the result of this reversing operation for basic equivalents is still quite distant from a genuine reverse bilingual dictionary formed according to lexicographic rules. This is why the further stage of work requires a number of compression operations, folding the entry back into a different combination of units. First, a list of words and word combinations available in the initial dictionary $Language_1 \rightarrow Language_2$ in the alphabetic order of $Language_2$ is created. In our case, basic equivalents extracted from the dictionary become the basis for the Ukrainian word list. The next step is the formation of word entries of the reverse dictionary. The equivalents extracted from the „dobry" entry, will appear in the entries containing relevant Ukrainian equivalent expressions: „добрий" („good"), „ласкавий" („kind"), „добровільно" („voluntarily"), „підхожий" („suitable"), „придатний" („fit"), „мати" („have"), „на" („on"), „брати" („take") and others. Clearly equivalents, for example for „мати", used either as a frequent functional verb or a noun ("have" or "mother"), will be gathered from various Polish headwords. To receive the basic (so-called dictionary) forms of words, the lemmatization procedure will obviously have to be used. The Ukrainian Grammatical Dictionary together with its supporting software developed at the ULIF NASU can serve for this purpose. Besides, it should be noted that main words of collocations should be determined during the compilation. These words will be the input to collocations in the reverse dictionary. If this choice is made and a system of grammatical identification of lexical units (lemmatization and paradigmatization) is available, the further creation of the inverse dictionary can be carried out automatically. Of course, some post-processing manual check and edition will be necessary anyway.

8 Database and an editing tool

After all basic cleaning and parsing stages the dictionary database is ready for further lexicographic work. A special editing environment is highly desirable for the more convenient work of the lexicographers, enabling them to introduce systematic changes into the dictionary. The lexicographical database of the explanatory dictionary of the Ukrainian language ("Словник української мови") developed at the ULIF NASU can be used as a model. In particular this system allows the user to view entries, directly access individual structural elements, as well as modify entries, replace elements, change the sequence of homogeneous structural elements, remove entries and add new ones to the dictionary. Thus, the lexicographical system is both a reference system for the user (an electronic dictionary) and an operating tool for lexicographers who compile or edit a dictionary. It should be noted that the structure of a bilingual dictionary differs significantly from a monolingual explanatory one, which turns the creation of a bilingual lexicographical database into a special independent task for which new solutions have to be found. An essential property of bilingual lexicographic systems is enabling users to enter the dictionary through either of the two languages' word list, which requires a reverse dictionary creation technology.

The approach presented here can produce an updated dictionary, as well as a lexicographic system as a computer tool set for further expansion and modification of the bilingual dictionary.

9 Future work

Lemmatization and paradigmatization allows us to conduct further interesting experiments. The word list of the Ukrainian part of the dictionary, with a frequency index, can be mapped against the word list of the explanatory Ukrainian dictionary. This can help us detect more outdated words, Russisms and Polonisms in an automatic way. It would also be interesting to see whether there are words of high frequency in the explanatory dictionary that are not used in the bilingual one and analyse this group.

On the other hand, we need to complete the bilingual dictionary with new terminology, e.g., of computer science, business, law, technology. Preliminary word lists for these fields to work with have already been extracted from the explanatory dictionary. Since bilingual terminology is usually presented by one-to-one correspondents, and our system allows for the reverse language direction to work with lexical entries, the source language of terms is no longer so important. Further work on existing lexical entries from the point of view of consistency of the grammatical description and presentation of semantic correlation of meanings within lexemes must be done as well.

Another practical task, important for language didactics, is extraction of automatic interlingual homonymy, or so-called translator's false friends.

We also plan to use Polish-Ukrainian corpus (PolUKR)[6] for acquisition of more translation equivalents, either automatically or manually.

References

[1] Kotsyba, Natalia and Magdalena Turska (2006). Polsko-ukraińska leksykografia – współczesny stan i perspektywy. In *Semantyka i konfrontacja językowa*, t. 3, red. V. Koseskiej-Toszewej, SOW, Warszawa.

[2] Słownik polsko-ukraiński we dwóch tomach (1958). Kolegium redakcyjne: A. I. Gęsiorski. T. Ł. Humecka (redaktor naczelny), M. Kiernycki, M.J. Onyszkiewicz, M.I. Rudnycki. Kijów.

[3] Trident Software. Polish-Ukrainian electronic dictionary and translator. `http://www.slownik.ukraincow.net/`

[4] Turska, Magdalena and Natalia Kotsyba (2007). Polish-Ukrainian Parallel Corpus and its Possible Applications. In *Proceedings of the International Conference 'Practical Applications in Language and Computers', 7–9 April 2005, Łódź*, Peter Lang GmbH.

[5] Universal (Pl-Ua) within ABBYY Lingvo x3 version (2008). Electronic version is based on: Polish-Ukrainian and Ukrainian-Polish dictionary by Anna Malecka and Zbigniew Landowski, edited by Vyacheslav Busel, ITF "Perun", 2007.

[6] Соломко, Валентин. Багатомовний електронний словник: `http://slovnyk.org`

[7] Шевченко И.В., Широков В.А., Рабулець А.Г. (2005). Электронный грамматический словарь украинского языка. // Труды международной конференции „Megaling'2005. Прикладная лингвистика в поиске новых путей". 27 июня–2 июля 2005 года. Меганом, Крым, Украина., p. 124–129.

[8] Широков В.А. (2005). Елементи лексикографії. Київ: Довіра.

[9] Широков В.А., Рабулець О.Г., Шевченко І.В., Костишин О.М., Якименко К.М. (2007). Інтегрована лексикографічна система „Словники України", версія 3.1. Київ. CD-видання.

[6] `http://www.corpus.domeczek.pl`

To a Question about Semantic Syncretism in Old Russian Language and Its Reflection in Modelling Semantics of an Old Russian Word*

Irina Nekipelova

Izhevsk State Technical University

Now the developments of the modelling of a word lexical meaning description and its semantic relations are important in the work of multifunctional web-modules of texts transcriptions. In this connection the creative group under V.A.Baranov's direction works at the creation of the automated lexical-semantic analyzer in the informational-analytical system "Manuscript" (`http://manuscripts.ru/`). This aspect is connected to a problem of the system use for the linguistic research in the field of the vocabulary and semantics and the development of the linguistic search system allowing the user to have an exact idea about a word lexical meaning and its semantic relations in language and texts of ancient manuscripts, kept in IAS "Manuscript" databases. Problems of the modelling of semantic, thematic and word-formation relations of words of the Old Slavonic and initial ancient Greek texts, the search of conformity, the storage of semantic relations in databases and their use are the most important.

Types of the lexical description are a basis of the lexical meaning and word semantics modelling in databases. Originally it is necessary to differentiate concepts a linguistic meaning and a lexical (nominative) meaning. Different in volume linguistic units - from a mark to a word-combination (a mark, a word form, a fixed expression, a word-combination) - have a linguistic meaning. Also, different in volume linguistic units - from a word to a fixed expression (a word, a speech formula, a fixed expression) - have a lexical (nominative) meaning [Nekipelova 2006: 140-147, 2006: 298-303]. Also, it is necessary to say that "All types of the meaning are understood as the additional ones to each other, i.e. as parts (the sides, aspects) of the whole" [Nikitin, 1997: 51]. It is significant because at the description of word semantics in its history it is necessary to take into account some facts complicating the research. First, the word (its word forms) in its modern state is examined only in a certain context / contexts that complicates the fixation of all possible word uses and its connections and relations with other linguistic units because the extant texts can not reflect all word relations which were realized during that period of the language development. Therefore, it is evidently the researcher examines not the whole semantic field. Thus, as the result of the lexical-semantic analysis of word functioning in a context it is possible to fix only certain semantic word relations, and only in exceptional cases it is possible to assume about the some elements existence connecting some linguistic phenomena because there is no full reliable information about all word relations and characteristics, no full list of word meanings, formula, fixed expressions, etc. from that period of the language existence. Many scientists are engaged in the reconstruction of these relations and their opinions about the ancient text interpretation do not always coincide.

Second, at the interpretation of words relations used in ancient texts, it is not always possible to speak about absolute adequacy of such an analysis because the description of word semantics is examined within the semantic word relations in the modern language which could not be in this lexeme at earlier stages of the language development. The basic complexity of the Old Russian texts studying was formulated by V.A.Baranov: "Unfortunately, till now we are not always sure that our understanding and interpretation of Old Russian texts from the point of syntagmatic relations, grammatic structure and semantics view is adequate to the text understanding by the ancient scribes" [Baranov 2003: 16]. The research of texts semantics is the most complicated because "... the system of Russian is just being formed and presents the other system, in many aspects different from modern one" [in the same place].

* The study and preparation of these results have received financing from the grant of president Russian Federation under grant agreement МК-4353.2008.6.

First, when we describe word semantics and its language relation we are guided by a context in which the word is used, and by the data of various linguistic dictionaries, fixing the use of a word in analogous, similar or other contexts. The use of dictionaries helps to reveal typicalness / atypicalness / occasional use and regularity / irregularity of the word use in a text / a context. It is important for revealing the regular and casual word use in the certain period of the language development.

Now rules of the meaning types description of each word are developed. Semantics modelling is the development of the semantic description typical structure, and the instantiation of this structure depends on the individual characteristics and relations of words.

The semantics modelling of an Old Russian word is submitted on a material of Color Triod text, in 11-12th and in 13th Centuries. (РГАДА, ф. 381 (Син. тип)), № 138, 173 p. Further the work with a material of other Triods lists, contained in the database IAS "Manuscript", and also Triods lists, being prepared for the publication is planned.

We developed the structure of the word meaning description as relations, reflecting hierarchical words connections. The structure of the semantic word description shown in the table demands some comments.

The basis of the characteristic of word relations and attributes is the description and differentiation of the linguistic typology of word meanings. The linguistic typology of meanings directly connects them with the way of the language words expression. “As a matter of fact, the linguistic typology of meaning has no direct relation to the contents and the character of an expressed meaning, and characterizes it on the linguistic unit level” [Nikitin 1997: 67]. The linguistic typology of meanings directly connects a meaning with the way, character of its language expression. The basic categories of the linguistic typology of meanings are grammatic, nominative and communicative, and, also, syntactic, morphological and word-formative (as types of grammatic meanings), lexical, phraseological, word-combinative (as versions of nominative meanings) meanings. Differences in the stratification nature of linguistic units are the base of differences in linguistic types of the meaning.

First of all, we develop the nominative type of the meaning because it directly reflects lexical-semantic word relations.

The definition of a lexical word meaning is the most important for the lexical word description when there are seven basic types of the lexical word meaning description: encyclopedic, defining, etymological, synonymic, antonymic, reference, homonymic.

There are no examples for the encyclopedic and etymological interpretation - the citations from the Triod because they reflect initial word relations. The encyclopedic word meaning is right only for the initial word meaning. The homonymous lexemes have no the encyclopedic meaning because dictionaries of this type do not contain meanings of homonyms. The same concerns the etymological word description all derivatives and homonyms have no the etymological characteristic. However, for the semantics description of the majority of them the field of word-formative relations – “reference meaning” → “to a primary word” (if it is possible use the data of the word-formative dictionary) is filled. Thus, we see the description of the phenomenon and word meaning from the different points of view which are not contradictory to each other.

Fields, where the phraseological meaning is fixed, are filled in process of the increase of researched materials volume. As known, the process of the conversion to a fixed expression has a long history, and those set phrases and expressions which are in tests of textual heritage of the 11-14th Centuries, are not yet fixed expressions. Mainly, scientists fix the functioning speech formulas this period.

V.J.Deryagin notes: “For the period of the usual business writing in the language aspect the formula is needed to be understood as a phrase of the nominative or communicative character, and also a word-combination, a phrase (the model of the sentense) with more or less constant lexical structure. On occasion the formula can consist of the several sentenses connected among themselves with the syntactic and semantic link” [Deryagin 1985: 243]. The formula is the basic unit of the stylistic analysis of the business text, it is the unit of a text level, but at the same time the formula can be determined in the terms used for

units of other levels, lower in the hierarchy: a formula - an offer (the certain type), a formula - a word-combination (the certain type), a phrase [In the same place: 244].

However, language formulas are not only in texts of business writing, but also in texts of other genres because the use of speech formulas is defined not only by the genre characteristic, but also by the common language processes. One of means of the speech formulas formation is the semantic tracing of the Greek metaphors resulting to their symbolization. V.V.Kolesov's the term "formula" first of all correlates the term to the form of the borrowed symbols of the Greek culture expression in Old Russian texts. "Most ancient [loan words] were not free from contexts in which they went to Slavs, and these contexts got to them in writing translated texts. The word-combination was adopted as a whole, that's why loan words became fixed" [Kolesov 2002 : 201]. Thus, formulas are word-combinations or sentenses connected by the syntactic (in a context), phraseological and semantic (by sense, the contents) links and characterized by the stability and reproducibility [Nekipelova 2005: 188].

The field "speech formulas" is constantly filled. It is necessary to note that those linguistic units which are marked in this field, do not always have the meanings fixed in various dictionaries, and from the point of view instantiation of citations from the Triod text the field "fixed expressions" is empty. It is possible these two fields do not coincide and have no common data.

Defining meaning is submitted as linguistic and contextual ones. Linguistic word meaning is the one of linguistic unit, that is the unit fixed in language of the certain period, regularly used in texts of various genres. The basic parameters of this fixedness are: 1) the high rate of the word use in ancient textual heritage; 2) fixation of a word and its meaning in the Dictionary of Russian of 11-17th Centuries; 3) the coincidence of its meaning with the etymological meaning. All other cases of word meaning representation (a derivation and a homonymy) are located in other fields. So all homonyms and the semantic derivative words fixed in dictionaries are represented. The meanings which have not been fixed in dictionaries, are represented in a field "contextual meaning" as independent lexical units with those meanings which they have in the given context, added by co-meanings and connotations.

This differentiation, in our opinion, is expedient because when the user finds the lexical meaning of an exact word he should get the meaning of exactly this word in this context, instead of all meanings in what the required word can be used. It is important also for the description of word-formative relations: in the description of word semantics the exact representation of the primary word and derivative words by the semantic way should be shown. The definition of contextual word meanings is important for the word interpretation. We develop some criteria of differentiation of linguistic and contextual meanings. One of criteria - presence or absence of the meaning description in dictionaries, in the first case we speak about the word use which has settled in language, in the second case that process of the concrete use fixedness is still being developed or in the casual use. The following criterion is the use degree in one text, in texts of one genre, in texts of different genres. The use frequency reflects the word fixedness in the language, the rare or individual use reflects incompleteness of the word fixedness or about its casual / atypical use. The third criterion is the fixing of an absolute word use as a lexeme independent of a context or an opportunity of the word use only in system relations with other words in the context. The opportunity of the absolute word use testifies to existence of this word as the high-grade unit in the language of the certain period. The opportunity of the word use only in a context can testify about its occasional use, the full dependence on the context, the symbolical character of the text meaning, the expansion of the word semantics and the initial stage of the formation process of the semantic derivative and, at last, about the process of the conversion to a fixed expression of word-combinations / statements. The successive use of these three criteria for the description of the word functioning most precisely allows to reveal the linguistic or contextual character of word meaning. This field from the point of view of the scrutiny level is the least investigated, therefore the special attention will be paid to the semantics description of these units.

Certainly, not all fields will be filled as a result of the analysis of different words semantics. The lexical meaning can be the reflection of a simple feature and no more, then it has the simple structure of

words, not decomposable on semantic features. Similar words have no definitions in explanatory dictionaries and they can be interpreted only indirectly - by synonyms or by the use. The list of these words till now is not clear for scientists. Thus, investigating the word semantic relations, also it is necessary to specify the interpretation through synonyms, antonyms, reference interpretation and interpretation through the word use. In many cases it will be carried out with the help of comments and supplementary information.

Not always a word have all types of meanings. All significant words potentially have all submitted characteristics, features and relations, however, only few words as much as possible realize them. As a result, at the analysis of different words semantics some fields will not be filled. Also, the main problem of the semantics research in a language history is connected with it. The subjects of the problem of the lexical-semantic model construction is connected with several aspects in linguistics - first of all, with the word-formation, lexicology and semantics, therefore methods of all these sections should be presented at the work. Alongside with the process of the new lexemes formation it is necessary to show functioning of these lexemes in the language and the text and also concretize those processes which occur in a word semantic structure. The data of modern Russian do not give the exact answer how these processes are realized, and the use of the historical material strongly complicates the research process because to a greater extent it promotes the analysis subjectivity, the attributing derivative those characteristics which are caused by an individual view of an author.

The complexity of the semantics description is connected with a number of factors: first, the basic problem is the semantic model in itself because we should take into account the data of all language levels in the model while semantics is not the language level – it is the content of the language; second, we should take into account the language development in the semantic model where the change of a semantic component is the most variable one. Thus, the model should represent time and spatial relation, vertical (diachronic) and horizontal (synchronic) relations) of language units.

Complexity is the definition of homonymic and derivational relations. It is connected with the problem of the polysemy, homonymy, and semantic derivation. However, the analysis of the language material shows it is not the opposition of the terminology for the same phenomena nomination, but it is the name and the definition of the different phenomena which are not valid and differentiated in linguistics because it is not taken into account that the mark, instead of a word [Kolesov 2002], is of many meanings.

Originally, in the early period of the language development the semantic syncretism of words was fixed when the word directly named a denotatum and meant something greater. The origin and activization of the process of the addition of a word meaning by connotations have resulted in the transition of Slavs mental thinking from the subject, concrete type of thinking in itself – the direct nomination of a denotatum - to the abstract type of thinking – the subaudition, meanings addition.

But not all words were syncrets. Some words did not develop additional meanings and even connotations. Their subject meaning was kept very long, and in some cases it has not changed by now.

Homonyms and semantic derivatives functioning is not attached that period because the word meant more than a concrete denotatum. That time the context (a word environment) starts to gain in importance because it had the basic semantic meaning.

This process is the result of extra linguistic factors because the complication of the language system became the result of the complication of cause-and-effect relations in the world and, hence, in Slavs view to the world. Occurrence of the new information about the world should be expressed by means of the language. And development of the new information inevitably occured by the comparison with known things about the world, and even moreover - on a basis of known. For the first time this thesis was stated by Nikolay Kuzansky. But the attention to it was paid in linguistics much later: in a number of the works F.I.Buslaev and A.A.Potebnya proved the anthropocentrism of the human thinking and the expression of this thinking in language, learning through already known facts. The complication of the language system,

including the complication of its semantic part, was connected with the complication of the process of Slavs extrapolation who became capable not only to name but also to assume and expect.

The semantic syncretism was based on the combining of additional meanings, the mark representing a symbol, became to name or designate many things. And the context was intended to help to determine the relevant symbol meaning, a mark in a situation. The symbol in itself as a language mark represented the system of the potential meanings and connotations, attributed to a word as a mark during this period of the language development. Any potential meanings could be realized in a context and become relevant ones, but only for this and similar contexts. As a result, the first linguistic and speech units with both syntactic and semantic links, with the certain constancy in use and functioning were formed, first, in the contextual use, then in the language. However, they had some freedom in the grammatic expression: the structure of these units, the completeness and sequence of the structure, the word-formative and morphological forms forming the compound name were rather free.

Semantic syncretism was based first of all on the expansion of a word semantic volume, but this expansion was the process of the meanings and connotations juxtaposition, but not result of a new word forming.

The end of words syncretism process was connected with the occurrence of the semantic derivatives which have appeared as a result of metonymical processes, based on the nomination by the contiguity.

Actually, it is impossible to divide time of the syncrets and metonymical derivatives functioning, it is right to say about the gradual change of thinking, and, then about the gradual prevalence of syncrets functioning, and then about the derivative one. However, always in any developing system including a language, transitive stages and the elements which are the reflection of these stages are fixed. Therefore, the disintegration of a word initial syncretism can be shown by means of the analysis of disintegrating syncrets, but only in a functioning system of these formations.

Also it is necessary to tell this process, like all processes in a language, is not absolutely universal as it is impossible to tell about the universality and objectivity of the human thinking. Therefore, it is not necessary to be surprised when we meet a metaphor a little before XIV century, and there are some examples of the semantic syncretism in the modern language.

The occurrence of semantic derivatives is connected with completely different ways of Slavs thinking and expression of their mentality. Actually, it is necessary to differentiate the metonymic derivation with the analogical processes and the metaphorical derivation.

The metonymic derivatives were formed as a result of disintegration semantic syncretism of words – the differentiation and division from each other those meanings which were combined concerning one mark and which in due time promoted the expansion of a word semantic volume. It is essentially a typical process as a result of which the certain word-formative models were generated. It is not a categorical breaking, but it is the categorical continuation of the development of those meanings which emerged in a syncretic words. The metonymy reflects the development of the abstract thinking, that is the definition and designation of a denotatum by means of the abstraction of new things from known and concrete ones, that initially was a base for the denotatum nomination.

Metaphorical derivatives were formed in the result of "gallop" in the thinking, a division of new from already known, the categorical breaking. Such linguistic process could be realized only in more developed thinking, rather than in abstract one. The occurrence of metaphors has resulted in the transition of Slavs thinking to more high level - abstract, it is not the abstraction, and a gap, therefore the metaphor is the essentially not typical process which should not be mixed up with the concept “atypical process”.

The formation process of the metaphor have been examined since Aristotle who has given its theoretical substantiation and definition for the first time. For all this time scientists even allocated typical models on which new metaphorical derivatives can be formed. But all these statements do not have the universal character. The language competence allows to native speakers to distinguish the whole categorical breaking which results in the metaphor, from the categorical breaking which results in

a mistake. For example, names of some parts of a body are used for the metaphorical name of the person to which the certain attributes and qualities are attributed on the basis of individual associations (similar nominations have appeared one of the first in language of the Middle Ages (a hand, an ear, an eye), the rethink of ancient Greek metaphorical loan words (language, a head) promoted to that, but some other names of parts of the body till now are not used for the metaphorical nominations (a leg, a side etc.). Linguists have described the process of the formation of the first things, but till now they are not capable to explain absence of it in the functioning of the second ones. Typicalness of a metaphorical derivation is an artificial association of derivatives on the basis of their belonging to the certain subjects, but the development of metaphorical models is a substitution of word-formative relations by the nominative approach to the characteristic of these derivatives.

The difference between the metonymical models and the metaphorical groupings is connected with that the formation of the metonymical models is the one of word-formative models with a high degree of potential fillability while the formation of the metaphorical groupings is a result of immediate individual processes. It confirms the analysis of the language data in the language history. The metonymical models are realized consistently and in the wide volume, the formation of the metonymy is essentially clear. Therefore, the realization of the metonymical derivations potentiality is the real, consecutive, possible phenomenon. And the potentiality of metaphorical derivations is the inconsistent, immanent, and probable, the occurrence of the metaphor only can be assumed instead of the prediction.

The best proof for this it is the use of the lexicographic materials: in dictionaries there are a great number of meanings based on the metonymical relations (as for the lexicography we cannot say about derivatives because only in case of the homonymy the words meanings belong to different words with the identical phonetic expression), thus the majority words in Russian, even in nonliterary its form, have the ramified meanings; and metaphorical names are not so frequent as metonymical ones. Therefore, we can not speak about change of the metonymical thinking to the metaphorical one on the boundary of 14-15th Centuries because there was the addition, instead of the replacement: the activization of the metaphorical thinking has not eliminated the metonymical thinking like the development of the abstract thinking has not become the proof of the refusal from the subject one (the direct denotatum nomination is often a basis for the nonderivative words).

It is wrong to identify concepts a derivative and a homonym. It is the result of different processes. The derivative is first of all a concept of the word-formation because it means word-formative relations, the homonym is the concept of the lexicology because it means relations between the words without word-formative relations.

Semantics penetrates all language levels, that is why at the construction of semantic models it is necessary to take into account the processes of other levels. For this reason, the semantic models are essentially different from other linguistic models. Here the instantiation of models is not the basic factor because the instantiation is not connected with the universality. While one of the basic criteria of a model, including the linguistic model, is the prediction of unknown, but possible behaviour of an object which should be proved by the data of the supervision or the experiment.

The modelling means the use of the abstraction and idealization. Reflecting the relevant essential (from the point of view of the research) properties of the original and distracting from insignificant ones, the model becomes some abstract idealized object. Any model is based on a hypothesis about the suggested structure of the original and it represents the functional analogue of the original that allows using knowledge about the model for the original. Ideally, the model should be formal (i.e. it should have initial objects, in an explicit form and identically defined, the relations, connecting them, and rules for use), it should have the explanatory power (i.e. not only to explain the facts or the data of experiments, inexplicable from the point of view of already the existing theory but also to predict unknown earlier, but possible behaviour of the original which should be proved later by the data of the supervision or experiments).

The modelling of a word semantics is the infinite process because the semantic model should be not simply a base for the description and classification of linguistic units semantics, but the structure of all possible semantic relations, even if any element will be single.

As a result of the lexical-semantic and grammar-semantic analysis of all words used in Triods, the semantic model, hierarchically designed and including all the word connections and relations from the Old Russian language should be formed because the complex of meanings and connotations forms the semantic system of the language (system of meanings).

References

[1] Baranov V.A. Formation of attributive categories in the history of Russian / V.A.Baranov. – Kazan: Publ. of Kazan State University, 2003. – 390 p.

[2] Deryagin V.J. Variation in formulas of Russian business language of the 15-17th Centuries. / V.J. Deryagin // The East Slavs: Languages. History. Culture: For the 85 anniversary of the academician V.I.Borkovsky. – M.: Science, 1985. – P. 243-249.

[3] Kolesov V.V. The Russian word philosophy / V.V.Kolesov. – S.-Petersburg, 2002. – 448 p.

[4] Nekipelova I.M. The metonymy and metaphorical derivation in the history of Russian (on the material of the business writing heritage of the 11-17th Centuries).The dissertation Cand.Phil.Sci. – Izhevsk, 2005. – 282 p.

[5] Nekipelova I.M. Problems of the word semantics modelling in databases // Works of the Internatioanl Conference “Corpus linguistics - 2006” (SPb., on October, 10-14th 2006). S.-Petersburg: Publ. of S.-Petersburg State University. – P. 298–303.

[6] Nekipelova I.M. The problems of the description and the word semantics modelling in databases // The modern information technology and textual heritage: from the ancient manuscripts to the electronic texts: materials of the International Scientific Conference (Izhevsk, on July, 13-17th 2006). – Izhevsk, Publ. of IzhSTU, 2006. – P. 140–147.

[7] Nikitin M.V. The course of the linguistic semantics: The tutorial for students, post-graduate students and teachers of linguistic disciplines at schools, lycées, colleges and high schools. / M.V.Nikitin. – S.-Petersburg: The scientific center of the dialogue problems, 1996. – 760 p.

[8] Potebnya A.A. Theoretical poetics. – SPb.-M.: Akademia, 2003. – 374 p.

Morphosyntactic Specifications for Polish. Theoretical Foundations. Description of Morphosyntactic Markers for Polish Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)*

Roman Roszko

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw

Abstract. In this paper the author presents the foundations for a scientifically-rigorous classification of lexemes into classes (parts of speech). Then he presents and analyses a portion of a new and already widespread classification into parts of speech (POS) authored by Zygmunt Saloni. Saloni's classification is also known from the tagger for Polish, TaKIPI (IPI PAN Corpus tagger). The analysis of Saloni's classification is aimed to develop morphosyntactic characteristics for all POS classes in the Polish language that would be in line with the morphosyntactic specifications used in MULTEXT-East. The author adjusts classification of Polish categories to the MULTEXT-East requirements. When necessary, he extends the already existing MULTEXT-East morphosyntactic specifications in accordance with its descriptive convention. The first stage involves development of morphosyntactic specifications for Polish nouns. Given the innovative subdivision into parts of speech, differing from traditional grammatical descriptions, and the existence of morphological, semantic and syntactic subcategories not found in other languages, the author expands the number of markers for Polish nouns. The following categories are the new morphosyntactic specifications: human, animate, post-prepositionality, stressability, depreciativeness. The category of gender has been rearranged. The author does not follow the elaborate gender system proposed by Saloni and retains the subdivision into masculine, feminine and neutral gender, as used in MULTEXT-East. Instead, he proposed new characteristics, human and animate, as independent, stand-alone attributes. The next step in the process will be to develop morphosyntactic specifications for the remaining parts of speech in the Polish language.

1 Introduction

The problem involving the degree of morphologisation of various meanings in natural language has a significant bearing on the grammatical description of that language. A high number of morphological categories, their transparency and absence of exceptions greatly facilitate such a description. However, Polish is not one of the languages where the degree of formalisation of meanings would facilitate grammatical description. Its evolution, including even only phonetic changes (which involve, e.g., simplifications, analogies or assimilation) as well as external influences and internal regional differentiation over the centuries, means that the contemporary Polish language is characterised by formal presence of multiple variants coupled with a generally modest number of morphologised meanings. It is enough to compare the declension system for Polish and Lithuanian nouns to see that the noun declension systems which were originally close to each other have remained simple, transparent and exception-free only for Lithuanian.[1]

My aim here is to adjust the grammatical description of the Polish language to the existing description within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004), developed for a larger group of languages (eleven, to date). Consequently, it must be immediately noted that one cannot talk about classes or parts of speech as a universal phenomenon, common to all natural languages. This suggests that a description of morphosyntactic characteristics for multiple languages is difficult and calls

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

[1] References to Lithuanian are by no means accidental. Together with prof. L. Dimitrova (IMI, BAS), prof. V. Koseska (ISS, PAS), Dr. D. Roszko (ISS, PAS) we are conducting preparatory work for a parallel Bulgarian-Polish-Lithuanian electronic corpus that would contain morphosyntactic specifications. We intend to expand the bilingual electronic Bulgarian-Polish dictionary by adding Lithuanian. Notably, Lithuanian is considered to be unique in the group of Indoeuropean languages as it is highly archaic.

for some satisfactory compromise. When building a simultaneous morphosyntactic description for many languages, one should first ask about the required outcome. Will the morphosyntactic description be used to build electronic parallel corpora, or electronic bilingual/multilingual dictionaries as well? Therefore, one needs to answer more questions: is the morphosyntactic description supposed to rely on word classes identified on the basis of inflection types (particularly important for inflective languages), or is it supposed to reflect types of inflection as well as meanings? Or, perhaps, is it only required to describe the meanings?

I do realise that there is no single subdivision into parts of speech for a language. One might design any number of classifications into parts of speech where any number of such parts is imaginable. For instance, the so-called adjectival participles in Polish (e.g. *czytający*, *składany*) would, in various classifications, be classified as: (1) verbs, or (2) adjectives, or (3) a separate part of speech: participles. In school grammars (e.g. Bąk [1]) adjectival participles are classified into the class of verbs. One might wonder whether it is a fortunate idea to combine verbal forms and participles together. One should notice that old past-tense active participles in Polish with *-ł-* have 'migrated' to the class of adjectives. In fact, those are rare remnants of those participles. Here are some selected examples: *zmarł-y* (*-a*, *-e*, *-i*, *-e*), *przeszły*, (*za*)(*nie*)*dbały*, *sparciały*, *naleciały*, *stopniały* etc. Some of them are used in contemporary Polish also as nouns, e.g. *zmarły*, *zgłodniały*, *ociemniały* etc. In view of the declension paradigm of participles, some researchers consider adjectival participles to be adjectives whereas others view them as a separate part of speech. The latter base their choices on function and direction of derivation.

Another problem for building a description of morphosyntactic characteristics is an unclear notion of 'word' which, as apposed to morpheme, is neither stable nor fixed. 'Word' continues to have arbitrary definitions. As a result, if a definition of 'word' is adopted, this is likely to exert significant influence on the final shape of such classification. Let us notice that 'word' may have a few meanings: phonological word, orthographical word, textual word, grammatical word (dictionary word, or lexeme) as well as other words which denote a limited/truncated set of forms and cannot be considered as items in the subdivision, for instance auxiliary word, empty word etc.

Phonological word – a string of phonemes delimited by pauses on both ends. In the Polish language both [widzi mi się] and [chyba] are phonetic words. Intuitively, the word [chyba] is simpler than [widzi mi się]. However, the latter example is perceived by an average user of Polish as a three-part element. This is because each of the components may appear separately in different contexts, e.g. *on widzi*, *daj mi*, *ubieraj się* etc. Certainly, a phonological word cannot be an object of a classification into parts of speech.

Orthographical word – a string of written text, delimited by spaces; it is an artificial creation and, as such, cannot represent the basis of classification into parts of speech. Let us consider the two functionally close examples: *_na_ pewno_* and *_naprawdę_*. The former receives the following description in the *Morfeusz* analyser [5]:

```
<tok>
<orth>na</orth>
<lex><base>na</base><ctag>prep:loc</ctag></lex>
<lex disamb="1"><base>na</base><ctag>prep:acc</ctag></lex>
</tok>
<tok>
<orth>pewno</orth>
<lex disamb="1"><base>pewno</base><ctag>adv:pos</ctag></lex>
</tok>
```

The latter receives the following description [5]:

```
<tok>
<orth>naprawdę</orth>
<lex disamb="1"><base>naprawdę</base><ctag>qub</ctag></lex>
</tok>
```

If *_na_pewno_* were spelt without a space (*napewno), then its morphosyntactic description might look as follows:

```
<tok>
<orth>napewno</orth>
<lex disamb="1"><base>napewno</base><ctag>qub</ctag></lex>
</tok>
```

Arbitrariness of spelling (words written separately or without a space) in the Polish language is reflected, for instance, in the recent spelling reform which recommends that *nie* with the so-called adjectival participles should be spelt as a single word. Before the reform particle *nie* with participles was spelt either separately or without a space, depending on the syntactic function of the participle. Let us take another example from Lithuanian. In Lithuanian, *ne* (equivalent of Polish *nie*) is spelt together with participles and verbs, as in the example below:

Lithuanian	Polish
nedirbu (praesentis)	***nie*** pracuję
nebuvau padaręs (perfectum)	***nie*** zrobiłem
neparašęs (participium praeteriti activi)	ten który ***nie*** napisał

Likewise, the Lithuanian *si* (equivalent of Polish *się*) is spelt together, as below:

Lithuanian	Polish
sveikina***si*** (praesentis)	żegna ***się***
at***si***sveikino (praeteritum)	pożegnał ***się***
juokiąsi***s*** (participium praesenti activi)	śmiejący ***się*** (z czegoś)
pa***si***juokęs (participium praeteriti activi)	ten który pośmiał ***się***

Textual word – it is related to rules that determine the word order in a sentence. More than a phonological word or orthographic word, this one could become an object of classification into parts of speech. A set of various textual words builds a grammatical word.

Grammatical word – much as the textual word, this one is related to (functional) syntax. In particular, the ability to enter into syntactic relations is considered to be specific to grammatical words. In the definition of a grammatical word [6, p. 646] syntactic characteristics are combined with semantic attributes that constitute the language-specific meaning and textual words. Possible links are made: a grammatical word identical with a textual word (e.g. *rzekomo* – *rzekomo*) and a grammatical word with a finite set of textual words (e.g. psycholog[nom. pl. masc.] – *psychologowie* / *psycholodzy* / *psychologi*). Sometimes a grammatical word is considered synonymous with a dictionary word, also called a lexeme.

Dictionary word (**lexeme**) – this is an established unit of dictionary descriptions, strictly linked with the adopted classification of words into parts of speech. While, theoretically, a dictionary word should be the object of classification, in practice it refers to some previous subdivision into parts of speech.

Summing up the above, let me point out that when making a subdivision into parts of speech, we must make the following important realisations: 1. What exactly is it that we are subdividing? and 2. What is the goal of this subdivision? A classification into parts of speech which is to be created should meet the criteria of scientific rigour. Therefore, a dichotomous subdivision (into two) is required at each stage. Also, clear, non-contradictory and uniform subdivision criteria are required. A criterion that has been already used at one level should not be used again at a lower level for a narrower set of lexemes. Moreover, the resulting subdivision should be easily verifiable, which means that, above all, it should cover the entire vocabulary. Is this kind of task feasible at all?

2 Theoretical foundations: *Słownik gramatyczny języka polskiego*, author: Zygmunt Saloni (Saloni in: [7])

It is for a reason that the theoretical foundations for the grammatical dictionary of the Polish language, authored by Z. Saloni, are the object of our interest. Importantly, Saloni's theoretical foundations became the point of departure for the very popular tagger for Polish, TaKIPI [5].

2.1 Formal foundations for identifying lexemes

As already mentioned in Section 1, the selection of lexemes is an important task and certainly has a crucial importance for further work on identifying classes of words, i.e. parts of speech, and their subclasses. Saloni believes (Saloni in: [7]) that Polish words should, above all, consist of Polish letters (or, in the auditory dimension, consist of sounds that are typical of the Polish language) and lexemes should be separated from one another with spaces. Another important criterion is the use of lexemes: they should appear more often and be repeated in modern times. As Polish is a inflective language, its words should fall into regular sets that operate within certain inflective types. A set of all inflective variants of the same core (stem) is a lexeme, for instance: ***dom***-, ***dom***-*u*, ***dom***-*owi*, ***dom***-*em*, ***dom***-*y*, ***dom***-*ów*, ***dom***-*om*, ***dom***-*ami*, ***dom***-*ach*. In dictionaries, a lexeme is usually represented by a single, default form, traditionally described as the dictionary form or base form. Consequently, various nouns are usually represented by the nominative case singular (e.g. *dom*), whereas verbs are represented by infinitives (e.g. *czytać*) etc.

2.2 Semantic foundations for identifying words and lexemes

An initial formal subdivision into Polish words is further analysed using morphological and semantic criteria. It is important to emphasise that unspaced spelling may sometimes lead to erroneous identification of words. According to Polish rules, unspaced spelling is required for some postpositional particles, abbreviated personal forms of praesenti for the verb *być* 'to be', for some operators or agglutinates. Based on J. Tokarski's *a tergo* dictionary [8] and Saloni's grammatical dictionary (Saloni in: [7, p. 19–21]), some cases of unspaced spelling for two words are given below.

Particles *ć*, *że*/*ż*, *li*, e.g. *pójdę-ć*, *dasz-li*, *już-że*, agglutinates or abbreviated personal forms of *praesens* for the verb *być* 'to be': *m*/*em*, *ś*/*eś*, *śmy*/*eśmy*, *ście*/*eście*, e.g. *ja-m*, *że-ś*, *skąd-eśmy*, *gdzie-ście*, conditional mode operator *by*, e.g. *jakkolwiek-by* or *jakkolwiek-by-m* (with the agglutinate *m*). There is also another form of the pronoun *on* 'he', common for the genitive and accusative case, spelt unspaced: *ń (do-ń, za-ń)*.

In opposition to the above, there is a case of separate (spaced) spelling of inflective forms, for instance *będę czytać* / *będę czytał.* Again, let us refer to the aforementioned examples of *na_pewno* (Lithuanian *tikrai)* and *śmiać się* (*śmiejący się*). While the Polish compound form *na_pewno* should be treated as a separate lexeme, the forms with *się*, even the ones which do not occur without *się* (such as *bać się*), are mere combinations of two lexemes. This is determined by their semantic properties. Saloni (Saloni in: [7, p. 21]) mentions more examples of so-called compound lexemes such as *po polsku* (Lithuanian *lenkiškai*) which are regularly derived from adjectives ending with *-ski*, *-cki*, *-dzki*. He challenges some phraseologisms and archaisms.

2.3 Foundations for identifying parts of speech in Polish

The problem of foundations underlying a classification into parts of speech (with examples) was presented at MONDILEX in a joint presentation delivered by Violetta Koseska-Toszewa and myself [3]. Let me now offer a brief overview of the most common criteria applied for identifying parts of speech: ontological/intuitive, morphological, semantic and syntactic. In the aforementioned article we write that there are hardly any classifications into parts of speech for Polish that would be based consistently on a single criterion (e.g. a semantic one). The aforementioned subdivision of Polish lexemes into parts of speech, as proposed by Saloni, is not consistent, either. It seems that Saloni relies most heavily on the criterion of morphology (inflection). When inflection fails to provide an answer, secondary criteria, semantic and syntactic ones, are employed.

3 Noun

In order for the morphosyntactic description of Polish nouns to fit with the description rules contained in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004), we should first review the definition of noun proposed by Saloni for the Polish language. Saloni (Saloni in: [7, p. 29]) identifies noun lexemes based on morphological criterion in the inflective category of case (= declension), inflective category of number and selective (i.e. not inflective) category of gender. Saloni specifies a number of characteristics which, in his opinion, are specific only to nouns but seem to have no influence on nouns being identified as separate parts of speech. They are just a specific addition to nouns as parts of speech, identified on the basis of inflective criteria of case and number and non-inflective criterion of gender. According to Saloni, nouns have the following additional specific characteristics: depreciativeness, uniformity, post-prepositionality and stressability. We will not focus on details here but will look at Saloni's seemingly controversial suggestions and on language-specific characteristics of nouns. Firstly, the author formally eliminates the class of uninflected nouns. Based on syntactic criteria and analogy to other, typical nouns (traditionally referred to as inflected nouns) he builds a paradigm for all nouns traditionally described as uninflected, as in the following example for *emu*:

Case	Singular	Plural
nominative	emu	emu
genitive	emu	emu
dative	emu	emu
accusative	emu	emu
instrumental	emu	emu
locative	emu	emu
vocative	emu	emu

Secondly, Saloni includes some forms traditionally considered to be pronouns onto the class of nouns: *ja* 'I', *ty* 'you', *on* 'he, she, it, they', *my* 'we', *wy* 'you', *kto* 'who', *ktoś* 'someone', *ktokolwiek* / *ktośkolwiek* 'anyone', *co* 'what', *coś* 'something', *cokolwiek* / *cośkolwiek* 'anything', *cóż* 'whatever', *nic* 'nothing', *się*$_1$ 'self', *się*$_2$ 'self', *wszyscy* 'everyone', *toto* 'this thing', *niecoś*, *śmo*, *wasze* 'yours' and other, a total of ca. 40 forms. It is important to stress that this group of nouns, in Polish referred to as nominal pronouns or 'nominal-pronominal lexemes' does not fit into the adopted paradigm. However, as we can see, this does not mean they cannot be considered as nouns based on non-inflective criteria. In this particular case semantic and syntactic criteria play a role. Let me point out that paradigmatic criteria have never been prevented lexemes such as *luty* 'February', *Kowalski* 'Kowalski' [surname], *przekątna* 'diagonal', *komorne* 'rent' and others which have a inflection typical of adjectives from being considered nouns. In this case inflection was not the criterion that determined the classification into the class of nouns.

3.1 Case

The category of case is identified on the basis of syntactic characteristics imposed on nouns, usually by verbs or prepositions. The following cases exist in the Polish language:

Case	Examples	
nominative	dom	domy
genitive	domu	domów
dative	domowi	domom
accusative	dom	domy
instrumental	domem	domami
locative	(pronoun +) domu	(pronoun +) domach

Locative case always occurs with a preposition, for instance *w domu*[2] 'at home', *o domu* 'about home'. This is a not a stand-alone case.

[2] The locative case takes no preposition in Lithuanian, for instance: Polish *w domu* – Lithuanian *namie* / *namuose*.

In the linguistic tradition vocative is considered as one of the cases. However, the use of vocative in a text does not confirm the existence of any government imposed on the vocative form by a verb or a preposition. Therefore, vocative should be viewed as a separate category. However, in MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) vocative is classified as one of the cases, which is why (according to the fallacious tradition) vocative is included here as another case in Polish, as in the example below:

Case	Examples
vocative	domu domy

3.2 Number

The identification of number is based on the semantic difference between a single object: (distributive = non-collective) set of objects, e.g. *dom* 'home' : *domy* 'homes'. The grammatical category of number sometimes slightly deviates from the aforementioned semantic opposition. Some nouns do not offer such a distinction, for instance, the so-called plurale tantum: *nożyczki*, 'scissors', *drzwi* 'door', *parzystokopytne*, *małżonkowie* (= *małżonka 'wife'* + *małżonek 'husband'*), *narzeczeni* (*narzeczona* 'fiancée' + *narzeczony* 'fiancé'), *Wadowice* (proper name). In fact, Polish has no singulare tantum nouns, yet they are sometimes mentioned in literature. One example of singulare tantum is *fizyka* 'physics' or *pierze* 'plumage' (the so-called collective nouns). However, for any singulare tantum a plural form may be created and a use for it may be found, as noted by Saloni (Saloni in: [7]).

There is no need to introduce the dual number in Polish. Contemporary Polish has few dual forms (e.g. in instrumental or locative case) for selected paired bodily parts such as hands, eyes or ears:

Case	Examples
instrumental	rękami/rękoma oczami/oczyma
locative	rękach/ręku

They should be treated as variants of plural forms.

3.3 Gender

Gender is identified on the basis of syntactic properties associated with the requirement that a specific form must occur next to a word that combines with a noun. Initially, gender was presumably a semantic category for some nouns which then spread onto all nominal and pronominal forms in language. The category of gender in nouns is selective.[3] All nouns in Polish have a fixed gender.[4] Traditionally, the following genders have been distinguished: masculine, feminine, neuter. Polish has all of these three genders, for instance: *dom* 'home' (masculine), *książka* 'book' (feminine), *dziecko* 'child' (neuter). This distinction into genders is specific to nouns in singular whereas the traditional notion of masculine, feminine and neuter is blurred in plural. In fact, one might talk about two groups of nouns in plural. The first group comprises masculine nouns that are human and pluralia tantum that are human. The second group covers all other plural forms of masculine, feminine and neuter nouns as well as pluralia tantum that are not human.

Apart from innovations in pluralis there are new phenomena in Polish which are characteristic of some classes of singular masculine and neuter nouns. The nature of those new phenomena is syntactic. Therefore, much as gender, the new phenomena are associated with the requirement to adopt a particular form in adjacency to words that combine with nouns.

In the Polish linguistic tradition, initiated by Witold Mańczak [4], three masculine genders are distinguished. They are also adopted by Saloni as three subgenders (Saloni in: [7]). In my view, an alternative solution is possible once we have introduced new categories, i.e. human and animate. In that case the gender classification adopted within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004) will be retained:

[3] Except for lexeme *on* 'he, she, it, they'.

[4] A small group of nouns may have no definite gender.

Type	Gender	Number	Case	Human	Animate	Examples		
comon	masculine	singular	nominative			profesor	pies	dom
			genitive			profesora	psa	domu
			dative			profesorowi	psu	domowi
			accusative	+		profesora		
			accusative		+		psa	
			accusative					dom
			instrumental			profesorem	psem	domem
			locative			profesorze	psie	domu
			vocative			profesorze	psie	domu

Type	Gender	Number	Case	Human	Animate	Examples		
comon	–	plural	nominative	+		profesorowie / / profesorzy		
			nominative				psy	domy
			genitive			profesorów	psów	domów
			dative			profesorom	psom	domom
			accusative	+		profesorów		
			accusative				psy	domy
			instrumental			profesorami	psami	domami
			locative			profesorach	psach	domach
			vocative	+		profesorowie / / profesorzy		
			vocative				psy	domy

Type	Gender	Number	Case	Human	Animate	Examples		
proper	masculine	singular	nominative			Roman	Burek (dog)	Płock
			genitive			Romana	Burka	Płocka
			dative			Romanowi	Burkowi	Płockowi
			accusative	+		Romana		
			accusative		+		Burka	
			accusative					Płock
			instrumental			Burkiem	psem	Płockiem
			locative			Romanie	Burku	Płocku
			vocative			Romanie	Burku	Płocku

Type	Gender	Number	Case	Human	Animate	Examples		
proper	–	plural	nominative	+		Romanowie / / Romany		
			nominative				Burki	Płocki
			genitive			Romanów	Burków	Płocków
			dative			Romanom	Burkom	Płockom
			accusative	+		Romanów		
			accusative				Burki	Płocki
			instrumental			Romanami	Burkami	Płockami
			locative			Romanach	Burkach	Płockach
			vocative	+		Romanowie / / Romany		
			vocative				Burki	Płocki

As shown in the table above, the ‘human’ category is visible in accusative singular and in accusative (and vocative) plural whereas ‘animate’ is visible only in accusative singular, as in the examples below:

Patrzę na profesora (acc = gen), *na psa* (acc = gen), *na dom* (acc = nom).

‘I am looking at a professor, a dog, a house.’

Patrzę na profesorów (acc = gen), *na psy* (acc = nom), *na domy* (acc = nom).
'I am looking at professors, dogs, houses.'

The information in brackets shows the characteristics coincidences between accusative and either nominative or genitive, depending on whether the noun is human or animate.

As far as lexemes of neuter gender are concerned, a split in collocation occurs only when such lexemes combine with numerals. The first group is small and has the following collocation pattern with numerals: ***czworo*** *szczeniąt*, ***troje*** *dzieci*. The second group is much more numerous and strongly supersedes the former group. Examples of collocations in this case are: ***cztery*** *pola*, ***trzy*** *lata*.

3.4 Depreciativeness

The category of depreciativeness is identified on the basis of syntactic properties associated with the requirement for a word to occur in a particular form next to a word that combines with a noun. In two cases in plural, nominative and vocative, (the latter being always identical with nominative) some masculine nouns have two forms that are used in parallel, e.g. *chłopacy* 'boys' and *chłopaki* '[contemptuously about] boys'. If it were not for syntactic differences associated with the use of one or the other form, one could talk about the existence of multivariants and so another subcategory[5] of depreciativeness would not need to be introduced:
*To są siln**i** chłopa**cy***.
and
*To są siln**e** chłopa**ki***. (both: 'These are strong boys')

We agree with Saloni that the subcategory of depreciativeness is an inflective one and one that enforces differing syntactic consequences.

As a rule, non-depreciative forms are neutral and considered to be basic. Depreciative forms should be seen as negatively marked, used to show a certain degree of disrespect. As usual in such cases, there are some exceptions such as neutralisation or even a reversal of marking, as described by Saloni. Let us add, however, that in some regions of Poland depreciative forms of some masculine human nouns (high-frequency ones) are considered neutral and are widely used.

The table below is an updated version of the relevant elements from the table provided in Section 3.3.:

Type	Gender	Number	Case	Human	Animate	Deprecjatywność	Examples
comon	–	plural	nominative	+			profesorowie /
				+			/ profesorzy
			nominative			+	profesory
			vocative	+			profesorowie /
				+			/ profesorzy
			vocative			+	profesory

The category of depreciativeness occurs in the group of masculine nouns which have the attribute of 'human'.

We do not think it is valid to introduce a separate category for nouns traditionally termed as bi-gendered, such as *ciapa* 'slowcoach', *łamaga* 'butterfingers', *niezdara* 'fumbler'.[6] The basis for distinguishing bi-genderness could only be semantic in this case. However, we view these forms as homonymous, i.e. *łamaga* described as masculine-human, and *łamaga* as feminine.

Notably, according to traditional descriptions the group of bi-gendered nouns also includes forms such as *psycholog* 'psychologist', *sędzia* 'judge' and many other. Such examples can be also described as homonymous forms of masculine-human and feminine. Recently, there has been a strong trend towards providing a formal distinction between such forms: *psycholog* – masculine-human and *psycholożka* – feminine, *sędzia* – masculine-human and *sędzina* – feminine[7].

[5] In this paper we use the term 'category' interchangeably with 'subcategory'.
[6] A similar phenomenon is also found in Lithuanian.
[7] In recent years Lithuanian has shown quite the opposite trend, i.e. equalisation of formal differences between such masculine and feminine forms. I see this phenomenon as an obvious influence of the so-called Western languages.

3.5 Uniformism

The category of uniformism, as distinguished by Saloni, is associated with stylistic differentiation of selected feminine nouns in genitive plural, for instance *kopalni* (neutral form) / *kopalń* 'coal mines'. As this differentiation does not affect syntactic relations in the sentence, we do not believe it is justified to include this category into the set of morphosyntactic characteristics. Moreover, the majority of marked forms are clearly formal or even archaic, sometimes bordering on humorous, as is the case with *racji* (neutral form) / *racyj* 'reasons'. For the purposes of the MULTTEXT East description, information about the existence of variants is sufficient.

3.6 The so-called nominal pronouns

Neither the contemporary grammar of Polish ([2]) nor Saloni (Saloni in: [7]) distinguish pronouns as a separate part of speech. Lexemes which were traditionally regarded as pronouns have been allocated to various classes based on semantic and syntactic criteria, respectively: nouns (e.g. *ja* 'I'), adjectives (e.g. *ten* 'this'), numerals (e.g. *wiele* 'many') and adverbs (e.g. *tam* 'there').

Pronouns included in the class of nouns have different morphosyntactic characteristics. Consequently, it was necessary to identify subgroups of nominal pronouns:

1. – This subgroup includes singularia tantum such as: *kto* 'who', *co* 'what', *cóż* 'whatever', *któż* 'whoever', *nikt* 'nobody', *nic* 'nothing', *to* 'this one', *tamto* 'that one', *owo* 'other', *wszystko* 'everything' and pluralia tantum such as *my* 'we', *wy* 'you', *wszyscy* 'everyone'. One characteristic of some forms within this group is that there are two forms of genitive case, their use being connected with occurrence after a preposition (more in Section 3.6.1 below).

2. – This subgroup includes singularia tantum: *ja* 'I', *ty* 'you' and *się*$_1$ 'self'. However, *się*$_1$ does not have a nominative form. This group is characterised by stressability (more in Section 3.6.2 below).

3. – This subgroup consists only of lexeme *on* 'he'. One characteristic of this lexeme is its inflective category of gender: *on* (masculine), *ona* (feminine), *ono* (neuter). This characteristic remains in contradiction to the initial assumption about selective gender. Nevertheless, as syntactic functions played a prevailing role, this lexeme was classified as a noun. Other characteristics of this group include post-prepositionality and stressability (more in Sections 3.6.1–2 below and example in Section 3.6.3).

4. – This subgroup covers uninflected lexemes: *toto, niecoś, śmo*, used in nominative and accusative; and *ichmość, wasze, się*$_2$ used in nominative. The form *się*$_2$ is the only one which combines with verbs that require a nominative form, for instance: *Układa się puzzle*; *Wybijało się szyby*.

Post-prepositionality This attribute is characteristic of a small number of lexemes: *co* 'what', *cóż* 'whatever', *nic* 'nothing, *on* 'he'. These lexemes have two forms of genitive, dative and accusative each. A selection of one of the two variants depends on whether this case form is required by a verb or a preposition.

Stressability This attribute characterises a small number of lexemes: *ja* 'I', *ty* 'you', *się*$_1$ 'self', *on* 'he'. These lexemes have two forms of genitive, dative and accusative, one of which is stressed. The other one is unstressed and is viewed as an enclitic: it forms a phonological word together with the preceding lexeme.

Post-prepositionality and stressability in examples Attributes of the lexeme *on* are as follows: post-prepositionality, stressability, case, number and inflective gender. Therefore, we will use this lexeme to demonstrate a paradigm representing key characteristics of the so-called nominal pronouns.

The source of those changes lies in the consistent formal differentiation of surnames (bearing administrative consequences, i.e. names written in passports), for instance Marcinka (masculine), Marcinkienė (feminine, a married woman, wife of Marcinka), Marcinkaitė (feminine, daughter of Mr. and Mrs. Marcinkai)

Type	Gender	Number	Case	Hum.	Anim.	Depr.	Post-Prep.	Stress.	Examples
common	masculine	singular	nominative						on
			genitive					+	jego
			genitive						go
			genitive				+	+	niego
			genitive				+		+ń
			dative					+	jemu
			dative						mu
			dative				+		niemu
			accusative					+	jego
			accusative						go
			accusative				+	+	niego
			accusative						+ń
			instrumental						nim
			locative						nim
			vocative						

Legend:

Hum. – Human
Anim. – Animate
Depr. – Depreciativeness
Post-Prep. – Post-prepositionality
Stress. – Stressability

Type	Gender	Number	Case	Hum.	Anim.	Depr.	Post-Prep.	Stress.	Examples
common	feminine	singular	nominative						ona
			genitive						jej
			genitive				+		niej
			dative						jej
			dative				+		niej
			accusative						ją
			accusative				+		nią
			instrumental						nią
			locative						nią
			vocative						

Type	Gender	Number	Case	Hum.	Anim.	Depr.	Post-Prep.	Stress.	Examples
common	neuter	singular	nominative						ono
			genitive					+	jego
			genitive						go
			genitive				+		niego
			dative					+	jemu
			dative						mu
			dative				+		niemu
			accusative						je
			accusative				+		nie
			instrumental						nim
			locative						nim
			vocative						

Type	Gender	Number	Case	Hum.	Anim.	Depr.	Post-Prep.	Stress.	Examples
common	–	plural	nominative	+					oni
			nominative			+			one
			genitive						ich
			genitive				+		nich
			dative						im
			dative					+	nim
			accusative	+					ich
			accusative	+			+		nich
			accusative						je
			instrumental						nimi
			locative						nich
			vocative						

References

[1] Bąk, P. (1977/2007). *Gramatyka języka polskiego*. Wiedza Powszechna, Warszawa.

[2] Grzegorczyk, R., Laskowski, R., Wróbel, H. eds. (1984/1998). *Gramatyka współczesnego języka polskiego*. Wiedza Powszechna, Warszawa.

[3] Koseska, V., Roszko, R. (2008). Remarks on classification of parts of speech and classifiers in an electronic dictionary. In *Lexicographic tools and techniques, Mondilex first open workshop, Moscow, Russia, 3–4 October, 2008, Proceedings*, pages 80–88, Moscow. Russian Academy of Sciences. Institute for Information Transmission Problems (Kharkevich Institute).

[4] Mańczak, W. (1956). Ile rodzajów jest w polskim? *Język Polski*, 1956(z.2):116–121.

[5] Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 2007(11):151–167.

[6] Polański, K. (eds.) (1999). *Encyklopedia językoznawstwa ogólnego*. Zakład Narodowy imienia Ossolińskich. Wydawnictwo, Wrocław – Warszawa – Kraków.

[7] Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R. (2007). *Słownik gramatyczny języka polskiego. Podstawy teoretyczne. Instrukcja użytkowania*. Wiedza Powszechna, Warszawa.

[8] Tokarski, J. (2002). *Schematyczny indeks a tergo polskich form wyrazowych*. Saloni, Z. (eds.), wyd. 2, Wiedza Powszechna, Warszawa.

Theory of Lexicographic Systems. Part 1.

Volodymyr A. Shyrokov

Ukrainian Linguo-Information Fund, National Academy of Sciences of Sciences

Abstract. The main foundations of the theory of lexicographic system have been developed. Information effects of natural languages have been researched and on this basis the lexicographic effect in information systems is stated. The latter makes a conceptual ground for building the formalized theory of lexicographic systems and lexicographic data models.

Theory of lexicographic systems provides a basic conceptual scheme for all the works of Ukrainian Lingua-Information Fund. In the lexicographic description of a language system, as far as may be inferred from our experience, the logical-linguistic status of this theory plays roughly the same role that the theory of formal grammars does in the grammatical description of language system.

1. Lexicographic effects in information systems

Theory of lexicographic systems has its phenomenological basis in the so-called lexicographic effect in information systems described in our works [1]. Let us dwell on its content.

Processes of lexicographicalization as a kind of intellectual activity and the phenomenology of dictionaries to serve a result of this activity are not constant over time. They evolve in accordance with the internal development of linguistic science and needs of practice. In some historical periods the factors external to linguistics' own objectives, as has not once happened in the history of science, are the main driving force to determine not only the development of lexicography as a separate section of linguistics but also the science of language in general. The direction and pace of the evolution of information and communication technology assures that the development will have a further impact on the progress of the information science. The development of the information society towards the knowledge society inspires us with confidence in this analysis.

Indeed, with the advent of computer technology a phenomenon unprecedented in the history of the world civilization has emerged, i.e. communication of human beings with non-living objects through the natural human language. However sceptical we may be in respect of the "mental potency" of computers, we cannot ignore the fact that linguistic reactions of modern computers in some cases are no more distinguishable from reactions of human. Taking into account the progress of language technologies during the recent twenty years, we can predict with certainty that computers will before long assume a substantial part of linguistic competence of human beings. In turn, this would create real prerequisites for the construction of computer systems based on natural language.

The above competence is an important component of the human thought-speech apparatus, and in the latter the linguistic structures are inextricably linked with the structures of thinking as already firmly established in psycho- and neurolinguistics. The mentioned relationship is considered to be so essential that it justifies the definition of intelligence as a form of personalization of a system having a linguistic status [2]. This definition can be naturally extrapolated on artificial intelligence systems. Thus, in the considered context the computer simulation of language is congenial and almost identical to the modeling of intelligence.

What role does the dictionary play in this concern? As noted by Robert Schenk [3], in the human thought-speech apparatus different "dictionaries" operate to be even more similar to "encyclopedias".

Studies carried out in the framework of the WordNet [4] methodology and based on principles of psycholinguistics not only confirm this thesis, but also throw light on the design of structures of the human subjective lexicon.

The fact that there is in principle the possibility of modeling human speech in non-living objects, such as computers, give us the idea that also in the nature of matter some mechanisms act to have common features with human language. Hence the phenomena which can be described as a manifestation of language, are not necessarily the exclusive prerogative of human (and even of any living creatures). Man in casting antropomorphous features on everything around him assigns the name of language to numerous events. A lot of the images are created in mythology and poetry where non-living substances are metaphorically endowed with the gift of speech. In general, the class of phenomena and objects characterized as linguistic is very diverse. First, it is the natural language that exists in the form of national languages. Now there are about six thousand languages, and often the relationships between them become dramatic [5]. There are also other natural semiotics-semantic systems, which somewhat metaphorically can be defined as "language" (e.g., "language" of genetic code). In addition, a number of artifacts of linguistic orientation, e.g. artificial languages like Esperanto that imitate certain "natural" languages, formal algebro-algorithmic structures known as formal languages recently drawing more and more closer to the natural languages (in mathematical linguistics there is even the term "almost a natural language"). The term system „Informatics" which embraces such items as programming language, information retrieval language, language classifier, query language, data description language, data manipulation language, data domain description language, a number of so-called "mark-up languages" (SGML, HTML, XML, VRML etc.) is just another prove of the prevalence of the term "language" in information science. In 1960s to 1970s an entire branch of informatics defined as the linguistic support of information systems did emerge.

Artificial interpretations of natural languages, in particular their written versions, play a significant role in culture and civilization, and sometimes the creators of artificial languages are so impressed with some features of natural languages that they make structures of their brainchildren inherit those properties from "usual" human languages.

All the above "linguistic" systems have common features of a basic, fundamental nature directly related to the definition of lexicographic processes and structures. For them to envision and to build a workable conceptual framework, we need a generalized notion of language that could be applied to any of the mentioned "linguistic" systems, because in the modern information environment, language ceases to be solely a prerogative of man, at least at a "technological" level.

What does linguistics say about this?

It asserts that to define the language is a difficult task since there are a lot – and diverse – of its definitions based on the different aspects of this multifaceted phenomenon. Despite their diversity, by summarizing their essential features, it is possible to conclude that most of them are variations of a theme proposed by W.Chafe [6]: language is a system that carries out a connection between the world of sounds and the world of meanings in a fairly sophisticated way.

Note that there are many other definitions of the language. A lot of them has since the time of W. von Humboldt been based on delimitation of the concepts of language and speech. After the publication of the book by F. de Saussure "Course of General Linguistics" this issue has become popular in linguistic circles, and the varied range of opinions keeps on fluctuating in an extremely wide range. They can be summarized in three main assertions: 1) speech and language are opposed to each other as completely autonomous objects that differ in a set of essential features, so that two separate areas of science – linguistics of language and linguistics of speech – are involved in their study, 2) language and speech are a single object of linguistics, and interpretation of the differences between them lies in the methodology and content of this complex science, and 3) between language and speech there is no difference at all.

L. Shcherba [7], as we know, singled out three main groups of linguistic phenomena. Namely, the first – the speech activity, the second – dictionaries and grammars compiled on the basis of speech recognition and understanding relevant to a certain historical period among certain groups of people, i.e. linguistic

systems, the third – all what those groups say or mean about – the speech material. He stressed that the speech is due to the complex linguistic apparatus of human or the individual's psychophysiological language organization having certain properties. Specifically, the linguistic organization:

a) cannot equal the sum of speech experience and must be a kind of its processing;

b) can be nothing else but a psychophysiological organization;

c) is a social product as well as speech itself;

d) serves as an individual manifestation of the language system as a result on the strength of the linguistic material;

e) the nature of this organization can be judged only on the basis of the speech activity of individuals.

L.V. Shcherba distinguished the notion of the mechanism (of speech) and the process (speech activity); the process and its product. The latter serves as an individual system of concepts and strategies used by individuals in the process of speaking and understanding, which is referred to as *language*.

We do not engage in a discussion about the correctness of any opinions and the extent of their compliance with the reality, because they all contain these or other features of the phenomenon we use the generalized name "language" for, features varying in different combinations in a large amount of works in general linguistics. We omit their description, because we believe that a correspondence between the world of sounds and the world of meanings lies in their basis explicitly or implicitly. Nor take we into account the definition of "language is the soul of the people" and the like, because they are unable to explore with scientific methods.

Note that the definitions we have dealt with are difficult to use for building a productive pattern formalized at least minimally. If the „world of sounds", can be somehow „localized", "the world of meanings" is much more complicated to treat. Indeed, where is this world focused? How can we get to it? How could it be „handled"? Actually, what sense is in the assertion about the existence of the "world of meanings"? How do the "world of contents", „world of images", and many other "worlds" relate to the "world of meanings"? So, by uniting notions strongly differing in the degree of abstraction ("the world of sounds" and the "world of concepts") in one definition, you cause an impression of a logical gap in this definition and raise more questions than give answers.

Nevertheless, the outlined version, and other attempts to determine the language, we consider to be useful, since each of them provides the material for the synthesis and the exposure of essential features of such a universal phenomenon as language. Having analyzed and looked ("listened") closely to a certain phenomenon we intuitively identify as a "language", taking into account "what it does" and "how it works", you can draw some conclusions. Language is a sort of "tool" (a kind of machine), to ensure the conversion of "forms" into the "content" and vice versa. But it concerns not arbitrary "forms" but "forms" of linear sequences of certain discrete objects (sounds and sound complexes, signs and signal complexes, etc.).

This assertion, not at all notable for its novelty [8], provides the basis for getting deeper into the relationship of phenomenology "form – content" (we denote it by RFC) and to find out its details to help us in revealing significant features of the language. In doing so, we believe that in its ontological dimension the RFC is not an a priori qualitative inherent property of the object as "a thing in itself", but rather a property that is disclosed ("given") to the subject in his/her interaction with the object. We strive to build a formal description of the RFC adapted to the creation of a specialized data model to represent effective procedures of exposure of essential properties of language with a technological orientation to support of the creation of dictionaries and other linguistic products. To analyze the details of the RFC displaying itself consider the chart which symbolically depicts the process of perception of an object by a subject:

$$S : D \to V(D). \tag{1.1}$$

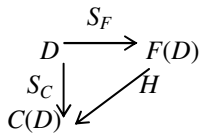
Here the letter D indicates something from the real (or imaginary) world, that serves as an object of perception (observation, study, attention, emotional experience, ...) on the part of some *S*, which we believe to be the subject of this process, the *V(D)* denotes the result of the process. Note that the *S* may be

either a person or a device designed by man or a man-machine system, or even something else endowed with qualities of perception and feeling ("reflection"). *S* may be a "collective entity" – a group of people, a social community, an ethnic group, a nation, a people, a collection of peoples or even the humanity in general.

Due to the physical, mental, intellectual and other limitations of the subject *S* all the properties of the object *D* for its perception are divided into two not very clear, ambiguous, volatile, and not entirely differentiable parts. To the first one, we reckon those properties of *D* directly perceived by the "sensory-perceptual" apparatus of *S*. Denote this part by the *F(D)*, and treat it as a set of formal properties of *D* in terms of the perceiving subject *S.* The second part contains those properties of *D,* which are not directly perceived by the perceptual-sensory apparatus of *S* but reflected in it indirectly. Denote this part by the *P(D)*, and we will treat it as a set of meaningful properties of *D* – again, in terms of perception by the subject *S.* In connection with this diagram (1.1) takes the following form:

$$D \xrightarrow{S_F} F(D) \xrightarrow{H} C(D), \tag{1.2}$$

where the symbol S_F designates the action of the "sensory-perceptual" apparatus of the subject *S,* The result of the action is a set of formal (in terms of *S*) properties of *D;* symbol *H* denotes a procedure to implement the connection between form and content and ensuring the integrity of the perception of an object *D* by the subject *S* (if it really succeeds in ensuring the above integrity). At the same time, allowing the existence of a procedure to enable the transition from *D* to *C(D)*, and defining the procedure by S_C,, we obtain a transformation of the diagram (1.2):



(1.3),

where, as we can see, there has taken place a "decomposition" of the subject *S* to S_F and S_C, that reconstruct the formal and semantic properties, respectively.

We are not inclined to absolutize the above pattern, because there is no clear boundary between *F(D)* and *C(D)*, as it does not exist really between form and content. Also, the properties of *S* are almost not examined in detail, though the decomposition of *S* to S_F and S_C has been made on general considerations. Thus, this approach is in all its signs phenomenological since it does not rely on assumptions about the possible "construction" of *S* and procedures for its functioning. Based on these considerations, it can be argued that the presented scheme is quite general – it does not set any „anzatzes" except for a single, specific feature: *F(D)* should be linear and therefore implemented by linear sequences of discrete objects a source of them is a certain finite set.

Having regard to the above the very possibility of the existence of such a phenomenon as the language results from the fundamental properties of *S* "to be a subject", i.e. someone for whom anything has its external side (form) and internal one (content).

The relationship between these different aspects of perception, symbolically shown by values S_F, S_C, *H*, varies considerably due to some properties fundamentally inherent in the perceiving subject S: variability, irregularity, variety, limitedness, fuzziness etc. For example, the shape of things for a substance capable of perceiving the world in the X-ray range of electromagnetic waves and in the ultrasonic range of mechanical vibrations, would be significantly different from our perception. Also, many properties of the things that we, with our inherent sensory-perceptual system, consider to be a "content", the hypothetical substance would be perceived as a "form".

An important aspect of the RFC is connected with the property of "attention" concentration by the subject of S on fragments of both the form and the content. This is achieved by "tuning" his/her perceptual-sensory systems on the details of what he/she perceives. Consequently, the initial RFC is modified: its certain semantic elements acquire properties of a form and to its formal elements some new details, previously unnoticed can be added. This class of properties embrace, in particular, those concepts

of internal and external forms of linguistic units so beloved by followers of the Humboldt-Potebnya school.

Note one more feature of the process of the RFC development. Contemporary culturology tends to treat it as a manifestation of post-modernism – namely, the influence of the subject to the object, i.e. the possibility of changing the state of an object D in the process of perception (observation, research ...) by the subject S. The point is that in order to set up the process symbolically depicted on the diagram (1.3), in many cases we need to "strengthen" the object D for it to "demonstrate" those of its qualities S "is interested in". In the classical scientific paradigm it was thought that such an initiation of the object can be made arbitrarily small and neglect it, believing that it does not significantly alter the state of the object. However, the development of science has shown that it is not so [9].

All described above, is entirely consistent with quite different phenomena and their formal models built for other reasons and other purposes. This is the definition of information, A.N. Kolmogorov [10], where the interaction takes place "content" with "shape" of linear sequences.

The introduction of the Kolmogorov's information measure means to specify the definition of information, firstly, without involvement of the probabilistic approach, and secondly, to make it possible to apply the measure to individual objects.

The main idea of the approach is that information about the object is considered to be obtained when a rearrangement of the object (model) is possible according to its final description (set of attributes). Building up the Kolmogorov's measure is based on such fundamental notions, as the algorithm, the Turing machine, the recursive function, and is derived from the ideas of the theory of computational complexity (complexity of algorhythms), which actually is a source for interpreting the information as a measure of complexity and structuredness of systems. Besides, the category of complexity is believed to be universal since any system, regardless of its nature, is characterized with some complexity and has a certain structure.

The relevant mathematical construction if not overburdened with details can be formulated as follows.

Consider some countable set of $X = \{x\}$. Let us assume that there is a isomorphism between X and the set D of binary words to begin with unity. In other words, let a bijective mapping be set:

$$n: X \rightarrow D, \tag{1.4}$$

so that to each $x \in X$ some $d = n(x)$, $d \in D$ corresponds, and vice versa. We consider that:

1. n (x) – general recursive function on D. Denote by $l(d)$ the length of the binary word $d \in D$, i.e. the number of zeros and ones in it. Then $l(n(x)) = l(x) + C$, where C – is a certain constant.

2. There is an monomorphism χ: $X^2 = X \times X \rightarrow X$, such one that for $\forall\ x \in X$, $y \in X$ $\exists z \in X$ that $z = \chi(x,y) \equiv (x,y)$, and $n(z) = n(x,y)$:

$$l(x,y) \le C_x + l(y),$$

where the constant C_x depends on x only.

Let us consider an isomorphism (1.4) ascertained, so that the set X will also be considered as a set of binary words.

Assume that there exists a general recursive function $\varphi(p, x)$, which brings binary word y to a binary word x, and at that p, $p \in D$ is interpreted as an algorithm (or a program), that "converts" x to y:

$$p: x \rightarrow y, \tag{1.5}$$

and φ presents here a method (a programming language). Without loss of generality assume that p for the given x is set by a certain binary word.

Denote:

$$K_\varphi(y|x) = \begin{cases} \min|_p\, l(p), \text{ if } \varphi(p, x) = y \\ \infty, \text{ unless a finite } p \text{ exists such one that } \varphi(p, x) = y \end{cases} \tag{1.6}$$

Thus, $K_\varphi(y|x)$ is the length of a minimum program p, that converts x to y under the specified method of programming. This value is called the complexity of y relative to x at a given φ. Of course, the dependence of the complexity's magnitude on φ is a drawback of the algorithm, but there is a theorem [11]

that maintains the existence of the "best" method of programming *A,* so that for any partial recursive function φ the inequality below is true:

$$K_A(y|x) \le K_\varphi(y|x) + C_\varphi, \tag{1.7}$$

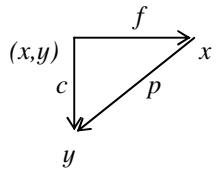
where the constant C_φ depends on φ only and does not depend on x and y.

The value $K_A(y) \equiv K_A(y|1)$ "normalized" relative to a singular element $x = 1$ is naturally considered to be the complexity of the elements *y*. At that the amount of information in the object *x* relative to the object *y* is defined as a difference:

$$I_A(x|y) = K_A(y) - K_A(y|x)\ . \tag{1.8}$$

That is the latter formula which defines a measure of information – the so-called Kolmogorov's algorithmic measure of information.

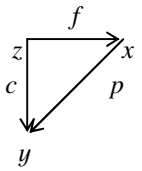
Formula (1.8) and the described approach for determining information via algorithmic complexity as a whole can be interpreted somewhat otherwise. To this end let us draw a triangular diagram:



(1.9)

where *x* and $y \in X$; the element $(x,y) \in X^2$, and, due to the existence of display $\chi : X^2 \to X$, $\chi(x,y) = z$, $z \in X$. On the diagram (1.9) the mappings *f* and *c* carry out the projection of the element *(x,y)* on the first and second factors, respectively, and the formulas (1.6) – (1.8) and the interpretation of complexity as an information measure are true.

Suppose now that an object z, representing the image of the Cartesian product *(x,y)*, at the mapping χ is in reality an independent object of the outside world, depending on neither *x* nor *y*. This assumption allows us to construct the following chart:



(1.10)

where $z \in Z$, Z – is the set, source of objects *z*, $x \in X$, $y \in Y$. Then the mappings f and *c* determine some interpretations of the object *z*, and, moreover, the mapping p interprets *x* via *y*. It is naturally to assume that the object *x* reflects "formal" properties of the object *z*, and *y* – its "substantial" properties, and the relationship between "form" and "content" is carried out by *p*. The requirement of minimality for *p* is here quite natural, because the "interpretation" of the form (the content is its result, in fact) must not contain "superfluous" components, random in respect of *x* (and *z*). The construct designed in this way and based on a triplet of objects (*z, x, y*) and a triplet of mappings (*f, c, p*) that form the diagram (1.10) admits a natural interpretation as a complex medium of the RFC.

So, the RFC proves to be "encrypted" in the algorithmic definition of complexity and information in this peculiar way and therefore, this ratio is closely related to the information processes and moreover, it is, as we understand it, a system-forming relationship. At the same time, it is a basic system-forming relationship of natural language. Consequently, the thought-speech objects, processes, constructs and structures are of information nature and for making their qualitative and quantitative analysis the Kolmogorov's formalism is quite applicable. The apparent similarity, affinity of the diagram (1.3) and the Kolmogorov's definition of information, as well as further considerations that have led to the construction of the diagram (1.10), suggest that the same regularities underlie both. The very form of representation of the information measure indicates a certain process that has led to the generation of an "alphabet" – a sign system to represent the object. The opening of mappings (*f, c, p*) in diagrams (1.9) – (1.10), to be compared with the elements of the RFC in diagram (1.3), incites us to confront the latter ones to the

components of the information processes, which in the Kolmogorov's theory are reduced to mathematic relations (algorhithms, recursive functions ,...), defined on discrete sets.

The described situation is considered to be so general that allows us to make a conclusion: in the grounds of any process, phenomenon, object, system and so on there lies a certain discrete class, we call it a class of elementary information units (EIU). In its definition the key role is played by the notion of the lexicographic effect in information systems, the essence of which is as follows.

A common feature of all processes and information exchange is transformation of information from one form to another, and the modern natural-science theories inambiguously confirm that the process of interaction and exchange is discrete ("quantum") in essence and therefore the process of description of reality undergoes to discretization in principle. The indicated discretization has at least one feature common to all known processes, namely: observing and generalizing the behavior of different systems we come to the conclusion that the process of evolution (dynamics, self-development) induces in a system of any nature a certain subsystem of relatively stable discrete essences (a subsystem of order) to act as the system's *elementary information units*, so that all the other fenomena are nothing more than combinations of these elementary information units arranged in a certain way. To illustrate we cite some examples.

Despite the fact that according to the modern concepts the universe is placed in the four-dimensional space-time continuum, all the observed values are essentially quantized. They on the whole depend on a small number of so-called "global constants" (Planck's constant, the speed of light, charge and mass of the electron, etc.), which make a kind of "alphabet" for physics, by means of which all the meaningful assertions about the behavior of physical systems (values of observed variables) are formulated. A similar situation has occurred also in the scientific description of other systems: all the substances are designated by certain «words» in the "alphabet" of chemical elements and their mutual transformations make "statements" in this "language". Proteins are mainly composed of residues of 20 amino acids, DNA molecules are constructed from four types of nucleotides, etc. The observed behavior is typical not only for theories that describe specific real-world objects (natural and technical), but also for the concepts to operate ideal objects, abstractions, speculative constructs. Indeed, even a description of the processes of discretization, signal quantization, Kotelnikov theorem conclusions can be interpreted not only as the possibility of producing a certain continual universe via discrete sets. It can be thought of as a law of nature, affirming the bound existence for any system of such a discrete subsystem of relatively permanent entities (in our terminology, a subsystem of EIU), which as a medium implements the possibility to submit any fact from the universe involved.

The mentioned subsystem has properties related to properties of the lexicographic subsystem of the natural language: it "generates" in its structure something like a sort of thesaurus and grammar with all properties inherent in such constructs: sign nature, meaning, content, polymorphism, homonymy (isomorphism), synonymy (homomorphism), polysemy (multiple meaning), ellipsis, metonymy and so on; it is the bearer of both "plan of expression" and "plan of content". Realization, interaction, mutual influence, interdependence of both plans in the system of basic information units are subject to certain regularities and the centuries-old controversy between realists and conventionalists in our opinion should be solved positively: examples of system of both types can be adduced, both those with the plan of expression determined by the plan of content and those with the relationship of a conventional type.

Classes of elementary information units like all the aggregates determined by objective processes (here it is the lexicographic effect) have a property of "substantiality", owing to which the indicated bodies possess relatively stable features to secure their localization in corresponding areas of system parameters.

The above description of phenomena makes the content of the lexicographic effect. On can assert that when examining any object domains experts in fact deal with examining lexicographic effects that occur in those domains or are typical for them. Thus, the lexicographic effect can be regarded not only from the phenomenological side but form the point of methodogy since it has a certain "potential of rationality" and stimulates in the course of modelling of phenomena the ascertainment and determination of the relevant EIU complexes taking into account their specific properties. In this aspect the concept of lexicographic effect appears to be a method of data abstracting.

We will not deal with the question of origin, type and classification of lexicographic effects, because much in this area is still unexplored. It is clear that their origin is connected with the processes of ordering and disordering of matter that relate to synergistic effects structure, complexity, heterogeneity of matter, i.e. characteristics of congenial information. Typology lexicographic effects, if it was known, opened the way for the construction of classification schemes of elementary information units.

The described set of phenomena is the content of lexicographic effect. It can be argued that when studying any subject domain specialists actually study the lexicographic effects occuring or specific to them. Thus, the lexicographic effect must be considered not only from the phenomenal part, but also from a methodological, as it has a certain "operational capacity" by stimulating the processes of modeling of various systems of establishing and defining the relevant sets of EIU, given and specifying their properties. In this hypostasis the concept of lexicographic effect serves as a method of abstracting the data.

Further on, a class of EIU of system D, which has evolved as a result of a lexicographic effect (or effects) *Q,* we denote by $I^{Q}(D)$ or *I(D)*, when the type of lexicographic effect is insignificant. The EIU system, being the carrier of a number of properties, has a certain structure. In particular, in any system of EIU a kernel can always be singled out as a subsystem $I_0{}^{Q}(D) \subseteq I^{Q}(D)$ and a generative procedure π is defined:

$$\pi: I_0{}^{Q}(D) \to I^{Q}(D). \tag{1.11}$$

We identify the triplet $(I^{Q}(D), I_0{}^{Q}(D), \pi)$ with the EIU system, and use this designation, as $I^{Q}(D)$, $I(D)$, $I_0{}^{Q}(D)$, $I_0(D)$ as equivalents, believing that generative procedure π is defined, known and understood from the context. We here give two examples to illustrate the definition (1.11).

For the set of chemical elements $I_0(CHEM)$ the aggregate of all the isotopes (which naturally contains $I_0(CHEM)$ themselves) makes $I(CHEM)$), while the generating procedure π in this case is the operation to append an allowable number of element neutrons (0, 1, 2, ...) to the nucleus of the respective atom.

For the set of tokens of an inflectional language in a canonical (original) form as $I^{WORD}(L)$ can act as an class of text word forms and in this case π is interpreted as the operator of paradigmatization (the operator to construct a complete word-inflection paradigm), i.e. an algorithm that juxtaposes a complete word-inflection paradigm to each lexeme in its canonical form.

Processes similar to those described, occur in all the socio-technical systems complicated enough and, more generally, in systems of any origin, in which there are sources, converters and consumers of information, and therefore some analogues of perceptual-sensory acts and thought-speech processes take place.

2. The structure and architecture of lexicographic systems.

We determine the basic constructive component in the strucrueres of the above type as so-called lexicographic systems (the abbreviation of L-system will be used below as well). The notion of L-system is the basic concept of this work, and its definition is based on phenomenology of the described lexicographic effect.

Lexicographic systems correlate with widespread formalized structures of the same kind, such as data models, formal systems, canonical calculi in finite alphabets, etc.

Note that particular cases (or implementations) of lexicographic systems in science and technology operate very long time. It serves as a large variety of information systems, databases and knowledge, which includes all the traditional dictionaries and computer-vocabulary system.

In terms of machine dictionaries, they can effectively perform its functions only if they adequately reflect the structure of the form and content of language units that are subject of lexicographing. The trend towards re-establishment of the completeness of actually observed only when the design of linguistic systems is based on a thorough study of the language of phenomenology, which itself "tells" the choice of adequate staff, as well as the construction of appropriate models. Despite the fact that the goal of information science is the interpretation of the subject industry (in our case, the linguistic facts), the language of data models, are the types of construction of these models should be the subject of industry

and perhaps more precisely tailored to specific linguistic phenomena. On the basis of the state that we have constructed a theory of lexicographic systems, based on the phenomenology of lexicographic effect, the consistent application of which provided an opportunity for the necessary systemic generalizations and establishing a methodology for constructing lexicographic models. Historically the starting point of analysis, resulting in the formulation of the theory of lexicographic systems, the study proved a significant number of structures actually existing traditional dictionaries, their generalisation and construction of appropriate models. To get a more detailed picture we study the general structure-making effects and elements of lexicographical systems which aside from the traditional dictionaries, becoming elements of infological models for lexicographic systems of "general position". This way led to the establishment of the *structure* of lexicographic system.

It is obvious that the structure of traditional dictionaries is not accidental, since it focuses on experience of generations of lexicographers. Therefore, it is usually free of subjective tastes and preferences of the developers of information systems. Lexicography experience as a kind of intellectual activity (to the extent of its accumulation), from systematization of the actual facts of philology (and even from the systematization of data on the lexical units) was gradually spread to the systematization of the data about the world, knowledge of which, in turn, are focused on natural language as an integrated information system.

The universality of the phenomenon of lexicographic effect gives rise to the trend we have noted not once: to undergo any linguistic phenomenon to lexicographing. This fact can explain the existence in lexicographical practice of dictionaries lexicographing even such language units which have no direct verbal expression. Thus, an attempt of lexicographing of the syntactic structures has been made, for example, in a work by G.A. Zolotova. Its introduction states: "As the physical world around us is made up of elementary particles, the smallest known particles of matter, similarly the syntactic structure of our language is organized by varied, though regular combinations of basic, or minimum, units no more divisible on the syntactical level. In linguistics at the present stage of its development the need has matured to understand the concept of elementary syntactic units to be, as it is ever more obvious, a base for other more complex ones to be built on." And further on: "We use the term “syntaxeme” for a minimum, no more divisible semantic-syntax unit of the Russian language that as both a carrier of basic meaning, and as a constructive component of more complex syntactic constructions, and therefore is characterized by a set of syntactic functions." [12]

Note the clear analogy (somewhere with almost a text match) to our formulation of lexicographic effect, its space of action being obviously much wider.

Similar attempts of lexicographing of semantic structures not only reflect the general trend of lexicographical description of linguistic phenomena but also meet the practical needs for the development of more sophisticated systems of language support.

From the above we derive the methodological correctness of the inclusion of units at any language level into elementary information lexicographic units of a certain lexicographical system. Thus, semantic, syntactic, cognitive, and other structures which typically do not have a direct verbal enbodiment in natural language do undergo the lexicographing. The works of the type are close to the compiling of dictionaries: ideographic, those of verbal management, word equivalents, phraseological units, etc. The latter two types of dictionaries adjoin a number of potential lexicographic works, not yet created, but theoretically having every right to exist [13]. In the abovementioned work suggestions are given for creating more than 50 different dictionaries in which units for lexicographing (elementary information units in respect of the relevant, sometimes very exotic lexicographic effects) are, for example, appeals, etiquette phrases, honoratives (expression of politeness), humiliatives (expression of boorishness), incentives, and reactions (echoing, consent, objection, refutation), etc.

The study of various structures in existing traditional dictionaries allows us to make some generalizations that can not only form the basis of theoretical lexicographical scheme, but also be used in the design of specific information systems of linguistic kind, as well as when creating the respective software. As lexicography has long been delineated the concepts of "dictionary" and "list of words", "list",

"index", "inventory", the dictionary as an abstract system necessarily has the lexicographic structure containing at least two required parts: register (left) and interpretation (right), as a manifestation of the relationship between form and content (RFC). That is the availability of interpretation (a carrier of the semantic component RFC) differentiates a dictionary from a usual list of words. However, the dictionary has a deeper structure, which is reflected in the structure of the register and interpretation of the dictionary as a whole, the individual word parts, as well as in the structure of inter-entry and inter-word mappings. Due to this the dictionary is a special kind of text, which in a systematic and structured way describes the units and relationships of a particular language (or an aggregate of languages). It is natural to consider the dictionary as a specific object of technology, namely, an information system, which designates certain linguistic effects by using some printing display, namely bolding, positional placement, special symbols, etc., which play the role of identifiers of the relevant information variables – elements of a dictionary metalanguage. Besides, the complexity of the dictionary structure is in the fact that not all elements of its structure are manifested by the above method. The structure of a real dictionary, as a rule, has a large number of implicit structural elements and to identify them is very often rather a difficult task. The process of abstraction of the dictionary (lexicographical) structure is a kind of decoding, the reconstruction of the specific lexicographic effect, which has caused formation of the given structure, and developes using several provisions though set out first in linguistics, but being in fact of a system-wide significance.

Building of the structural model of lexicographic (dictionary) systems is focused on many aspects of the representation of the sign nature of lexical units as the most compact and most informative in natural language. From the standpoint of the theory of lexicographic effect, this means the extraction from the studied language system of a subsystem of elementary information units (EIU) and the identification of the set of system-structural parameters.

The next point is in taking into account the dichotomical structure of every EIU (and of their full aggregate), what is reflected in a multidimensional relationship between form and content, the ascertained EIU is a carrier of.

The multi-aspect representation of the sign nature of natural language units in the traditional dictionaries (or EIU in the general lexicographic systems) is provided for by accounting semiological, linguistic (phonetic, morphematic, grammatical, semantic, stylistic, etc.) and cognitive features of objects to undergo the lexicographing depending on the type of vocabulary and characterization depth of the lexicographic effect to be studied in each case. In the information-lexicographical model a certain number of data and/or knowledge sets match the specified features.

Note that in the language (speech) flow, the ontological nature of language is not divided into separate components, as is the case in conceptual interpretations. This fact gives rise to the desire of creating "integrated" dictionaries and, hence, the need for comprehensive (integrated) models of linguistic phenomena. Therefore, in designing computer systems for the language processing raises the challenge of creating formalized models that are configured for the effective presentation of the integration processes of language and at the same time take into account the specifics of linguistic objects. Thus, the criterion of a plurality of aspects in the representation of sign nature of language makes it possible to build comprehensive, integrated data models fit for the unification of conceptual representations linguistic phenomena different in its nature.

The dichotomic structure of the EIU in the information lexicographical model (as is the case in most traditional dictionaries) is manifested in the structural organization of lexicographical system and derives from the fundamentals of modern linguistics operating concepts of form and content, internal and external forms of linguistic units, their phenomenology being deeply studied on the linguistic material.

As V.M. Rusanivsky [14] has noted, language has a dualistic function: on the one hand this is the material ground on which the thinking is based in the process of its operation, and on the other, the material in which it is recorded to become an accomplished fact. The objects of study of the thought-speech flow constituents are both physical ("material") and notional ("ideal") sides. So, the sound substance of the speech can be regarded as its form and properties of information as its content. From this perspective, a sound implementation of the speech can be divided into elements agregatated at a different

degree: integral units of intonation (intonems), combinations of vowels and consonants (syllables), vowels and consonants themselves (sounds), etc. This process is infinite, because the selection and classification of sounds of speech depends on a variety of reasons, including the progress in acoustics, phonology, etc. The physical process of speech refers to the irreversible (like many other acoustic phenomena) dissipative processes. These properties of the physical substance of speech, together with the properties of the speech apparatus determine its external information characteristics.

In turn, the written form of language models its oral form. Therefore, in general, the sequence:

<model of reality $\rightarrow$ thinking
$\rightarrow$ pattern of thinking $\rightarrow$ oral language
$\rightarrow$ model of oral language $\rightarrow$ written language

is quite correct. As shown those models are physically implemented in a single system (related to individuals, social communities, systems of culture, etc.), their interaction and mutual influence are natural and necessary. Thus, the written version of language also serves as both a pattern of thinking and a model of reality.

The direct manifestation of language in speech activity, as well as the existence of writing and other ways of fixing the language acts on physical media, different from natural language expressions, is a property of language "to have an external form." The external form is possible due to the ability of language to be a "representative" part of the phenomenal side of reality, and, because the speech is a sort of reality, it has facilities for denoting itself.

The system which is a representative of the phenomenal side of reality should be organized in a special way. Since the difference between the phenomenon and the essence is relative, but there is no clear boundary between the phenomenal and substantive aspects, language as a model of reality should not have such a boundary either. This fact is realized in the property of word to have an internal form associated with the place of its noumenon part in the language system. The external and internal forms are therefore interrelated and together become the form of word, as opposed to its content as a sum of specific values.

All this gives rise to the claim that the RFC (including the notion of internal and external form of linguistic units) are general in nature and represent a universal property of EIU, induced in the development of a lexicographic effect. As formalized in the form of data models they are able to form a substrate of models for information systems of arbitrary nature and origin. For the language-oriented models, in general, the RFC is necessary. The mentioned concepts, in our opinion, have a potential of constructivity, because the content exists only in a certain formal shell that allows us to apply a uniform approach to the construction of their representatives in the scientific theory.

Consider a fragment of reality D and present its conceptual description in the form of a specific lexicographical system. Since we are interested first of all in linguistic facts that we consider a natural language, or a set of natural languages, or a subsystem of (certain aspects of) natural language to represent D here.

According to the above, a certain hierarchy of lexicographic effects is inherent in the system D. Thus, for the system of natural language we can give a number of lexicographic effects which result in the selection of individual phonemes, syllables, suggestions, etc from the flow of speech. All of the units serve as components of EIU relative to certain types of lexicographic effects in natural language.

Later on we regard the lexicographic system (L-system) as a special information environment in which a certain lexicographic effect (or a combination of lexicographic effects) is developed.

To construct a practically useful scheme for modelling the abovementioned phenomena it is necessary to determine a set of information constructives that specify the structural elements of L-systems to allow you to develop specific applications. In turn, this requires building a constructive theory of L-systems. It is based on the lexicographic model, developed in the works [15], and their conventional symbols and results are used below.

In accordance with the information interpretation of perception [16] we determine the result of the reception by a subject S of a class of elementary information units (EIU) $I^Q(D)$ in the form of a certain set of $V(I^Q(D))$ – set of descriptions of units belonging to the class of $I^Q(D)$; this set is the result of process:

$$S : I^Q(D) \rightarrow V(I^Q(D)), \tag{1.12}$$

That is why for each element $x \in I^Q(D)$ its description of $V(x)$ as an element of the set $V(I^Q(D))$: $V(x) \in V(I^Q(D))$; $Sx = V(x)$ is uniquely defined. Therefore, it is logical to assume that for each element $V(I^Q(D))$ has the form of aggregation:

$$V(I^Q(D)) = \cup_{x \in I^Q(D)} V(x). \quad (1.13)$$

In accordance with the information concept RFC, each $V(x)$ is represented in the form of a word (a text) in some finite alphabet $\boldsymbol{A} = \{a_1, a_2, ..., a_n\}$, i.e. finite sequence of symbols from $\boldsymbol{A}$. Further on, words in the alphabet $\boldsymbol{A}$ are called $\boldsymbol{A}$-words. For example, if we consider the Dictionary of the Ukrainian language, the alphabet $\boldsymbol{A}$ consists of the following elements:

standard Ukrainian alphabet (big and small letters), punctuation marks, Arabic numerals, Roman numerals, spaces and paragraph symbols, special characters (/ /, Δ, ▲, ◊, ♦ ,...); font types, etc. .

Description of any EIU in this way is presented in $\boldsymbol{A}$-word of the following form:

$$V(x) = v_1(x)v_2(x)... v_{k(x)}(x), v_i(x) \in A, i = 1, 2, ..., k(x), k(x) \geq 1. \quad (1.14),$$

where each "letter" $v_i(x)$ ($\boldsymbol{A}$-letter) is taken from the alphabet A. Note that the length of $k(x)$ $\boldsymbol{A}$-word $V(x)$ depends on x. Formula (1.14), by definition, provides a complete, in a sense, exhaustive description of the elementary information unit x in the lexicographic system. Using the mapping S between the class of the EIU ($I^Q(D)$ and numerous descriptions of $V(I^Q(D))$ establishes a certain isomorphism. In other words, a set of descriptions of $V(I^Q(D)$ is an own subset of $W(\boldsymbol{A})$: $V(I^Q(D)) \subset W(\boldsymbol{A})$, and the set $W(\boldsymbol{A})$ is a set of all words of finite length from $\boldsymbol{A}$, i.e. sequences of $v_1v_2...v_q$, $q<\infty$, $v_i \in \boldsymbol{A}$, $i = 1, 2, ..., q$. We believe that the word of zero length – 0 also belongs to $W(\boldsymbol{A})$: $\forall\, a \in W(\boldsymbol{A})\ \exists\, 0 \in W(\boldsymbol{A})$, such that $a * 0 = 0 * a = a$, where «*» is a concatenation. The closure relative to the operation of concatenation, i.e. the requirement: $a, b \in W(\boldsymbol{A}) \Rightarrow \exists\, c \in W(\boldsymbol{A})$, $c = a * b$, as well as the associativity about it: $\forall\, a, b, c \in W(\boldsymbol{A}) \Rightarrow a * b * c = (a * b) * c = a * (b * c)*$ c) makes $W(\boldsymbol{A})$ a semigroup with the semigroup operation «*» and the unit element 0.

The choice of the alphabet A, which is realized by $W(\boldsymbol{A})$ and $V(I^Q(D))$, is not justified and specified here what corresponds to the algebraic tradition. However, note that its generation is a consequence of a certain lexicographic effect developing in the system of speech (acoustic) and its information and graphical interpretation. If we consider conventional dictionaries, the interpretation of the $\boldsymbol{A}$-word $V(x)$ as a word-entry text with the register unit x is natural.

In general the semigroup structure is poor enough, and the construction of $W(\boldsymbol{A})$ is too large to effectively identify in it characteristic features of language systems. To achieve this goal, it is necessary to introduce some additional assumptions and constraints, that are instrumental to single out substructures typical exactly for natural language in the structure of $W(\boldsymbol{A})$. This is achieved as follows.

Since each $V(x)$ is an adequate and unambiguous description of the corresponding element x of the system $I^Q(D)$, its structure with sufficient fullness must reflect the properties of that element. Given the linear character of $V(x)$ as a linear sequence of symbols from $\boldsymbol{A}$, we come to the conclusion that the only possible natural source of its structure is a certain set of his $\boldsymbol{A}$-subwords and certain relations between them. $\boldsymbol{A}$-subwords in the description of $V(x)$ is defined as $\boldsymbol{A}$-words, consisting of those symbols of alphabet $\boldsymbol{A}$ contained in the description of $V(x)$ and located in the $\boldsymbol{A}$-subwords in the order induced by the location of letters in the description. Obviously, the set of all $\boldsymbol{A}$-subwords of the $\boldsymbol{A}$-word of length n (i.e. the $\boldsymbol{A}$-words which consist of n $\boldsymbol{A}$-letters) contains 2^n elements. We denote the set of all $\boldsymbol{A}$-subwords of $\boldsymbol{A}$-word $V(x)$ by $B[V(x)]$.

The structure on the set of descriptions is introduced as follows. Assume that all descriptions of $V(x)$ there is one rule by which any of the $\boldsymbol{A}$-word $V(x)$ can be singled out a set of $\boldsymbol{A}$-subwords $\beta(x) = \{\beta_i(x)\}$ with the following properties:

– the element x belongs to the set $\beta(x)$: $x \in \beta(x)$;

– the whole description $V(x)$ is an element of the set $\beta(x)$: $V(x) \in \beta(x)$;

– the rule that singles out the elements of the set $\beta(x)$ is the same for all $V(x)$.

Thus, from $V(x)$ elements of the set of the $\beta[V(x)]$ of the values ($\boldsymbol{A}$-subwords) $\beta_i(x)$ of the following form can be marked out:

$$\beta[V(x)] \equiv \{\beta_i(x),\ i = 1, 2, ..., q\} \subseteq B[V(x)], \tag{1.15}$$

где $B[V(x)] = \{v_{i_1} v_{i_2} ... v_{i_p}\ ,\ 1 \le\ i_1 < i_2 < ... < i_p \le\ k(x),\ \ p = 1, 2, ..., k(x)\ ,$ and:

$$v_{ij} \in \{v_{1(x),}\, v_{2(x),\, ...,}\, v_{k(x)}(x)\};\ x \in \beta[V(x)];\ V(x) \in \beta[V(x)],\ \beta_{k(x)} \neq \beta_{m(x)}\ when\ k \neq m. \tag{1.16}$$

Suppose, by definition:

$$\beta[V(I^Q(D))] = \cup\ \beta\ [V(x)]. \tag{1.17}$$
$$x \in I^{Q(D)}$$

Obviously: $V(I^Q(D)) \in \beta[V(I^Q(D))]$. Designate:

$$\beta_i = \cup\ \beta_i(x),\ i = 1, 2, ..., q, \text{ and also } \beta = \cup_{i}^{q}\ \beta_i\ . \tag{1.18}$$
$$x \in I^{Q(D)}$$

Undoubtedly, $\beta \equiv \beta\ [V(I^Q(D))]$. Note that some of the elements $\beta_i(x)$, $i = 1, 2, ..., q$ may be empty under certain values of $x \in I^Q(D)$; In this case they are omitted in the formulas (1.15) - (1.18).

By $\boldsymbol{\sigma}[\beta]$ denote some kind of structure defined on β, and therefore on $V(I^Q(D))$. Futher on we call $\boldsymbol{\sigma}[\beta]$ a macrostructure $V(I^Q(D))$; restrictions $\boldsymbol{\sigma}[\beta]$ on $V(x)$: $\boldsymbol{\sigma}[\beta]\,\big|_{V(x)} \equiv \boldsymbol{\sigma}(x)$ generates the microstructure of $V(x)$. A strong formulation of this fact is to establish procedures (operator, process ...) $\boldsymbol{\sigma}$ that generates the structure $\boldsymbol{\sigma}[\beta]$ on β:

$$\boldsymbol{\sigma}: \beta \rightarrow \boldsymbol{\sigma}[\beta]. \tag{1.19}$$

A range of a number of non-isomorphic structures $\boldsymbol{\sigma}[\beta]$ can be generated on β. Those structures can be represented by any of the known data models (hierarchical, network, relational, object-relational, etc.), logical-mathematical models (in particular, logical calculi like the predicate logic), expressions of formal grammars, etc.

The following method can be one of the possible procedures to form the structure. Let us build a table:

β_1	β_2	...	β_q
$\beta_1(x_1)$	$\beta_2(x_1)$	...	$\beta_q(x_1)$
$\beta_1(x_2)$	$\beta_2(x_2)$	...	$\beta_q(x_2)$
.	.		.
.	.		.
.	.		.
$\beta_1(x_M)$	$\beta_2(x_M)$	...	$\beta_q(x_M)$

Some of the elements, obviously, can be empty, therefore the length of the chart columns, generally speaking, can be different. Values β_i, i =1, 2, ..., q, are interpreted as attributes (attribute names) and sets $\text{Dom}\ \beta_i \equiv \{\beta_i(x_1), \beta_i(x_2), \ldots, \beta_i(x_M)\}$, i =1, 2, ..., q, as the domains of these attributes. Then the structure $\boldsymbol{\sigma}[\beta]$ can be realized in the form of a relational algebra R, defined on the Cartesian product:

$$\mathbf{X}_{i=1}^{q}\ \text{Dom}\beta_i = \beta^{\otimes}\ . \tag{1.20}$$

In other words, if the structure $\boldsymbol{\sigma}$ is identified with a certain relational algebra R over $\beta^{\otimes}$, then the triplet $\{V(I^Q(D)),\ \beta,\ R\ [\beta]\}$ is nothing different from a relational model, and the quintuple $\{I^Q(D),\ D,\ V(I^Q(D)),\ \beta,\ R\ [\beta]\}$ specifies some object-relational model [17]. Besides, $I^Q(D)$ is a class of objects of the model, β_i are interpreted as attributes (attribute names), with the domains $\text{Dom}\ \beta_i$, $I^Q(D)$ with elements $\beta_i(x)$, $x \in I^Q(D)$. It is clear that the set $\{x\}$ makes individual domains by itself (to abridge the description, without further detail, we identify an element x as belonging to the class $I^Q(D)$, with its "name" in the

$V(x)$) and $V(I^{Q}(D))$ – as the set $\{V(x)\}$. The relationship of the respective arity are defined, as always, in the form of certain subsets of the set $\beta^{\otimes}$. Their tuples are elements of the form:

$$(\beta_{i1}(x_{ji1}), \beta_{i2}(x_{ji2}), ..., \beta_{ir}(x_{jir})), \; i1 < i2 < ... < ir \; ; \; x_{jim} \in I^{Q}(D), \; m = 1, 2, \ldots, r . \qquad (1.21)$$

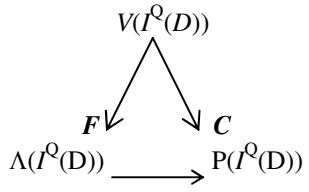
The relational calculus in this model is determined as usual (see, for example [18]).

Due to the interpretation of $I^{Q(}D)$ as classes with objects of any origin as their elements, an object-oriented interpretation of the model looks quite natural. Th relationships between the elements of the class $I^{Q(}D)$ are induced by a system by unary relations $r\,[\beta_1]$ on $\beta_1 = \{x\}$ and the mapping

$$\Delta: V(I^{Q}(D)) \rightarrow I^{Q}(D); \qquad \Delta\, r\,[\beta_1]. \qquad (1.22).$$

Thus, the set $I^{Q(}D)$ represents the ontological essence of reality to be modelled, while $V(I^{Q(}D))$, β, $\boldsymbol{\sigma}[\beta]$ presents its conceptual side.

Further on we focus on the RFC that develop and implement in the medium $\{I^{Q(}D)$, S, $V(I^{Q(}D))$, β, $\boldsymbol{\sigma}[\beta]\}$. This combination of properties of the complex $I^{Q(}D)$ is divided into two parts not well-defined and harldy separable. Note that the conceptual scheme, implemented in the description of $V(I^{Q}(D))$, these parts must be separated. In other words, a necessary condition for the construction to be correct is the existence of a procedure making such a separation. This is shown in the commutative diagram:



(1.23)

$\boldsymbol{F}V(I^{Q}(D)) = \Lambda(I^{Q}(D))$; $\boldsymbol{C}V(I^{Q}(D)) = P(I^{Q}(D))$; $\Lambda(I^{Q}(D)) \cap P(I^{Q}(D)) = \varnothing$, and

$\boldsymbol{H} \circ \boldsymbol{F} = \boldsymbol{C}$, where the symbol «∘» marked composition of mappings.

$$\Lambda(I^{Q}(D)) = \bigcup_{x \in I^{Q}(D)} \Lambda(x); \qquad P(I^{Q}(D)) = \bigcup_{x \in I^{Q}(D)} P(x). \qquad (1.24)$$

On $\Lambda(I^{Q}(D))$ и $P(I^{Q}(D))$ macrostructures are induced:

$$\boldsymbol{F}\boldsymbol{\sigma}[\beta] = \lambda[\beta] \text{ i } \boldsymbol{C}\boldsymbol{\sigma}[\beta]) = \rho[\beta] \qquad , \qquad (1.25)$$

and the corresponding macrostructures:

$$\lambda[\beta] \,\big|_{V(x)} \equiv \lambda(x); \; \rho[\beta] \,\big|_{V(x)} \equiv \rho(x) \qquad (1.26)$$

as a restriction $\lambda[\beta]$ and $\rho[\beta]$ on $V(x)$.

Note that the diagram (1.23) (i.e. the objects $V(I^{Q}(D))$, $\Lambda(I^{Q}(D))$, $P(I^{Q}(D)$ and the mappings $\boldsymbol{F}$, $\boldsymbol{C}$, $\boldsymbol{H}$) are build independently of the structure $\boldsymbol{\sigma}[\beta]$. The origin and content of its constituent elements is quite different. Namely: $\Lambda(I^{Q}(D))$ corresponds to that part of the description $V(I^{Q}(D))$, which in a sense, represents the form of $I^{Q}(D)$, while $P(I^{Q}(D)$, respectively, corresponds to that part of the description of $V(I^{Q}(D))$ responsible for the content of $I^{Q}(D)$. The above specification confirms the idea that the RFC are universal, inherent in objects of any origin.

Definition 1. The octad of objects $\{I^{Q}(D)$, S, $V(I^{Q}(D))$, β, $\boldsymbol{\sigma}[\beta]$, $\boldsymbol{F}$, $\boldsymbol{C}$, $\boldsymbol{H}\}$ designate the basic lexicographic data model and its concrete realization – the elementary lexicographic. Sometimes, for abridgement, if no discrepancy emerges, we denoted by $V(I^{Q}(D))$ all the elementary lexicographical system.

Note that any element (or any of their aggregates) belonging to the structures β, $\boldsymbol{\sigma}[\beta]$, $\lambda[\beta]$, $\rho[\beta]$, can be interpreted as a elementary L-system. Hence we get an opportunity to single out in the basic structure of the original elementary L-system a number of information-linguistic substructures we interpret as separate L-systems. Thus redefining the structure of the original L-system, we obtain the general L-model and L-system (not elementary ones). It appears in the form of a certain number of elementary L-systems with possible mappings and links among them. Thus, the general position is of the form of a graph $G = \{V = \{V_i\}; R = \{R_{kl}\}\}$, where $V = \{V_i\}$–G is a set of vertices made by L-systems constituent of V_i, members of the G, and $R = \{R_{kl}\}$ – the set of edges of the graph G, R_{kl} combines V_k and V_l.

In particular, nothing prevents us from regarding $\Lambda(I^{Q}(D))$ и $P(I^{Q}(D))$ as separate, autonomous elementary L-systems to make it possible to build the following:

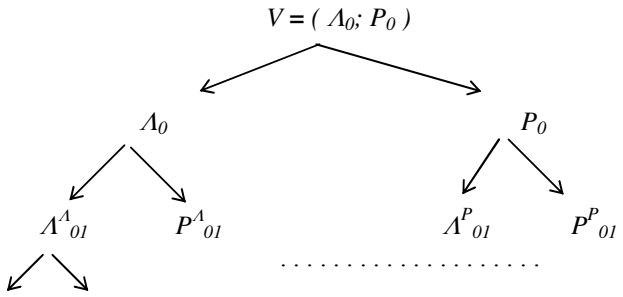
$$V(I^{Q}(D)) = (\Lambda(I^{Q}(D)) \equiv \Lambda_0(I^{Q}(D)) \xrightarrow{H_0} P_0(I^{Q}(D)) \equiv P(I^{Q}(D))$$

$$\mathrm{F}^{\Lambda}_{01} \quad \mathrm{C}^{\Lambda}_{01} \qquad\qquad \mathrm{F}^{\mathrm{P}}_{01} \quad \mathrm{C}^{\mathrm{P}}_{01} \tag{1.27}$$

$$\Lambda^{\Lambda}_{01}(I^{Q1}(D)) \rightleftarrows P^{\Lambda}_{01}(I^{Q1}(D)) \qquad \Lambda^{P}_{01}(I^{Q2}(D)) \rightleftarrows P^{P}_{01}(I^{Q2}(D))$$

$$H^{\Lambda}_{01} \qquad\qquad H^{P}_{01}$$

Note the type of the lexicographic effect on the second floor has changed – instead of Q now we have Q_1 and Q_2 respectively. So, we come to a set of objects $I^{Q1}(D)$ and $I^{Q2}(D)$. Continuing this process we obtain the recursive development of the lexicographical system $V(I^{Q}(D))$:



(1.28)

We call this process *the recursive reduction of the lexicographical system*. It recalls a kind of information "microscope" revealing ever more subtle details of lexicographical system.

Further on, we will denote the recursive reduction process of L-system $V(I^{Q}(D))$ by $RR{\downarrow}[V(I^{Q}(D))]$. The definition of this process includes the characterization of all the operators ***F, C, H***, at all the levels of the recursive reduction, together with the results of their actions, as well as all the macro- and microstructure σ, λ, ρ.

The described construction makes the content of a general definition of lexicographical data model:

$$\{I^{Q}(D), S, V(I^{Q}(D)), \beta, \boldsymbol{\sigma}\,[\beta], RR{\downarrow}[V(I^{Q}(D))]\} \tag{1.29}$$

and of the lexicographical system:

$$\{I^{Q}(D), S, V(I^{Q}(D)), \beta, \sigma\,[\beta], RR{\downarrow}[V(I^{Q}(D))], \boldsymbol{\Sigma}\}, \tag{1.30}$$

where the symbol Σ designates its architecture as an information model.

The three-level architecture $\boldsymbol{\Sigma}$ is usually chosen to conform to the standard architecture of information systems, introduced back in 1975 and named ANSI/X3/SPARK or just ANSI/SPARK [19]. We use the main components of the architecture ANSI/SPARK in the following interpretation:

$$ARCH\_LS = \{CM, EXM, INM; \Phi, \Psi, \Xi\}, \tag{1.31}$$

where the symbol CM designates the conceptual model of the lexicographical system LS. The symbol $EXM = \{\mathrm{ex}M\}$ identifies a set of its external models conforming to the conceptual model of the CM, and $INM = \{\mathrm{in}M\}$ – the corresponding set of its internal models. By CM we denote the set of mappings of *CM* into *EXM*:

$$\varphi : CM \to \mathrm{ex}M, \text{ where } \mathrm{ex}M \in EXM; \tag{1.32}$$

respectively, $\Psi = \{\psi\}$ – set of mappings of the *CM* into *INM:*

$$\psi : CM \to \mathrm{in}M, \text{ where } \mathrm{in}M \in \mathrm{INM}; \tag{1.33}$$

$\Xi = \{\xi\}$ – the set of mappings of INM into EXM:

$$\xi\,(\mathrm{in}M) = \mathrm{ex}M. \tag{1.34}$$

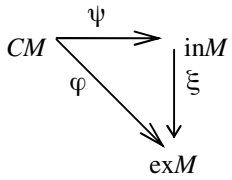
Next we dwell on the interpretation of architecture elements.

A conceptual model (conceptual level of presentation) of the subject area is semantic, sign model integrating notions of different experts in the subject field in an unambiguous, finite and consistent form.

The internal model (internal level of presentations) defines types, structures and formats of data presentation, preservation and manipulation, an algorithmic base and an operating software environment the model is immersed in when being implemented.

The external model (external level of presentation) reflects the views of end users and, hence, application programmers, to the information system. It means a system of tools is implemented, to enable the user to make the permitted contacts and manipulate the data provided on the internal level.

Mappings are constructed in such a way that the diagram:



(1.35)

is commutative: $\xi \circ \psi = \varphi$. The requirement of commutativity of the diagram is essential since it ensures a consistency among all the levels of the system architecture.

References

[1] В.А.Широков. Феноменологія лексикографічних систем., – К.: Довіра, 2004, с. 327. Розділ 1.

[2] Ibid.

[3] Шенк Р., Бирнбаум Л., Мей Дж. К интеграции семантики и прагматики // Новое в зарубежной лингвистике : Вып 24. Компьютерная лингвистика.—М., 1988.— С. 33.

[4] http://cogsci.princeton.edu/~wn/.

[5] Дьячков М. В. Миноритарные языки в полиэтнических (многонациональных) государствах.— М., 1996.— 116 с.

[6] Чейф У. Значение и структура языка.— М., 1975.— 432 с.

[7] Щерба Л. В. Языковая система и речевая деятельность.— Л., 1974.— 428 с.

[8] Different aspects of the correlation «form-content» have been actively explored in linguistics from W.von Humboldt and then F. de Saussure; and we have also in the previous section given an account of the established linguistic views at this correlation.

[9] Post-modernism flatly interprets this phenomenon as a creative act, namely a process of creating an object by the subject. Note the phenomenon has been known in quantum physics within more than 70 years (the device's impact on an observed object can make the latter's behaviour utterly uncontrolled; at due time this circumstance gave cause to adherents of a vulgar trend in materialistic philosophy for accusing physicists of subjective idealism). It may have been the first ever precedent of considering the "substance" of the subjective in a positive science.

[10] А.Н.Колмогоров. Три подхода к определению понятия количества информации. В кн. Теория информации и теория алгоритмов. – М.: Наука, 1987. – С.220.

[11] А.Н.Колмогоров. Три подхода к определению понятия количества информации. В кн. Теория информации и теория алгоритмов. – М.: Наука, 1987. – С.220.

[12] Золотова Г. А. Указ. труд.— С. 3–4.

[13] Девкин В. Д. О неродившихся немецких и русских словарях // Вопр. языкознания.— 2001.— № 1.— С. 85–97.

[14] Русанівський В. М. Структура лексичної та граматичної семантики.— К., 1988.— 240 с.

[15] Широков В. А. Інформаційна теорія лексикографічних систем.; Широков В. А., Рабулець О. Г. Формалізація у галузі лінгвістики // Актуальні проблеми української лінгвістики: теорія і практика.—К., 2002.— Вип. 5.— С. 3–27.

[16] A.V.Palagin, V.A.Shirokov. Principles of cognitive lexicography // International journal «Informational theories & application».— 2000.— Vol. 9, № 2, pp. 43–51.

[17] Коннолли Томас, Бегг Каролин, Страчан Анна. Базы данных: проектирование, реализация и сопровождение. Теория и практика. – 2-е изд. :Пер. с англ. – М.: Издательский дом "Вильямс", 2000. - 1120 с. : ил.

[18] Ульман Дж. Основы систем Баз данных. М.: Финансы и статистика, 1983. – 334 с.

[19] ANSI/X3/SPARK Study group on data base management systems: interim report, FDT, 7:2, ACM.— New York, 1975.

A Knowledge-rich Lexicon for Bulgarian*

Kiril Simov

Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria
kivs@bultreebank.org

Abstract. In contrast to the morphological and syntactic processing, the semantic annotation, based on ontology, is still underdeveloped for Bulgarian. On the other hand, the prerequisites for an ontological annotation are already available. These are as follows: a morpho-syntactic tagger for Bulgarian with more than 95% accuracy; a dependency parser with more than 84% accuracy; a general chunker and a named entity grammar. Therefore, the next logical step is the semantic annotation. As a minimal set of semantic resources we consider the following ones:

- a lexicon for Bulgarian aligned to an upper ontology as a mechanism to cover the common lexica in domain texts, and aligned to domain ontologies to cover domain terminology;
- a corpus, annotated with ontology information in order to train machine learning component for automatic word sense disambiguation;
- an annotation grammar for Bulgarian, based on syntactic knowledge of Bulgarian and conceptual information from the ontology.

In this paper, we will focus on the description of the lexicon.

1 Introduction

Semantic Annotation (Tagging) is a natural further development in the area of language resources after the creation of morphologically and syntactically annotated corpora. The importance of Semantic Annotation became a hot topic within the initiative for creation of Semantic Web. Although much work is already done in the area, the term "semantic annotation" is not yet well defined – see [8] and citation therein. In our work we consider the text as consisting of two types of information: (1) ontological classes and relations, and (2) world facts. The ontological part determines generally the topic and the domain of the text. We call the corresponding "minimal" part of ontology implied by the text ontology of the text. The world facts represent an instantiation of the ontology in the text (here higher order entities like beliefs, claims, etc. are also included). Both types of information are called uniformly 'semantic content of the text'. Both components of the semantic content are connected to the syntactic structure of the text. Any (partial) explication of the semantic content of a text will be called semantic annotation of the text. Defined in this way, the semantic annotation could contain also some pragmatic information and actual world knowledge.

In order to support this kind of semantic annotation we rely on a knowledge-rich lexicon to determine the content of the semantic annotation. The lexicon is aligned to an upper ontology which covers the general meanings of the lexical items. In addition to the upper ontology the lexicon might be aligned to domain ontologies in order to support more precise domain annotation. In the paper a special focus is put on the role of the regular polysemy and metonymy. They are encoded as special patterns extracted from a semantically annotated corpus and reflecting the conceptual structure of the ontology. The lexicon is also connected to an annotation grammar which establishes a relation between the ontology and the text. In this paper we will not discuss the grammar and the annotation process.

The structure of the paper is as follows: the next section discusses the structure of a domain ontology, its connection to an upper ontology; the third section provides a model of ontology-to-text relation which is a motivation for the creation of the a knowledge-rich lexicon of Bulgarian; the next section discusses the extensions of the ontology-to-text relation with respect to general lexica and coverage of some phenomena; the fifth section compares our work with some other works; and the last section concludes the paper.

* This work is partially supported by LTfLL project (Language Technology for Lifelong Learning – IST-212578).

2 The Structure of the Domain Ontology

Independently from the methodology for ontology creation, the end result has the following structure:

- Domain layer. At this layer we have the real domain concepts and relations representing the main notions in the domain. These concepts and relations are used in solving different tasks such as representation of domain knowledge, representation of common conceptualization for information exchange in the domain, semantic annotation of domain texts, etc.
- Upper layer. The alignment of the domain layer to an upper ontology is an obligatory step in each ontology creation methodology. This alignment ensures several properties of the domain ontology: (1) consistency with the design of the upper ontology; (2) inheritance of the knowledge represented in the upper ontology.
- Middle layer. This layer contains concepts and relations which are not part of the upper or the domain layers, but play important role for the alignment between them.
- Language layer. It is supposed that the domain ontology (together with middle and upper layers) is language independent, formalized in some ontology representation language. In practise such an ontology needs has to be aligned to some language resources. This is necessary in order the ontology to be presented to users who do not know much of ontology and to support analysis of texts. Aa a minimum it is necessary to have a lexicon aligned to the concepts and the relations in the ontology.

We have used this structure of the ontology in three European projects – LT4eL, AsIsKnown and LTfLL. In each of them we have used as an upper ontology DOLCE Ontology [10] for several reasons: (1) it is constructed on rigorous basis which reflects the OntoClean methodology [6]; (2) it is represented in OWL-DL; (3) the authors of the ontology provide us comments and help on the alignment of the domain ontology to DOLCE. For the middle layer we have used OntoWordNet [4] – a version of WordNet aligned to DOLCE. OntoWordNet facilitates the alignment between the upper ontology and domain layer. This is ensured by providing more understandable concepts (more specific and closer to the domain) and the mapping between the concepts is easier. In the middle layer we include from OntoWordNet only those concepts that are necessary to support the alignment between the domain layer and the upper layer. The domain layer is created for each domain. The result of three layers is a domain ontology with a better structuring of the concepts and relations. Also relations and axioms are inherited from DOLCE to the domain layer.

Language layers in each of the projects were created on the basis of the model of the ontology-to-text relation presented in the next section.

3 Ontology-to-Text Model

In this section we represent the two main components that define the ontology-to-text relation. These components are: lexicon and concept annotation grammar.

The lexicon plays twofold role in our architecture. First, it interrelates the concepts in the ontology to the lexical knowledge used by the grammar in order to recognize the role of the concepts in the text. Second, the lexicon represents the main interface between the user and the ontology. This interface allows for the ontology to be navigated or represented in a natural for the user way. For example, the concepts and relations might be named with terms used by the users in their everyday activities and in their own natural language (e.g. Bulgarian). This could be considered as a first step to a contextualized usage of the ontology in a sense that the ontology could be viewed through different terms depending on the context. For example, the color names will vary from very specific terms within the domain of carpet production to more common names used when the same carpet is part of an interior design. Thus, the lexical items contain the following information: a term, contextual information determining the context of the term usage, grammatical features determining the syntactic realization within the text. In the current implementation of the lexicons the contextual information is simplified to a list of a few types of users (producer, retailer, etc). With respect to the relations between the terms in the lexicon and the concepts in the ontology, there are two main problems: (1) there is no lexicalized term for some of the concepts in the ontology, and (2) there are lexical terms in the language of the domain which lack corresponding concepts in the ontology, which represent the meaning

of the terms. The first problem is overcome by writing down in the lexicon also non-lexicalized (fully compositional) phrases to be represented. These different phrases or terms for a given concept are used as a basis for construction of the annotation grammar. Having them, we might capture different wordings of the same meaning in the text. The picture below shows the mapping varieties. It depicts the realization of the concepts (similarly for relations and instances) in the language. The concepts are language independent and they might be represented within a natural language as form(s) of a lexicalized term, or as a free phrase. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language[1]. Some of the free phrases receive their meaning compositionally regardless of their usage in the text, other free phrases denote the corresponding concept only in a particular context. In our lexicons we decided to register as many free phrases as possible in order to have better recall on the semantic annotation task. In case of a concept that is not-lexicalized in a given language we require at least one free phrase to be provided for this concept. The following picture shows the mapping from the ontology to the lexicon:

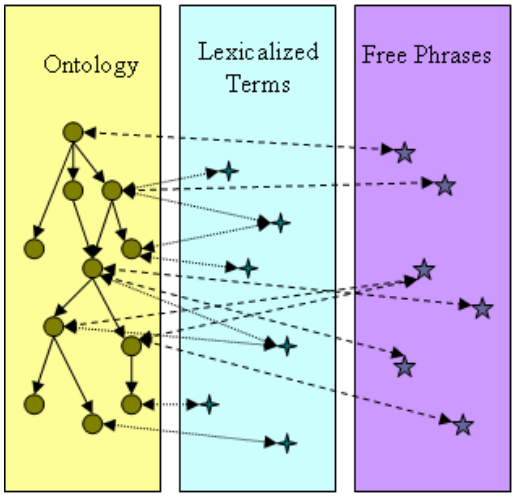


Fig 1. Ontology-to-Lexicon Relation.

The picture depicts the realization of the ontological concepts in a natural language. The concepts are language independent and they might be represented within a natural language as form(s) of a lexicalized term (or item), or as a free phrase. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language. Some of the free phrases receive their meaning compositionally regardless their usage in the text, other free phrases denote the corresponding concept only in a particular context.

In order to solve the second problem (missing concept for a lexical item) we modify the ontology in such a way that it contains all the important concepts for the domain. However, this solution requires a special treatment of the "head words" in the lexicons, because such phrases allow bigger freedom with respect to their occurrences in the text. Variability is a problem even with respect to the lexicalized cases and the idea is to represent the most frequent (based on a corpus) variants for each concept. The specific solutions for the lexical terms without appropriate concept in the ontology are the following:

More detailed classes in the ontology. In cases where it is possible, we are creating more specific concepts in the ontology. For example, the concept of 'shortcut' in the domain of Computer Science for End Users, is denoted by different lexical items in English depending on the operating system, because each operating system (MS Windows, Linux, etc) as a rule introduces its own terminology. When the notion is borrowed in other languages, it could be borrowed with different granularity, thus, we introduce more specific concepts in the ontology in order to ensure correct mapping between languages.

[1] The presence of free phrases in the lexicon is also motivated by the fact that the lexicalization is not a discrete feature. There are many different degrees of lexicalization. Thus the free phrases are the extreme end of the scale.

More complex mapping exists between the ontology and terms in some language. Our initial idea was that each meaning of a lexical item in any language is mapped to exactly one concept in the ontology. If for some lexical item this one-to-one mapping is not appropriate or it requires very complicated changes in the ontology, we realize a mapping based on ontology expressions instead of a single concept. This mechanism allows us to keep the ontology simpler and more understandable, and to handle cases that do not allow appropriate mappings. Currently, such cases are not detected in domains for which we applied this model.

We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in form of concepts and relations with appropriate axioms; and the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Of course, the ways in which a concept could be represented in the text are potentially infinite in number, thus, we could hope to represent in our lexicons only the most frequent and important terms and phrases. Here is an example of an entry:

```
<entry id="entry-34">
  <owl:Class rdf:about="http://www.asisknown.org/AIKHT#CarpetOWN">
    <rdfs:comment>a piece of thick heavy fabric (usually with nap or pile)
    used to cover a floor</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Class rdf:about=http://www.asisknown.org/AIKHT#FloorCovering/>
    </rdfs:subClassOf>
  </owl:Class>
  <def>a piece of thick heavy fabric (usually with nap or pile)
  used to cover a floor</def>
  <termg lang="en">
    <term shead="1">carpet</term>
    <term>carpeting</term>
    <term>rug</term>
    <term type="nonlex">textile floor covering</term>
    <def>a piece of thick heavy fabric (usually with nap or pile)
    used to cover a floor</def>
    <gramline>reference to finite state grammar</gramline>
  </termg>
</entry>
```

Each entry of the lexicons contains the following types of information: (1) information about the concept from the ontology which represents the meaning for the terms in the entry; (2) explanation of the concept meaning in English; (3) a set of terms in a given language that have the meaning expressed by the concept; and (4) relation to grammar rules. The concept part of the entry provides minimum information for formal definition of the concept. The English explanation of the concept meaning facilitates the human understanding. The set of terms stands for different wordings of the concept in the corresponding language. One of the terms is the representative for the term set. Note that this is a somewhat arbitrary decision, which might depend on frequency of term usage or specialist's intuition. This representative term will be used where just one of terms from the set is necessary to be used, for example as an item of a menu. In the example above we present the set of English terms for the concept 'carpet'. One of the terms is non-lexicalized – attribute `type` with value `"nonlex"`. The first term is representative for the term set and it is marked-up with attribute `shead` with value `"1"`. The elements `gramline` provide links to linguistic features of the terms like lemmatized variants of the terms, implementation as regular expressions to be compiled as finite state automata, etc.

Here we present a (part of) DTD for the lexicon:

```
<!ELEMENT OntoLexicon (entry+)>

<!ELEMENT entry
      ((owl:Class|rdf:Description|rdf:Property), def, termg+)>
```

```
<!ELEMENT def (#PCDATA)>

<!ELEMENT termg (term+,def?,gramline*)>
<!ATTLIST termg
      lang  (bg|cs|de|en|fr|hu|it|mt|nl|pl|pt|ro|ru) # REQUIRED
>

<!ELEMENT term (\#PCDATA)>
<!ATTLIST term
      type    (lex|nonlex)        "lex"
      shead   (1|0)               "0"
      gram    CDATA               #IMPLIED
>

<!ELEMENT gramline (#PCDATA)>
```

The lexicon consists of entries. Each entry consists of a concept, relation or instance (partial) definition, followed by a definition of the concept content in English and one or several term groups. Each term group represents all the available lexical terms or free phrases for the corresponding concept (relation or instance) in a given natural language (determined by the attribute lang). Optionally, the term group for a given language could contain a definition of the content of the concept in that language. Each term represents a normalized form of the term. Additionally, we could state whether: the term is a lexicalization of the concept in the language or it is a free phrase (attribute type); the term is representative for the concept in the language (the attribute shead) or not; and which grammar rules recognize this term (related to the concept (relation or instance) of the entry) in text. The format of the currently implemented grammars is given below.

The second component of the ontology-to-text relation, the concept annotation grammar, is ideally considered as an extension of a general language deep grammar which is adopted to the concept annotation task. Minimally, the concept annotation grammar consists of a chunk grammar for concept annotation and (sense) disambiguation rules. The chunk grammar for each term in the lexicon contains at least one grammar rule for recognition of the term. As a preprocessing step we consider annotation with grammatical features and lemmatization of the text. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context such as topic of the text, discourse segmentation, etc. Currently we have implemented chunk grammars for several languages. We have implemented a very simple disambiguator which uses an unigram model.

For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System [13]. The structure of each grammar rule in CLaRK is defined by the following DTD fragment:

```
<!ELEMENT line (LC?, RE, RC?, RM, Comment?) >

<!ELEMENT LC (#PCDATA)>

<!ELEMENT RC (#PCDATA)>

<!ELEMENT RE (#PCDATA)>

<!ELEMENT RM (#PCDATA)>

<!ELEMENT Comment (#PCDATA)>
```

Each rule is represented as a line element. The rule consists of a regular expression (RE) and a category (RM = return markup). The regular expression is evaluated over the content of a given XML element and could recognize tokens and/or annotated data. The return markup is represented as an XML fragment which is substituted for the recognized part of the content of the element. Additionally, the user could use regular expressions to restrict the context in which the regular expression is evaluated successfully. The LC element

contains a regular expression for the left context and the RC for the right one. The element Comment is for human use. The application of the grammar is governed by XPath expressions which provide additional mechanism for accurate annotation of a given XML document. Thus, the CLaRK grammar is a good choice for the implementation of the initial annotation grammar.

The creation of the actual annotation grammars started with the terms in the lexicons for the corresponding languages. Each term was lemmatized and the lemmatized form of the term was converted into regular expression of grammar rules. Each concept related to the term is stored in the return markup of the corresponding rule. Thus, if a term is ambiguous, then the corresponding rule in the grammar contains reference to all concepts related to the term.

The following picture depicts the relations between lexical items, grammar rules and the text:

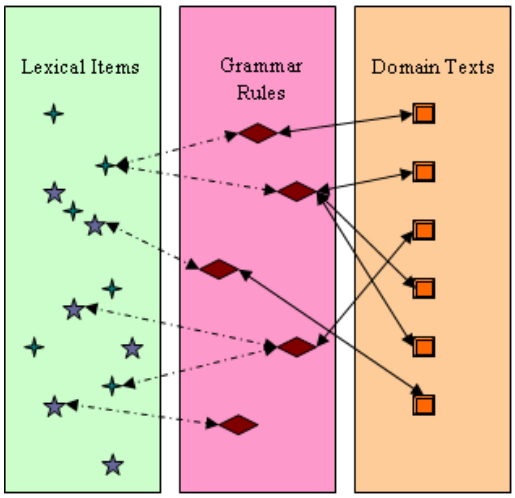


Fig 2. Lexicon-to-Text Relation.

The relations between the different elements of the models are as follows. A lexical item could have more than one grammar rule associated to it depending on the word order and the grammatical realization of the lexical item. Two lexical items could share a grammar rule if they have the same wording, but they are connected to different concepts in the ontology. Each grammar rule could recognize zero or several text chunks.

The relation ontology-to-text implemented in this way provides facilities for solving different tasks, such as ontology search (including crosslingual search), ontology browsing, ontology learning. In order to support multilingual access to semantic annotated corpus we have to implement the relation for several languages using the same ontology as starting point. In this way we implement a mapping between the lexicons in these languages and also comparable annotation of texts in them.

We have been using the relations between the various elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is relied on for the concept annotation of the text. In this way we established a connection between the ontology and the texts. The relation between the lexicon and the ontology is used for definition of user queries with respect to the appropriate segments within the documents. The annotation of texts in different languages on the basis of the same ontology could facilitate the definition of similarity metrics between such texts.

4 A Knowledge-rich Lexicon of Bulgarian

The main problem with the model of the ontology-to-text relation, described in the previous section, is the fact that the annotation of domain texts with domain concepts is very sparse. For example, in the domain of Computer Science for End Users we have annotated 8 concepts within 100 tokens (with 14.8 tokens per sentence = 1.19 concepts per sentence at average). This sparse annotation blocks possibilities for using better methods for word sense disambiguation. This holds when the lexical items in the domain lexicon

are ambiguous among themselves or with respect to the general lexica. For example, the concepts 'key-of-keyboard', 'key-of-database' and 'key-for-door' have the same wording in English and the last one is not from the domain ontology.

We consider two solutions to this problem: (1) better annotation grammar, and (2) Interaction with general lexica. The first can be done by exploiting coreferential relations and lexical chains. The second via connection to lexicons like WordNet. In order to benefit from these solutions, we have to tune them to the model of ontology-to-text relation. First, in order to construct lexical chains and coreferential relations in which the domain terms in the text to participate we need these terms and the surrounding general lexica to share their semantic annotation. In order to ensure this we have to align the general lexica with appropriate semantic information.

Ideally, each meaning of the general lexicon has to be presented in the ontology in order to use the model of ontology-to-text relation from the previous section. Unfortunately such an ontology does not exists yet. Thus, we have to use a smaller ontology and to change the implementation of the ontology-to-text relation.

From our experience within the projects mentioned above we can conclude that there exist a relatively stable upper and middle part of each of the domain ontologies. Thus, for the creation of an appropriate lexical resource for semantic annotation we consider as a first step the building of an upper-middle layer ontology which to provide the necessary semantic information for the tasks of word sense disambiguation. In our case this is a mixture of DOLCE and the upper part of OntoWordNet. Such an ontology can be used for several tasks: (1) representation of general meaning of lexical items in a language; (2) basis for construction of domain ontologies and lexicons.

In the previous model we have used `equality` relation between the conceptual information in the ontology and the meaning of the corresponding lexical items. In this new lexicon this will not be possible because there will be no enough concepts in the ontology. Thus, the first difference from the previous model is that we will allow also the relation `subsume` to be used. The lexicon entry for each lexical item will specify what the relation is between the meaning of the lexical item and the corresponding concept. The requirement for the mapping via `subsume` relation is as follows: the concept that is used with in the ontology to be the most specific one available.

In addition to the mapping to the ontology we want to represent also information necessary for some of the more important phenomena for the task of word sense disambiguation: polysemy, metonymy[2] and verb representation. The first two phenomena – polysemy and metonymy are treated in similar way. First of all, the word senses are represented in the ontology. Thus, the lexical representation is done via appropriate mappings to corresponding concepts in the ontology. Let us consider the case of metonymy. In general, metonymy is defined as a trope in which one entity is used to stand for another associated entity[3]. Thus, we can consider metonymy to be encoded via a composition of ontology relations encoded in the lexicon. For example, let us suppose that we have to annotate the sentence "She was wearing stripe." First we annotate 'stripe' as a kind of a `property` and as such it is connected to 'cloth' via `property-of` relation and 'cloth' is annotated as `material` and it is connected to 'clothing' via the `made-of` relation. The concept 'clothing' is of the relevant type for the object of the verb 'to wear'. Thus, the understanding of the sentence is something like: "She was wearing a clothing made from a textile with a stripe design." The composition of the corresponding relations is stored in the lexical entries for the corresponding lexical items. In the case of metonymy this is a better option, because the possible patterns are (potentially) infinite in number. Representing each metonymy usage as a separate meaning will result in many strange meanings for the lexical items. In this way we separate the most frequent metonymy uses as inference patterns and the actual inference during the analysis of the discourse where the lexical item is used metonymically. Similarly we treat the polysemy. The different meanings are represented in the ontology as different concepts and these concepts are connected via appropriate relations. The main difference here is that for each of the meaning we construct a separate lexical entry. Thus, always during the analysis of the text we have to disambiguate between these senses. In some cases more that one of the senses are visible via one usage of the lexical item. For example, in the sentence "This large book is very interesting." the word 'book' is used simultaneously as a `physical object` selected by 'large' and as an `information object` selected by 'interesting'.

[2] The treatment of metaphorical uses are recorded as separate entries in the lexicon.

[3] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsMetonymy.htm

Encoding of verbs is also very important for the task of semantic annotation. We assume that the appropriate information is also represented in two ways: (1) in the ontology each verb is connected to concept representing the event related to the meaning of the verb. In the ontology all the participants (irrespectively whether they are considered as arguments, adjuncts, etc.) are represented as such via appropriate relations; (2) the linguistic behavior is encoded in the lexicon as a set of frames. These frames determine the role of each participant in the a given event.

The actual lexicon is under construction. It is based on several machine-readable dictionaries: a Morphological Dictionary, a Valence dictionary and an Explanatory Dictionary of Bulgarian. The selection of the lexical items is on the basis of construction of the lexicon aligned to the upper and middle parts of the ontology where we encoded about 3000 lexical entries. The rest of of lexical items are selected on the basis of their ranking in a large Bulgarian corpus (72 million running words from BulTreeBank text archive). The ranks are calculated via automatic morphosyntactic analysis of the corpus and then lemmatization. For each lemma we consider the frequency in the corpus and in how many documents the lemma occurs.

5 Discussion

The need of a knowledge-rich lexicon of Bulgarian is motivated by the need to introduce more world knowledge in the semantic analysis of the text. As it was mentioned in [7], the most lexical relations necessary to determine the semantic content of the lexical items are non-classical in contrast the classical ones, i.e. `hyponymy`, `meronymy`, `antonymy`. The non-classical relations are specific for some classes of meanings, i.e. `made-of`, `used-for`, etc. In our case we assume that these relations are represented in the ontology. Thus, they are formally defined, can be used in inference process and can be used for representation of some language phenomena like polysemy, metonymy, etc.

From point of view of the complexity and precision of ontology according to Nicola Guarino ([5]) we have the following classification of ontologies:

- Lexicon: *Machine Readable Dictionaries; Vocabulary with NL definitions*
- Simple Taxonomy: *Clasifications*
- Thesaurus: *WordNet; Taxonomy plus related-terms*
- Relational Model: *Light-weight ontologies; Unconstrained use of arbitrary relations*
- Fully Axiomatized Theory: *Heavy-weight ontologies*.

The classification starts with less formal and knowledge-poor ontology – simple lexicons and ends with heavily constrained theories about the world. Our attempt is to move the current semantic lexicons from the level of thesaurus to the level of light-ontologies (as a minimum).

Our approach gains in many respects from such works as WordNet [3], EuroWordNet [14], SIMPLE [9]. The mapping between the language specific lexicons was facilitated by the ontology. Our model shares common features with other lexicon models: with WordNet-like ([3]; [14]) lexicons we share the idea of grouping lexical items around a common meaning and in this respect the term groups in our model correspond to synsets in WordNet model. The difference in our case is that the meaning is defined independently in the ontology. With SIMPLE model [9] we share the idea to define the meaning of lexical items by means of the ontology, but we differ in the selection of the ontology which in our case represents the domain of interest, and in the case of SIMPLE reflects the lexicon model. With the LingInfo model ([1]; [2]; [12]) we share the idea that grammatical and context information also needs to be presented in a connection to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools. At the end we would like to mention the work on Ontology Semantics ([11]) which is very similar to our model except that we use existing ontologies like DOLCE and we allow for an incremental construction of the lexicon.

6 Conclusion

In this paper we presented a further developed model for ontology-to-text relation connecting the conceptual information in an ontology to the lexical items and grammatical rules for realization of this information

in tests. We started with domain ontologies and lexicons and then extended the model and their coverage to general lexica. The model represents also phenomena like polysemy, metonymy, verbal frames. The resulting lexicon will ensure better semantic annotation of texts. Our future goals are to implement a system for automatic word sense disambiguation, metonymy usage discovery. Also, the lexicon together with the ontology could be used for the creation of domain ontologies and lexicons.

References

[1] Buitelaar P., Declerck Th., Frank An., Racioppa St., Kiesel M., Sintek M., Engel R., Romanelli M., Sonntag D., Loos B., Micelli V., Porzel R., Cimiano Ph. (2006) LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: *Proceedings of OntoLex06, a Workshop at LREC*, Genoa, Italy.

[2] Buitelaar, P., Sintek, M., and Kiesel, M. (2006). A Lexicon Model for Multilingual/Multimedia Ontologies In: *Proceedings of the 3rd European Semantic Web Conference (ESWC06)*, Budva, Montenegro.

[3] Fellbaum Chr. (1998). Editor. *WORDNET: an electronic lexical database*. MIT Press.

[4] Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In: *Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference*, Springer.

[5] Guarino N. (2000) *Invited Mini-course on Ontological Analysis and Ontology Design.* First Workshop on Ontologies and lexical Knowledge Bases – OntoLex 2000. Sozopol, Bulgaria.

[6] Guarino, N., and Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean*. Communications of the ACM, 45(2): 61–65.

[7] Morris J. and Graeme Hirst Gr. (2004). Non-Classical Lexical Semantic Relations. In: *Proceedings of the HLT Workshop on Computational Lexical Semantics*. Boston, Massachusetts, USA. pp 46–51.

[8] Kiryakov At., Popov B., Terziev Iv., Manov D., and Ognyanoff D. (2005). *Semantic Annotation, Indexing, and Retrieval*. Elsevier's Journal of Web Sematics, Vol. 2, Issue 1.

[9] Lenci A., Busa F., Ruimy N., Gola El., Monachini M., Calzolari N., Zampolli A., Guimier E., Recourcé G., Humphreys L., von Rekovsky U., Ogonowski A., McCauley Cl., Peters W., Peters Iv., Gaizauskas R., Villegas M. (2000). *SIMPLE Work Package 2 – Linguistic Specifications, Deliverable D2.1*. ILC-CNR, Pisa, Italy.

[10] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). *Ontology Library (final)*. WonderWeb Deliverable D18, December 2003.
`http://www.loa-cnr.it/Publications.html`

[11] Nirenburg S. and Raskin V. (2004). *Ontological Semantics.* MIT Press.

[12] Romanelli, M., Buitelaar, P., and Sintek, M. (2007). Modeling Linguistic Facets of Multimedia Content for Semantic Annotation. In: *Proceedings of SAMT07 (International Conference on Semantics And digital Media Technologies)*, Genova, Italy, pp 240–251.

[13] Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK – an XML-based System for Corpora Development. In: *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK.

[14] Vossen P. (1999). Editor. *EuroWordNet General Document. Version 3, Final, July 19, 1999*.
`http://www.hum.uva.nl/~ewn`

Non-Technical Computer Thesaurus versus Specialized Computer Thesaurus

Dr. Olena Siruk

Kyiv National Taras Shevchenko University

Abstract. This paper is devoted to a comparative analysis of the **Computer Thesaurus of Ukrainian Verbs** and the **Specialized Thesaurus of Computer Ideography**. These two dictionaries are representative examples of a general language (non-technical) computer thesaurus and a specialized computer thesaurus. We focus our attention on the entries of each thesaurus, its macrostructure, microstructure, compilation and use.

1 Introduction

One of the most important tasks of modern lexicography is the design of dictionaries that would satisfy the exigent demand of today's information-aware society for systematized linguistic information at the level of world standards. As a consequence, **thesauri** attract the special attention of specialists as dictionaries which not only inventory but also systematize lexical units within the limits of the required linguistic subsystem. The level of development of information technologies in Ukraine allows, and the users' necessities require, concentration on the development of **computer thesauri** of different types: non-technical as well as specialized terminological thesauri. This work is conducted by the employees of the Laboratory for Computational Linguistics of the Institute of Philology (Kyiv National Taras Shevchenko University) within the framework of scientific research dedicated to formalization in linguistics [2: 3–10; 4: 84–87].

For want of Ukrainian ideographical dictionaries (not only computer but also paper ones), and also because of the current state of lexicographic research, virtually no work has been done so far on developing the terminology of such a 'young' linguistic industry as computer ideography. This is why, during the composition of the Computer Thesaurus of Ukrainian as a dictionary which would satisfy the necessity of Ukrainian lexicography in non-technical computer thesauri at least to some extent, the terms of this linguistic area had to be defined and systematized. For the purposes of both projects a review of the literature of linguistic semantics and lexicography was performed, and ready-made linguistic products available in libraries and on the Internet (15 paper thesauri and more than 50 computer thesauri) were also analysed.

The **Computer Thesaurus of Ukrainian Verbs** and the **Specialized Thesaurus of Computer Ideography** are examples of computer ideographical dictionaries with different thematic orientation (non-technical and specialized thesauri, respectively). The thematic orientation of an ideographical dictionary is one of its principal characteristics, affecting its **composition, structure, design features and use**.

2 Non-technical thesauri

Non-technical thesauri, as primarily non-alphabetical dictionaries which reflect systematic semantic relations between units explicitly, represent the vocabulary of the entire language and, as a rule, are voluminous (for example, the computerized Roget Thesaurus, Merriam–Webster Online Thesaurus, Visual Thesaurus, CARMEN, SWD, EuroWordNet, BalkaNet, RussNet etc.). A non-technical thesaurus of the order of thousands of words and expressions counts as concise. **Specialized dictionaries** or **dictionaries of sublanguages** represent terminological systems of individual branches of science. Here we may name such computer systems as NASA Thesaurus of aeronautics (NASA Thesaurus), Agricultural Thesaurus (AGROVOC), Thesaurus of archaeological objects (Archaeological Objects Thesaurus), The Astronomy Thesaurus, Bioethics Thesaurus, Cambridge Life Sciences Thesaurus (Cambridge Scientific Abstracts), Thesaurus of biology of animals (Tesauro ICYT de biología animal (CINDOC)), INFODATA Thesaurus of

information and documentation (INFODATA. Thesaurus für den Bereich der Information und Dokumentation), POPIN Thesaurus (Population Multilingual Thesaurus), dictionary and thesaurus of military terminology (CALL Dictionary and Thesaurus (US Government)), Thesaurus of the Terminology of Gender Research by A. Denisova etc. On the Internet such terminological resources are implemented in the form of dictionaries of concepts and terms with links between them. The fundamental purpose of a dictionary of this type is to help in the process of information retrieval: the query is expanded on the basis of the links in the thesaurus, and the navigation based on them facilitates the accurate formulation of the query. A specialized thesaurus which contains 150–200 units is considered complete.

3 Units of a Thesaurus

The Specialized Thesaurus of Computer Ideography comprises 75 terms. In comparison, in the semantic field of speech in the Computer Thesaurus of Ukrainian Verbs the verbal lexical-semantic variants alone number about two thousand. It should be noted that although the Computer Thesaurus of Ukrainian Verbs and the Specialized Thesaurus of the Computer Ideography are autonomous constituents of larger projects (namely the Computer Thesaurus of Ukrainian and the Thesaurus of Applied Linguistics), the substantial difference in the quantity of units in favour of the non-technical thesaurus will remain or even grow due to the increase of the register of the Computer Thesaurus of Ukrainian by addition of other parts of speech, in particular nouns, which are substantially more numerous in the language than verbs. The Specialized Thesaurus cannot count on a considerable increase by verbs due to the nature of its units. The units of the Specialized Thesaurus of Computer Ideography are characteristic of this kind of dictionary. The terms are represented by nouns and two- to four-word noun-noun or adjective-noun compounds (N+N, Adj+N, Adj+Adj+N etc.), also in the form of abbreviations. The overwhelming majority of units only relate to the indicated area of knowledge (*комп'ютерний тезаурус (КТ) 'computer thesaurus (CT)', розширений КТ 'extended CT', методика укладання КТ 'methods of composition of CT'*), but there are also terms shared with other linguistic domains (*тезаурус 'thesaurus', ідеографічний словник 'ideographical dictionary'*—with lexicography; *семантичне поле 'semantic field', сема 'seme', лексико-семантичний варіант 'lexical-semantic variant', антонімія 'antonymy', гіпонімія 'hyponymy', синонімія 'synonymy'*—with lexicology; *база даних 'database', лінгвістичний процесор 'linguistic processor', лінгвістичний алгоритм 'linguistic algorithm'*—with computational linguistics in general). A minority of terms are united by relations of synonymy as well as subsumption. They mostly denote concepts already established in the literature, shared with other sections of linguistics (*гіпернім 'hypernym'* and *гіперонім 'hyperonym'*, *ідеографічний словник 'ideographical dictionary'* and *тезаурус 'thesaurus'*, *семантичне поле 'semantic field'* and *лексико-семантичне поле 'lexical-semantic field'*, *семна структура 'structure of semes'* and *семний набір 'set of semes'*, *ядро семантичного поля 'nucleus of the semantic field'* and *центр семантичного поля 'centre of the semantic field'*, *ядерна сема 'nuclear seme'*, *концептуальна сема 'conceptual seme'* and *центральна сема 'central seme'*).

The fact that a specialized thesaurus is usually restricted to nouns (the part of speech that is prevalent in terminology) whilst in a non-technical thesaurus practically all parts of speech are represented along with set expressions (phraseological units and proverbs) is yet another difference between these dictionaries and it draws attention to the difference between the composition of verbal and nominal vocabulary.

Since significative semantics prevails in the meaning of a verb and verbs belong to the analytical vocabulary, verbal meaning is not correlated directly with a subject domain but explicates a the relation between objects [6: 51]. This feature directly influences the method of working with verbal (as opposed to substantival) material. In light of this for verbs

1. an *internal, significative* concept selection strategy *based on the analysis of meaning* is more acceptable;
2. an *inductive approach* to ordering lexemes is more adequate;
3. relations based on *word-formation type* (derivation hyponymy) and *valency potential* (a basis for connections between parts of speech) are *essential*;
4. taxonomy, whole–part relations are *irrelevant*.

The experience from working with English, Spanish, German, Russian thesauri on the Internet shows that verbs are included in the different types of thesauri considerably less often than nouns, and especially seldom in terminological thesauri.

The basis for the semantic scheme of nouns is the external picture of connections between objects and phenomena, adopted from objective extralinguistic reality. The categorization of nouns on a denotative basis is predefined by the categorial nature of nouns, which are predominantly oriented to the reflection of objective reality [1, 180–181]. Consequently, for a noun

1. *external, denotative* choice of concepts is characteristic;
2. a *deductive approach* to structuring the material is mostly applied;
3. word-formation and the valency potential of a noun *are not very important* for the creation of the synoptic scheme;
4. whole–part relations *are substantial*, taxonomy *is prevalent*.

It is precisely the noun that holds the garland in ideographical dictionaries of different languages with respect to the development of foreign-language thesauri of any type.

All these characteristics are reflected in the theoretical principles of the dictionary's construction, which correlate with the micro- and macrostructure of computer thesaurus, in particular they predetermine the filling of fields in its entries. Although the general structure of the vocabulary entry for nouns and verbs is of the same type and consists of three main components (headword and lexemes related to the headword by interverb/internoun and inter-part-of-speech relations), there is substantial differentiation at a deeper level. Not only are there verbs connected by relations of synonymy, antonymy and hyponymy (which is also true of nouns), verbs are also characterized by a high frequency of phonetic variants, a ramified net of derivational relations based on the semantics of modes of action, a network of relations on the basis of the valency potential of the verb, and dependence of the structure of the vocabulary entry on the derivational structure of the verb (i.e., on whether it is derived or not).

The fact that a specialized thesaurus is based on the dominant scientific conception, whereas the synoptic scheme of a non-technical thesaurus is constructed under the influence of ideological and world-view factors, constitutes another substantial difference between the dictionaries. As previously noted, there are differences of principle in the character of the lexical material presented in these dictionaries. This implies that the reflection of the lexical system in a specialized thesaurus is predetermined by external circumstances, by the term system of the described domain, whereas a non-technical thesaurus chiefly models the semantic system of the language, putting aside the linguistic picture of the world.

Jury Karaulov endeavours to find the intersection of the construction principles of non-technical thesauri and the design rules of specialized thesauri for information storage and retrieval [3]. Both types have certain common, analogous and uniting features:

1. both dictionaries represent more or less completely the relations between units;
2. both dictionaries either have an explicit synoptic scheme, that is a division of the universe into thematic classes, or such a scheme is present implicitly;
3. the rubric (a class of synonymous words in non-technical thesauri and a descriptor article in specialized thesauri) serves as interpretation, or as context, in both dictionaries;
4. there are cross-references between entries in both dictionaries.

The features of the lexical semantics of verbs conditions the difference between an ideographical dictionary of nouns and an analogous dictionary of verbs with respect to the organization of its external structure (macrostructure), in the methods of display and description of the lexical categorization of nouns. Verbs have been categorized primarily on a semantic basis, using the method of component analysis and stepwise identification of verbal meanings.

4 Macrostructure of thesauri

The interface of the Specialized Thesaurus of Computer Ideography have two windows. In the left-hand window is the permutation index of the dictionary. It has the form of a tree of terms whose the levels can be expanded if there is a '+' mark on the left. The zeroth level of the specialized thesaurus is represented by

the term *комп'ютерна лексикографія* 'computer lexicography', hyperonym of the concept of the first level *комп'ютерна ідеографія* 'computer ideography'. The second level contains 4 concepts: *одиниці КТ* 'units of CT', *відношення між одиницями КТ* 'relations between units of CT', *комп'ютерний тезаурус* 'computer thesaurus' and *укладання КТ* 'CT design', which contain 5, 8, 10 and 6 terms of the third level respectively. The maximal depth of the hierarchies in the Specialized Thesaurus of Computer Ideography is six intervals, and in the Computer Thesaurus of Ukrainian verbs it is seven, which corresponds to the conventional constant of depth of any thesaurus [3, 186–187]. The entries of both thesauri are in a thematic-alphabetical order.

5 Microstructure of thesauri

An entry of the Specialized Thesaurus of Computer Ideography consists of a head term, located in the window on the left, and a definition. To find the definition of a required term one has to select it with the mouse and push the button 'Тлумачення' ('Interpretation'). Thereupon some text will appear in a window on the right. The definition mostly consists of a hyperonym specified by differentiating semes (*Багатомовний КТ – комп'ютерний тезаурус, орієнтований на ідеографічну структуру одночасно декількох мов 'Multilingual CT: a computer thesaurus oriented simultaneously to the ideographical structure of several languages'*), but can also be more extended, approaching an encyclopaedic definition, when characterizing a concept (*Комп'ютерний тезаурус (КТ) – представлений за допомогою комп'ютера ідеографічний словник. Під цим терміном об'єднуються комп'ютерна версія тезауруса та власне комп'ютерний тезаурус. КТ може бути загальномовним або спеціалізованим (за тематичною спрямованістю), одномовним чи багатомовним (за мовою виконання), мінімальним або розширеним (за повнотою викладу). Окремим видом КТ є авторський комп'ютерний тезаурус. Дослідженням КТ займається комп'ютерна ідеографія 'Computer Thesaurus (CT): an ideographical dictionary presented with the aid of a computer. This term subsumes computer versions of thesauri and computer thesauri proper. A CT can be general or specialized (by its thematic orientation), unilingual or multilingual (by its language of implementation), minimal or extended (by the completeness of its exposition). A separate type of CT is the author computer thesaurus. The research of CT is a topic within Computer Ideography'*). The semantization of the headword in the Computer Thesaurus of Ukrainian Verbs is performed through a definition from the 11-volume explanatory dictionary of Ukrainian. If the definition is a logical explanation of the concept, a statement of its content and distinctive features as is characteristic of encyclopaedic and terminological dictionaries, the interpretation exposes/reveals the meaning of the linguistic unit from the point of view of the naive picture of the world. The dictionary entry of the Computer Thesaurus of Ukrainian Verbs is set up in a separate window. It can be either only verbal *(simple)* or broadened, as a result of the integration of the verbal part of CT into the Computer Thesaurus of Ukrainian, by the relationships of the verb with substantival, adjectival (participial) and adverbial (gerundial) vocabulary *(extended)*. Such relations appear on the basis of the existence of additional semes: "actor", "instrument of action", "product of action", "process", "place of action", "reified action, abstraction", "one characterized by the action", "in accordance with the qualities of the action". Fig. 1 shows the extended vocabulary entry of the verb *базікати* 'jabber', where apart from the interverbal relations represented by the hyperonym *вимовляти* 'pronounce', 9 synonyms (*нести*, *верзти*, *варнякати*, *просторікувати*, *ляпати*, *торочити*, *плескати*, *молоти*, *патякати*) and 2 verbs denoting modes of action (the cumulative *набазікати* and the supercompletive *добазікатися*) one can see the relations between the verb and nouns (1 'actor' *базіка* 'chatterer' and 1 'process' *базікання* 'jabber') highlighted by a red background, and between the verb and a participle (1 'attribute' *балакучий* 'talkative') marked by a green background.

The basic **form of presentation** of both Thesauri is **on the computer**. There are databases in Microsoft Access format and a program written in C#. The Specialized Thesaurus of Computer Ideography exists in parallel on paper and online, on the pages of the Linguistic portal of MOVA.info in the section 'Dictionaries' `http://www.mova.info/toc.asp?PP=16&tocPath=1`.

The **advantages** of computerizing thesauri can be seen in such areas as sorting material in a database (a computer dictionary is an open system: a database can be augmented and edited, a paper version cannot), the speed of the work with the dictionary (thanks to the multiple entrances, especially to the search system,

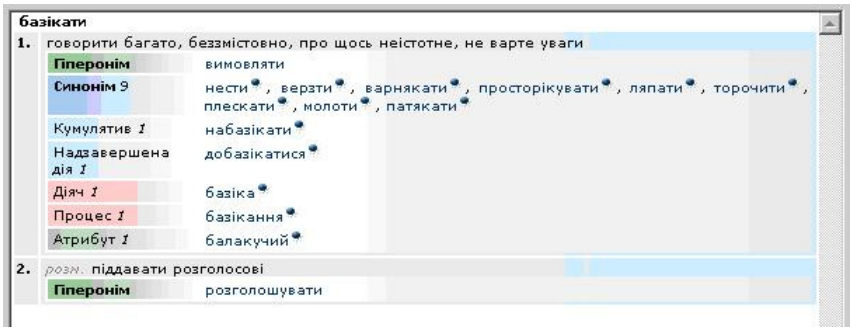


Fig. 1. An extended vocabulary entry.

cross-references in definitions and the possibility to complete and edit the database) and the integration of the product into a network of linguistic software (peculiar to computer dictionaries).

6 Search system

Computer thesauri have two entrances: for a synoptic scheme (permutation index) and system of search for a lexeme and its parts, which substantially simplifies and speeds up the work. Apart from this, the definitions of the Specialized Thesaurus of Computer Ideography contain italicized cross-references to terms they use (**Укладання КТ** — процес створення *КТ*, або розроблення *макроструктури КТ*. Складається з трьох основних завдань: створення *бази даних*, *лексикографічного процесора* та вироблення формату *словникової статті*, або *мікроструктури КТ* '**Composition of CT**: the process of creation of CT, or of development of the *CT's macrostructure*. Consists of three basic tasks: creation of a *database*, a *lexicographic processor* and developing the format of the *dictionary entry* or the *CT's microstructure*'), regardless of which concept's boundaries they are in. In the dictionary entry of the Computer Thesaurus of Ukrainian Verbs all the semantic variants marked by relations towards the headword are references to the corresponding articles.

7 Application

The Specialized Thesaurus of Computer Ideography is intended for scholars and students of philology. It can be used as an information system or for the purposes of education. The Computer Thesaurus of Ukrainian Verbs has a wider audience: thanks to its specification, it can be used as a multi-level information system and as a base for further linguistic research. Due to the possibility of integration into a network of linguistic software, the Computer Thesaurus of Ukrainian Verbs was used, together with a package of additional utilities, for the analysis of the features of the style of the distinguished Ukrainian writers Lina Kostenko and Vasyl Stus [5: 246–251].

Experience with the analysis of lexicographic materials and a significant number of ideographical dictionaries on the Internet enabled us to 1) systematize the terminology of computer ideography in the form of a Specialized Thesaurus of Computer Ideography; 2) develop a formalized method of composition of the non-technical Computer Thesaurus of Ukrainian Verbs as an information and research system. The comprehensive comparison of these dictionaries as examples of a non-technical and a specialized thesaurus can find a place in lectures, advanced courses and specialized seminars on the problems of creating computer dictionaries and the formalization of lexical semantics and on the whole will be advantageous both for philologists, in particular practising lexicographers, and for general users.

References

[1] *Бабенко Л. Г.* Принципы категоризации именной лексики в толковом идеографическом словаре существительных русского языка // Русский язык: исторические судьбы и современность: II Международный конгресс исследователей русского языка. Труды и материалы / Составители М.Л. Ремнева, О.В. Дедова, А.А. Поликарпов. — М., 2004. — С. 180–181.

[2] *Дарчук Н., Денисенко І., Сірук О., Сорокін В.* Ідеографічний тезаурус української мови // Вісник Черкаського університету, Вип. 24. Серія "Філологічні науки". — Черкаси, 2001. — С. 3–10.

[3] *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. – М., 1981.

[4] *Сірук О.* Два підходи до побудови комп'ютерного тезауруса дієслів української мови // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 31. — К., 2004. — С. 84–87.

[5] *Сірук О.* Статистичний аналіз художнього тексту за допомогою комп'ютерного тезауруса: недоліки і переваги // Мовні і концептуальні картини Світу. Вип. 12. Ч–на 2. — К., 2004. — С. 246–251.

[6] *Уфимцева А. А.* Семантика слова // Аспекты семантических исследований. — М., 1980.

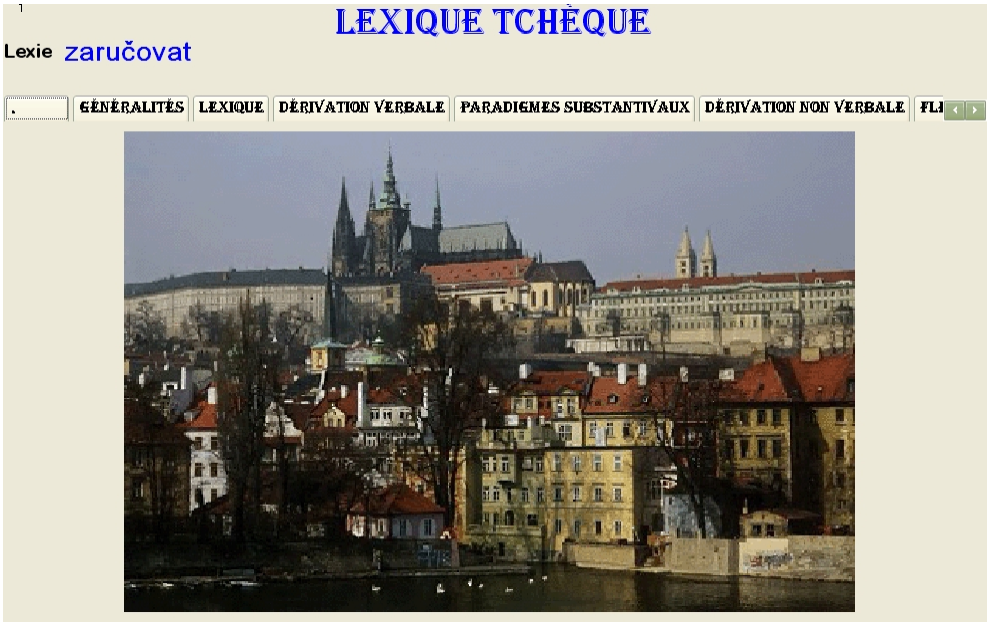
Définition d'un prototype général de bases de données (étude des langues slaves de l'Ouest dans une visée multilingue)

Patrice Pognan

Professeur à l'INALCO
LALIC (Université de Paris-Sorbonne et INALCO)

Ces bases de données auront le français comme pivot de l'outil (bases de données langue étrangère – français, puis français – langue étrangère) mais aussi comme langue d'interface homme – machine. Il est appréciable pour le monde francophone d'avoir la confrontation simultanée de plusieurs langues face au français (plutôt qu'à l'anglais, ce qui est relativement fréquent).

Elles seront construites avec le souci constant du multilingue et même celui du traitement de langues individuelles dans une visée multilingue (nous développons des travaux sur l'enseignement des langues par groupes linguistiques – essentiellement, pour le moment, pour les langues slaves de l'Ouest, mais l'appréhension d'une langue comme le berbère ne peut se faire que dans le cadre d'études plurielles sur les différents parlers).



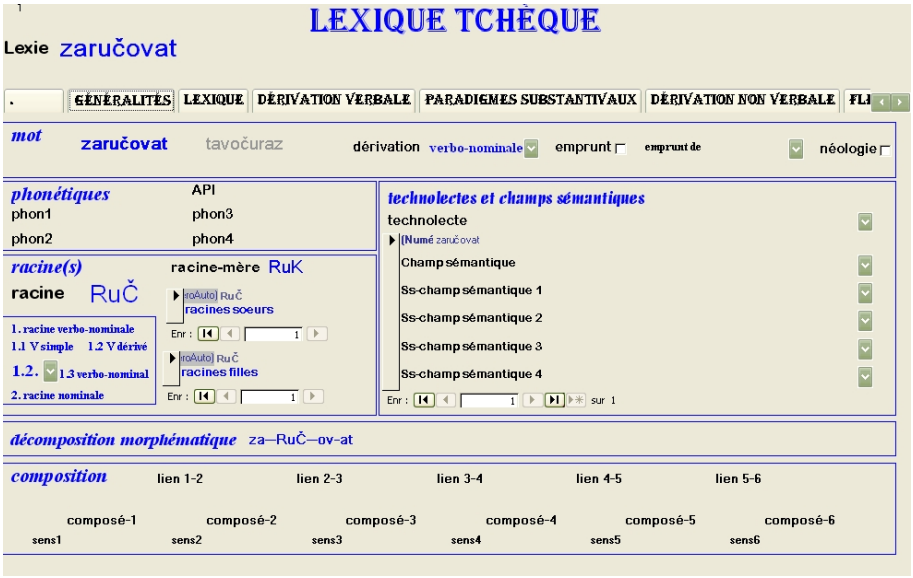
Les bases de données élaborées auront de manière essentielle une structure commune, constituée de plusieurs composantes importantes:

1. Informations générales

Elles seront données dans une classification à partir des mots, mais étant donnée la présence de langues chamito-sémitiques, la structure donnera la possibilité simultanée de fournir un classement par racines. Quelle que soit la langue, y compris indo-européenne, la production de lexiques et de dictionnaires classés aussi par racines est bienvenue. A ces fins, les informations générales présenteront un découpage morphématique qui permettra de mieux définir la racine et ses variantes.

Seront aussi données des informations sur le statut éventuel d'emprunt (si oui, à quelle langue) ou de néologie ainsi que sur une éventuelle composition. Des indications sur le champ terminologique éventuellement concerné permettront de produire des lexiques spécialisés.

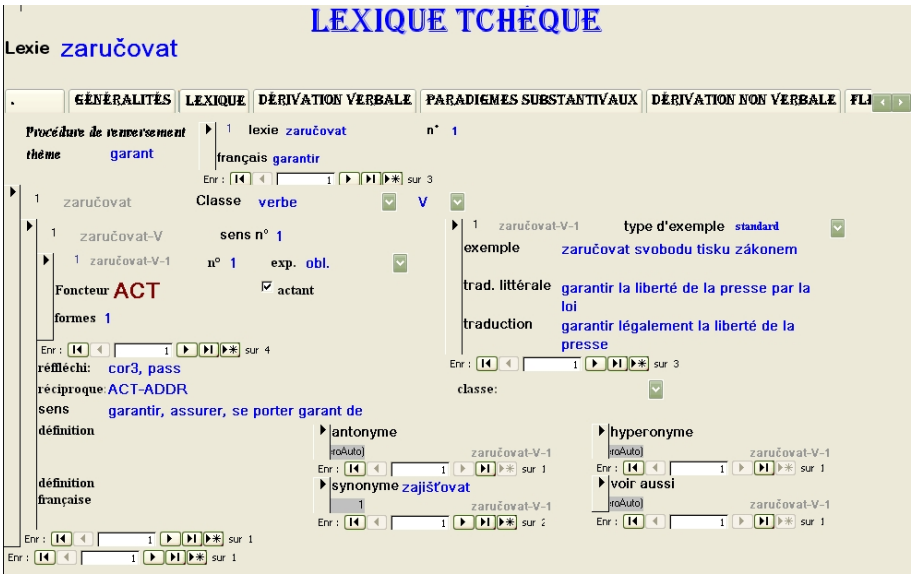
Pour chaque entrée, la base de données calcule immédiatement la forme inversée du mot pour la construction d'un dictionnaire rétrograde (« a tergo ») très utile dans l'étude des marques de catégories lexicales et des éléments fonctionnels (suffixations de langues agglutinantes, désinences de langues à flexion externe).



2. Structuration lexicale

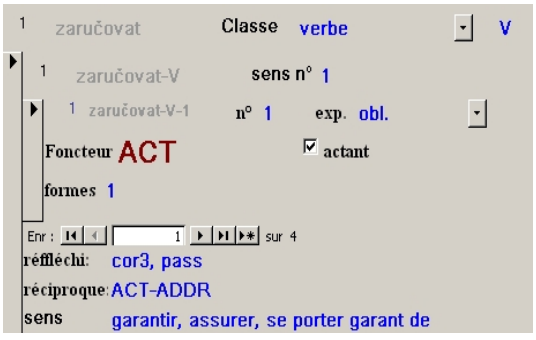
C'est une structure qui permet essentiellement la production de dictionnaires. Nous devons prévoir une structure de la base de données telle qu'elle enregistre toutes les catégories lexicales d'un mot donné, puis toutes les significations possibles pour une catégorie lexicale d'un mot. Chacune de ces significations doit pouvoir englober d'autres tables: celle des exemples avec les traductions, y compris mot à mot, celle des synonymes, celle des antonymes, … suivant l'organisation donnée ci-après :

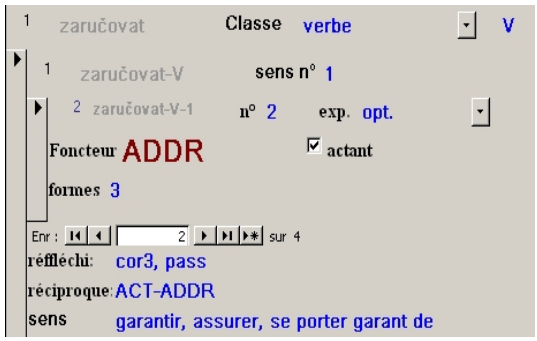
1. mot et racine afférente
    2. catégorie lexicale (un même mot peut en avoir plusieurs)
        3. significations (il peut y en avoir plusieurs par catégorie lexicale)
            4. exemples (avec les traductions) par signification
            4. cadre verbal syntaxico-sémantique (par catégorie lexicale – signification)
            4. synonymie (par catégorie lexicale – signification)
            4. synonymie floue ou analogie (par catégorie lexicale – signification)
            4. antonymie (par catégorie lexicale – signification)
            4. hyperonymie (par catégorie lexicale – signification)

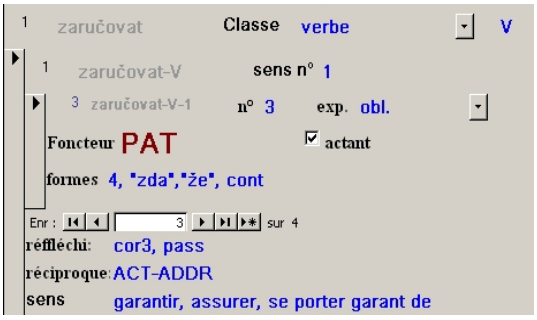


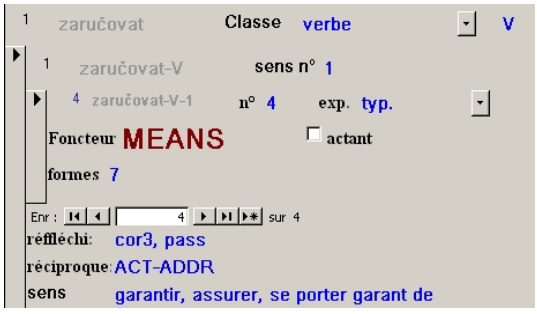
Les études de syntaxe profonde seront menées en coopération avec nos partenaires tchèques à l'aide du dictionnaire de valences lexicales (Vallex), résultat de longues recherches sur la division thème-rhème, l'ordre systémique et le cadre verbal.

rang	mcatsens	fonct-num	foncteur	exposant	formes	actant
1	zaručovat-V-1	1	ACT	obl.	1	☑
2	zaručovat-V-1	2	ADDR	opt.	3	☑
3	zaručovat-V-1	3	PAT	obl.	4, "zda","že", cont	☑
4	zaručovat-V-1	4	MEANS	typ.	7	☐









3. Renversement de la base

Associée à la structure lexicale, nous avons développé une procédure simple de renversement de la base. Cette procédure permet de construire une nouvelle base de données français – langue étrangère qui contient tout le matériau lexical et grammatical de la base langue étrangère – français. Cette procédure assure la quasi équivalence de la masse lexicale dans les deux bases de données et donc des deux dictionnaires bilingues qui en découlent.

La nouvelle base de données français – langue étrangère doit répondre à la conception que l'on se fait d'un tel dictionnaire à l'usage de francophones. Il est donc nécessaire de construire la structure adéquate et de reconstruire le dictionnaire correspondant. Le fait de pouvoir puiser dans une (ou des) table(s) englobée(s) dans la base de données accélère de manière sensible la construction du dictionnaire français – langue étrangère. Cette procédure sera présente sur toutes les bases de données. Nous l'avons testé sur un modèle ancien de base de données slovaque – français pour créer le modèle français – slovaque (contrat Lingua II – ALPCU – Découvrir et pratiquer le slovaque, 2007). La dissymétrie entre les deux lexiques a été très nette: sur la base du lexique slovaque – français qui avait 1200 entrées, nous avons obtenu un lexique français – slovaque qui n'en avait plus que 1000.

4. Composantes flexionnelles

Elles seront toutes assurées par des tables secondaires imbriquées dans la table principale. Elles permettront la production d'ouvrages utiles, par exemple de conjugaison du type « 201 / 301 … verbes x ». Les tables proposées sont :

- table de conjugaison
- table de déclinaison des substantifs
- selon les cas, éventuellement table de déclinaison des adjectifs.

Les langues étudiées, et particulièrement les langues slaves de l'Ouest, sont hautement flexionnelles et il serait fastidieux, voire impossible de renseigner ces tables annexes concernant la flexion (conjugaison, flexion des substantifs et, éventuellement, des adjectifs) si nous ne faisions pas appel à une composante de génération automatique des formes régulières.

Cela signifie la définition et la réalisation d'une véritable composante de génération automatique de formes pour l'intégralité d'une langue. Il convient donc, dans ce cadre, de réviser et définir précisément tous les paradigmes flexionnels qui peuvent répondre des générations régulières. L'opération de génération automatique est simplifiée grâce à un ordre de circulation dans la base de données tel que, lors du déclenchement de l'opération de génération des formes, il y ait une consultation automatique de champs préalablement renseignés comme par exemple l'indication d'un paradigme de flexion. Les champs flexionnels sont laissés accessibles de manière à ce que d'éventuelles erreurs de génération puissent être corrigées manuellement. L'idiosyncrasie, en particulier au niveau de verbes ou de substantifs hautement irréguliers, ne sera pas générée automatiquement, mais laissée au soin du constructeur de la base.

La sous-classe des verbes tchèques terminés en -ovat (3ème classe, 2ème sous-classe), qui est parfaitement régulière tant dans les formes personnelles qu'au niveau des participes, des gérondifs et des adjectifs qui en découlent, nous a servi de banc d'essai pour tester la faisabilité. Les opérations flexionnelles pouvant atteindre une complexité certaine, nous emploierons un langage de programmation adéquat pour le faire.

Nous présenterons deux exemples de complexité de flexion:

Le premier concerne la génération de la forme masculine animée au nominatif pluriel des adjectifs durs où la désinence molle en « –í » provoque la palatalisation (mutation faisant passer une consonne dure à la consonne molle correspondante) de la dernière consonne ou du dernier groupe consonantique du radical : « pražský doktor » (un docteur pragois) devient au nominatif pluriel « pražští doktoři », ce qui met en évidence l'alternance « sk » / « šť » avec une règle de réécriture du dernier segment en « št » devant « –í ».

Le second exemple est nettement plus complexe: il concerne la génération du génitif pluriel des féminins et neutres durs. Ces formes ont une désinence « zéro ». Lorsque le radical du mot se termine par un groupe de consonnes (c'est-à-dire au moins deux), il faut calculer s'il est nécessaire d'insérer un « e » intercalaire « épenthétique ». Nous avons pu établir un ensemble de 4 règles qui permet ce calcul. L'une d'entre elles est le caractère étranger du mot qui interdit l'insertion, ce qui veut dire qu'il faut trouver les critères qui marquent l'emprunt.

C'est un ensemble de 12 sous-modules du module « reconnaissance des emprunts », présenté précédemment, qu'il faut intégrer dans le module de flexion. A leur tour, certains de ces sous-modules requièrent la présence d'une composante de phonologie historique. Ainsi, un mot tel que « embargo » sera reconnu comme étranger d'abord par la présence de « g », puis du « e » en tête de mot et enfin par la trace de la nasale française en « emb » et la procédure de reconnaissance des emprunts communiquera à la procédure de flexion une qualité d'emprunt, ce qui interdira l'insertion d'un « e ».

A partir du programme qui sera mis en place dans la base de données (en particulier pour le tchèque), nous aurons la possibilité d'obtenir une composante de génération flexionnelle parfaitement autonome et réutilisable dans la construction de systèmes de TAL.

Remarquons que le cumul des formes calculées dans ces tables permet d'obtenir un dictionnaire de formes d'une langue, ce qui trouve de nombreuses applications en TAL.

5. Composantes « sciences classificatoires »

Lorsque l'entrée est un mot qui appartient à un domaine de sciences avec une tradition classificatoire, en particulier des sciences de la vie (p. ex. zoologie, botanique, mycologie), des sciences de la terre ou de la chimie, cette entrée sera également décrite dans le cadre approprié de sa discipline à l'aide de tables annexes spécialisées.

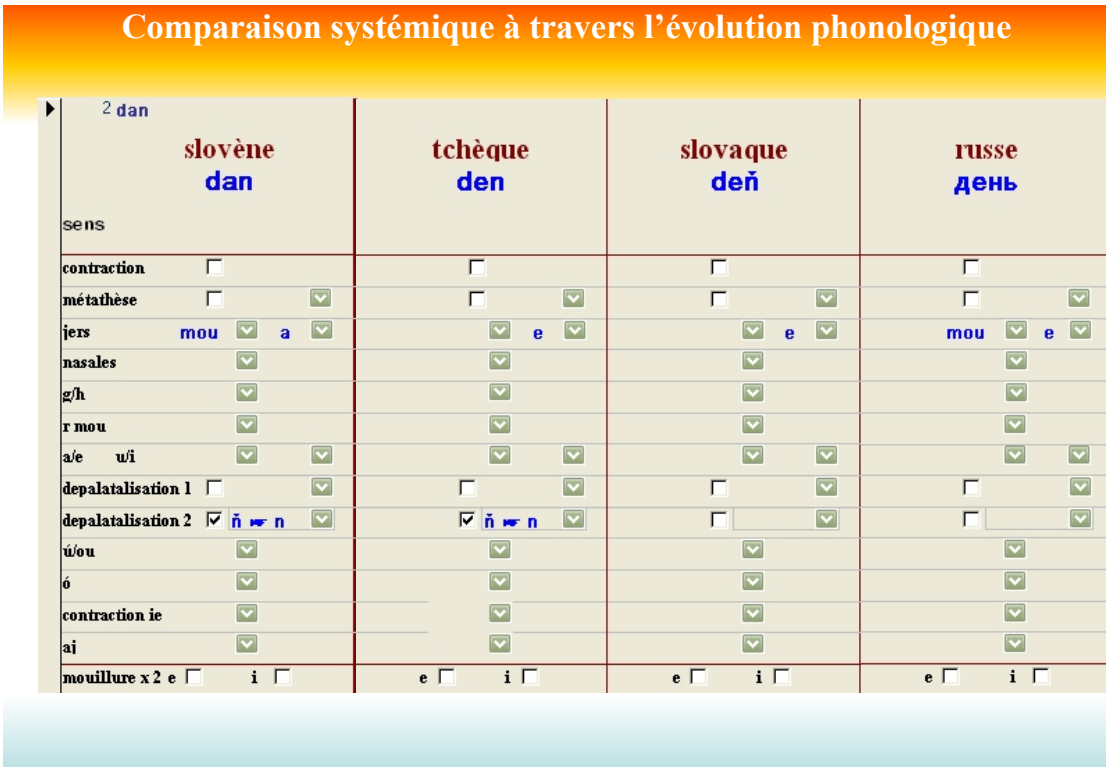
En sciences de la vie, la table annexe présentera, à côté du report de l'entrée, la nomenclature appropriée (binôme de Linné), les désignations populaires et l'ensemble des termes de la classification en latin (la langue de référence dans cette table), en français et dans la langue étrangère étudiée. En zoologie, un cours bilingue tchèque – français professé pendant 6 ans nous donne déjà une part du matériau nécessaire.

6. Composante « onomastique »

Initiée par le congrès « Primer Col-loqui Internacional sobre la Toponimia Amaziga » à Barcelone en 2008, elle sera définie et réalisée en relation avec plusieurs organismes. Sa définition est en cours de test dans le prototype berbère. Elle sera ensuite implantée dans le prototype général et dans toutes les bases dérivées.

7. Composante multilingue

Dans le cas des langues slaves de l'Ouest qui nous sert de référence pour ces travaux, la prise en compte de l'évolution historique de la phonologie permet une radiographie très précise des lexiques des langues concernées Elle permet de percevoir nettement les phénomènes de distanciation progressive des langues d'un même groupe entre elles et de pouvoir en extraire des éléments déterminants d'un apprentissage global du groupe de langues. Les faits de phonologie historique mis en évidence sont d'une grande importance pour l'analyse automatique de l'état synchronique d'une langue.



Dans tous les cas, la base de données doit structurer la configuration la plus étendue du système linguistique. Chaque langue est ensuite définie et enregistrée dans un ensemble égal ou inférieur à celui du système. Cette démarche permet d'assimiler le système linguistique global, les limites particulières de chacune des langues du groupe et les phénomènes linguistiques diachroniques ou synchroniques qui déterminent la variation lexicale. Il nous semble que, menée à terme, une telle démarche est susceptible d'engendrer des moyens d'apprentissage d'un groupe de langues en évitant les phénomènes de confusion qui parsèment les apprentissages successifs sans lien les uns avec les autres.

8. État du projet

Issus d'un prototype général, les prototypes de base de données pour le tchèque et le slovaque sont prêts, en dehors de la composante de génération automatique seulement testée. Sur la base du tchèque seront dupliquées les bases de données pour le polonais, le haut-sorabe et le bas sorabe, le slovène (le choix des paradigmes de déclinaison et de conjugaison n'est pas terminé) et le russe.

Signalons que le même prototype de base de données débouche sur la réalisation de bases de données pour le berbère tachelhit (chleuh) et le berbère tamazight (Maroc central) avec un très gros projet, en cours de réalisation, liant traitement automatique et bases de données et portant sur le dictionnaire raisonné berbère – français de Miloud Taïfi (près de 7200 racines).

Références

[1] Lamprecht, A., Šlosar, D. & Bauer, J. (1986), „*Historická mluvnice češtiny*“. Státní Pedagogické Nakladatelství, Prague.
[2] Lopatková, M., Žabokrtský, Z., Kettnerová, V. (2008), „*Valenční slovník českých sloves*“. Karolinum, Prague.
[3] Pognan (1983), “*Une reconnaissance automatique des mots étrangers dans les textes scientifiques. Un essai en langue tchèque*”. The Prague Bulletin of Mathematical Linguistics n° 40. Prague.
[4] Pognan (1998), “*Histoire de l'écriture et de l'orthographe tchèques*”. Histoire, Epistémologie, Langage. Paris.
[5] Pognan (2001), “*Introduction aux systèmes d'écriture des langues slaves de l'Ouest (polonais, bas-sorabe, haut-sorabe, tchèque, slovaque)*”.Slavica occitania. Toulouse.
[6] Taïfi, Pognan (2008), “*Dualité toponymique au Maroc: projet de répertorisation et description linguistique*”. Primer Col-loqui Internacional sobre la Toponimia Amaziga. Barcelone.

Authors

Igor Boguslavsky, Laboratory of Computational Linguistics; Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia & Madrid Polytechnic University, Spain.

Ivan Derzhanski, Department for Mathematical Linguistics, Institute for Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Viacheslav Dikonov, Laboratory of Computational Linguistics; Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.

Ludmila Dimitrova, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Ralitsa Dutsova, Veliko Tărnovo University & Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Tomaž Erjavec, Department of Knowledge Technologies at the Jožef Stefan Institute, Ljubljana, Slovenia.

Tatyana Frolova, Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.

Radovan Garabík, Slovak National Corpus department, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia.

Ján Genči, Technical university of Košice, Slovakia

Leonid Iomdin, Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.

Oskár Kadlec, Civic Association For Slovak Medical Terminology, Bratislava, Slovakia

Maria Khokhlova, Faculty of Philology and Arts, St. Petersburg State University, Russia.

Violetta Koseska-Toszewa, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.

Natalia Kotsyba, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.

Antoni Mazurkiewicz, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Irina Nekipelova, Izhevsk State Technical University, Russia.

Karel Pala, Center for Natural Language Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic.

Rumyana Panova, Veliko Tărnovo University & Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Patrice Pognan, LALIC-CERTAL (Université de Paris-Sorbonne et INALCO), France.

Metalanguage and Encoding Scheme Design for Digital Lexicography

MONDILEX third open workshop
Bratislava, Slovakia, 15–16 April, 2009

Editor: Radovan Garabík
Technical editor: Marek Ivančík

Published by: Tribun EU s.r.o., Gorkého 41, 602 00 Brno, Czech Republic
1st edition
ISBN: 978-80-7399-745-8