COGNITIVE STUDIES | ÉTUDES COGNITIVES, 15: 353–367
Warsaw 2015

DOI: 10.11649/cs.2015.024

LUDMILA DIMITROVA[1,A] & **VIOLETTA KOSESKA-TOSZEWA**[2,B]

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

[A]ludmila@cc.bas.bg ; [B]amaz@inetia.pl

DIGITAL CORPORA AND THEIR APPLICATIONS IN SEMANTIC STUDIES AND LEXICOGRAPHY

Abstract

The paper describes the first Bulgarian-Polish digital resources — parallel and comparable corpora, and their applications in the semantic studies and lexicography for creation of Bulgarian-Polish digital dictionary, a significant part of these bilingual resources. Some examples show how valuable the links between the bilingual aligned corpus and the digital dictionary are. The first Bulgarian-Polish digital resources are the main result of collaborative work between the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI-BAS) and the Institute of Slavic Studies of the Polish Academy of Sciences (ISS-PAS), established for the first time in 2006.

Keywords: digital corpora; aligned corpus; bilingual corpus; digital dictionary; digital lexicographic resource; semantic study

1. Introduction

A parallel bilingual corpus differs fundamentally from a monolingual one because of the language material collected in it: the parallel texts have to be at the synchronous level and must reflect the current state of the two languages. Keeping in mind the richness and diversity of natural languages, we point out that the selection of texts in a parallel corpus is essential, especially for linguistic purposes. The advantage of processing a bilingual parallel corpus is to obtain context specific information about the syntactic and semantic structures and the usage of words in two languages. Thus, the Bulgarian-Polish parallel corpus is useful to linguists-researchers for various research purposes, for instance in contrastive and comparative studies of the Bulgarian and Polish languages. The first Bulgarian-Polish aligned corpus, a part of the Bulgarian-Polish parallel corpus, is the main tool for investigation and study of semantic properties of the some universal language categories: definite-

ness/indefiniteness, category of time and aspect, and modality in Bulgarian and Polish (Dimitrova & Koseska-Toszewa, 2014).

2. First Bulgarian-Polish Digital Resources

At the start of the collaborative project between IMI-BAS and ISS-PAS no bilingual Bulgarian-Polish digital resources existed. To realize the goals of the project language materials were gathered in order to create bilingual resources. The first Bulgarian-Polish digital resources include the first Bulgarian-Polish corpus and the language material for the first digital Bulgarian-Polish dictionaries. Both languages belong to the Slavic language family: Bulgarian belongs to the South-Slavic and Polish — to the West-Slavic language family and linguistic and contrastive studies of the two languages can be carried out based on bilingual digital resources (corpora and dictionaries). Furthermore, the Bulgarian-Polish aligned corpus serves as the main source of vocabulary for Bulgarian-Polish digital dictionaries.

2.1. International standards for development and applications of bilingual resources

In 1995, the international project Text Encoding Initiative (TEI) (Ide & Sperberg-McQueen, 1995), one of whose goals was to develop a guide for the preparation and exchange of texts in digital form for research purposes (Sperberg-McQueen & Burnard, 2002), proposed the usage of standards for text representation. The TEI group chose Standard Generalized Markup Language (SGML), a meta-language defined in 1986 with the international standard ISO 8879 for the applications to language engineering (Burnard, 1995). SGML and later XML (Extensible Markup Language) provide the multiple uses of marked texts for different types of processing, independent of the natural language. That is why the SGML/XML-annotated texts serve as multi-use language resources for various multilingual systems. The next step to the standardization is the preparation of a Corpus Encoding Standard (CES) (Ide, 1998) for different kinds of language resources. The annotation of the text data in accordance with international standards provides efficient exchange of digital language resources and language technology between researchers in linguistics, informatics, the humanities and social sciences.

2.2. Corpus annotation

Bi- and multilingual corpora are large digital repositories of natural language data with an important role in natural language processing. These valuable multilingual datasets are intended for language engineering research and development. They are widely applicable to contrastive studies in a multilingual context, as well as in education for the purpose of language learning or training of translators especially for the training of software tools for machine translation.

Corpus annotation is the process of adding linguistic or structural information to a text corpus. One common type of annotation is the addition of labels or tags that indicate the word class for the words in the text. This is the so called **part-of-speech tagging** (or POS tagging). Apart from POS tagging, there are other types of annotation, for example, **structural annotation**, which corresponds to different structural levels of a corpus or text. Written texts contain a number of different structural forms or divisions. Novels have a complex hierarchy and are

divided into parts and chapters, newspapers are divided into sections, reference works — into articles, etc. The most common division in this hierarchy is the paragraph.

2.3. First Bulgarian-Polish Corpus

What are the reasons for the development of the Bulgarian-Polish corpus?
The first Bulgarian-Polish corpus was developed due to the need for research material for contrastive studies in these two languages. The MTE-model for corpora, developed under the EC COP project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* (Dimitrova et al., 1998), is the formal model that is used in the development of Bulgarian-Polish corpus. The structure of the first Bulgarian-Polish corpus corresponds to the MTE-model for multilingual corpora: it contains two sub-corpora — **parallel and comparable** (Dimitrova & Koseska, 2009a).

Linguistic contents and numbers
The corpus contains texts (data) in the national languages: Bulgarian and Polish. It is constantly growing, since new data are being added, or are planned for addition (other texts are in various stages of completion).

At the moment the first Bulgarian-Polish parallel corpus contains digital data for approximately 10 million words. A small part comprises literary texts by Bulgarian and Polish authors — short stories, novels, children's literature, science fiction. These are original Bulgarian texts with Polish translations or *vice versa* and texts in other languages translated into both Bulgarian and Polish. Recently, a big part of available through the Internet texts of official documents of the European Union (EU) and European Commission (EC) were included as well. All texts in the corpus are texts published in and distributed over the Internet or were provided by the authors for research purposes only. The *Bulgarian-Polish parallel corpus*, depending of the content, includes two parallel sub-corpora: *core* and *translated*.

The *core Bulgarian-Polish corpus* consists of original texts in Polish — excerpts from novels, science fictions and short stories by Polish writers and their translation in Bulgarian; and original texts in Bulgarian — short stories by Bulgarian writers and their translation in Polish.

The *translated Bulgarian-Polish corpus* consists of texts in Bulgarian and in Polish of brochures of the EC, documents of the EU and the EU-Parliament, published on the Internet; Bulgarian and Polish translations of Antoine de Saint-Exupery's "The Little Prince"; Bulgarian and Polish translations of Michael Bulgakov's "Master and Margarita", Bulgarian and Polish translations of George Orwell's "1984".

Structural Annotation of Bulgarian-Polish Corpora
Some texts in the ongoing version of the Bulgarian-Polish parallel corpus are annotated at the paragraph level (manually or using ad-hoc tools), according to the standards of **TEI** and **CES** with paragraph level boundaries (Tab. 1). Other texts are annotated at the segment level (usually sentence level — Tab. 2). Such structural annotation allows texts in both languages (Bulgarian/Polish and vice versa) to be segmented and compared at the paragraph or at the segment level. The segmentation allows drawing of a broader context in both languages. This approach is

more correct — we are not comparing "word" with "word", we compare word-forms in a broader context (level paragraph, segment, or sentence), which allows us to obtain a more adequate meaning of the word. Some examples see Tab. 1–2.

Table 1: Annotation at the paragraph level (excerpts from K. Gyulemetov's "The Glorious Frog").

Bulgarian: Нямаше какво да отлага. Нарами торбичка с храна, пристегна маратонките си и потегли към Синия вир. Там държеше малка лодка, скрита в коренищата на върбите. Мислеше първо да посети приятелите, да поостане някой и друг ден, пък ако иска някой да го придружи в славния поход — добре дошло.	Polish: Pan Bo schował kąpielówki do swej torby podróżnej, włożył swoje ulubione szare spodnie i czerwona koszulę w paski, zawiązał adidasy i ruszył w kierunku Niebieskiego Wiru. Tam, schowana w korzeniach starej wierzby, czekała na niego mała łódka. Pan Bo miał zamiar odwiedzić najpierw najbliższych przyjaciół; może któryś z nich zdecyduje się towarzyszyć mu w podróży?
Bulgarian: Бо потърси лодката с поглед. Нямаше я никаква. Това не бива да ни учудва, както не учуди и Жабока. Той ставаше много нехаен, ако му се случеше такова приятно развлечение като днешното. Затова, без ни най-малко да се огорчи за изгубената лодка, тръгна към долните вирове.	Polish: Pan Bo poszukał wzrokiem łódki. Dookoła spokojnie falowała woda, ale łódki nigdzie nie było widać. Nie martwcie się tym, skoro nawet Pan Bo nie zmartwił się zbytnio. Po tylu przyjemnych chwilach żabka nie chciała popsuć sobie miłego nastroju. Ruszyła zatem drogą prowadzącą wzdłuż rzeki i zanuciła nową piosenkę.

Table 2: Annotation at the segment level (excerpts from documents of the EU).

Bulgarian: УПРАВИТЕЛНИЯТ СЪВЕТ НА ЕВРОПЕЙСКАТА ЦЕНТРАЛНА БАНКА , като взе предвид Устава на Европейската система на централните банки и на Европейската централна банка , и по - специално член 27 . 1 от него , като има предвид , че : ####	Polish: RADA PREZESÓW EUROPEJSKIEGO BANKU CENTRALNEGO , uwzględniając Statut Europejskiego Systemu Banków Centralnych i Europejskiego Banku Centralnego , w szczególności art . 27 ust . 1 , a także mając na uwadze , co następuje : ####
(1) Отчетите на Европейската централна банка (ЕЦБ) и на националните централни банки на Еуросисте-	(1) Sprawozdania finansowe Europejskiego Banku Centralnego (EBC) oraz krajowych banków centralnych Eurosys-

мата се проверяват от независими външни одитори , препоръчани от Управителния съвет на ЕЦБ и одобрени от Съвета на Европейския съюз . ####	temu podlegają badaniu prowadzonemu przez niezależnych zewnętrznych audytorów rekomendowanych przez Radę Prezesów EBC i zatwierdzanych przez Radę Unii Europejskiej . ####
(2) Съгласно член 37 , параграф 1 от Федералния закон за Oesterreichische Nationalbank , всяка година общото събрание на Oesterreichische Nationalbank (OeNB) избира двама одитори и двама заместник - одитори . ####	(2) Zgodnie z art . 37 ust . 1 ustawy federalnej o Oesterreichische Nationalbank , Walne Zgromadzenie Oesterreichische Nationalbank (OeNB) corocznie dokonuje wyboru dwóch audytorów oraz dwóch audytorów zastępczych . ####
Заместник-одиторите могат да бъдат упълномощени само в случай , че одиторите са възпрепятствани да извършат одита . ####	Audytorzy zastępczy mogą otrzymać mandat jedynie w przypadku niemożności wykonywania obowiązków przez audytorów . ####

2.4. Bulgarian-Polish Aligned Corpus

For a parallel corpus to be useful, it must be treated with a special program for "alignment". The term "alignment" refers to the process of connecting pairs of words, phrases, terms or sentences in texts from different languages that are translated equivalents of each other. "Alignment" is a form of annotation carried over parallel corpora to facilitate the construction and evaluation of translation models stored in memory and used in support of computer-assisted translation. Although many parallel corpora are manually "aligned" (or "annotated"), automatic "alignment" forms the core of parallel corpora processing and tool development for "alignment" with a high degree of accuracy.

The aligned Bulgarian-Polish corpus is developed at the Department of Mathematical Linguistics at IMI-BAS. Texts of the parallel Bulgarian-Polish corpus serve as input data. The texts of the aligned corpus are automatically annotated for language (Bulgarian or Polish) and sentence/segment boundaries. Two language-independent freely available programs are used to align Bulgarian-Polish parallel texts at the sentence level:

1. Memory Translation 2007, a computer aided tool TextAlign (`http://mt2007-cat.ru/index.html`),
2. Bitext Aligner/Converter (bitext2tmx aligner, available at `http://bitext2tmx.sourceforge.net/`).

These software packages have applications in computer-assisted translation. Both tools align bilingual texts without bilingual dictionaries, but still require human editing. The resulting aligned texts are similar. The aligned Bulgarian-Polish texts are manually checked for the correctness of bilingual links.

Table 3 shows excerpts from the aligned at the sentence level texts of the Bulgarian-Polish aligned corpus, using the tool TextAlign.

Table 3: Excerpts from sentence-aligned Orwell's "1984"

1-1 aligned sentences	<tuv xml:lang="bulgarian"> <seg>На всяка площадка от стената срещу шахтата на асансьора в него се взираше огромното лице от плаката.</seg></tuv> <tuv xml:lang="polish">" <seg>Na każdym piętrze, na wprost drzwi windy, spoglądał ze ściany plakat z ogromną twarzą. </seg></tuv>
5-1 aligned sentences	<tuv xml:lang="bulgarian"> <seg>Те приютяваха четирите министерства, между които бе поделен целият апарат на властта. Министерството на истината се занимаваше с информацията, забавленията, образованието и изкуствата. Министерството на мира се занимаваше с войната. Министерството на любовта поддържаше законността и реда. А Министерството на изобилието отговаряше за икономиката.</seg></tuv> <tuv xml:lang="polish">" <seg>Mieściły cztery ministerstwa, składające się na aparat rządowy: Ministerstwo Prawdy, któremu podlegała prasa, rozrywka, oświata i sztuka, Ministerstwo Pokoju, które zajmowało się prowadzeniem wojny, Ministerstwo Miłości, które pilnowało ładu i porządku, wreszcie Ministerstwo Obfitości, sprawujące pieczę nad gospodarką.</seg></tuv>
1-2 aligned sentences	<tuv xml:lang="bulgarian"> <seg>Уинстън тъкмо заемаше мястото си в средните редици, когато двамина, които познаваше по външност, но никога не бе разговарял с тях, неочаквано влязоха.</seg></tuv> <tuv xml:lang="polish"> <seg> Winston właśnie zajmował miejsce w jednym ze środkowych rzędów, gdy na salę weszły niespodziewanie dwie osoby. Znał je z widzenia, lecz nigdy nie zamienił z nimi słowa.</seg></tuv>

3. Applications of Digital Corpora

The aligned parallel corpora are useful for many natural language processing applications: in systems for machine-aided human translation, or for training of software for machine translation. They are a prerequisite for contrastive studies or other linguistics research, and can also be used for the retrieval of linguistic information, for producing concordances, etc.

3.1. Applications in Contrastive Studies

It's a well-known fact that an aligned corpus itself provides to researchers more language material than examples presented in theoretical studies and articles. The usage of a given word-form in a wide context, like a set of sentences from bilingual aligned texts, show specific features of this word-form, such as gender and number for nouns; tense, aspect, and mode for verbs, etc. The aligned Bulgarian-Polish

corpus is a good tool with wide applications in the contrasting of semantics problems in the two languages (Dimitrova & Koseska, 2012). The aligned corpus is annotated at the sentence-level and therefore represents the formal structure of the text, it is an appropriate tool for contrast problems, typical for the semantic structure of sentences in Bulgarian and Polish. The corpus gives us a possibility to contrast such semantic categories like different kinds of modality, the semantic category of time and especially the quantification of time. Some examples of temporal meanings (understood like in Koseska & Mazurkiewicz, 2010) and verbal forms used to express them, extracted from the Bulgarian-Polish aligned corpus, follow:

(1) **Polish praeteritum of perfective verbs // Bulgarian aorist form of perfective verbs** — represent the *unique quantification of an event*:

PL *Stewardessa poprowadziła mnie między rzędami foteli na sam przód.*

BG *Стюардесата ме поведе напред между редицата от кресла.*

PL *Zjechałem na dół, chyba kilka pięter, i wyszedłszy na ulicę dolnego poziomu zdziwiłem się, zobaczywszy znów nad sobą niebo.*

BG *Спуснах се надолу, може би няколко етажа, и когато излязох на улицата на долното ниво, учудих се, че отново виждам небе.*

(2) **Polish praeteritum of imperfective verbs // Bulgarian imperfect form of imperfective verbs** — represent the *unique quantification of a state*:

PL *Chciał jeszcze coś powiedzieć.*

BG *Той искаше да ми каже още нещо.*

PL *Nie mogłem go znaleźć i nawet szukać nie próbowałem.*

BG *Не можех да го намеря и дори не се опитвах да го търся.*

(3) **Polish praeteritum form of imperfective verbs // Bulgarian aorist form of imperfective verbs** — represent the *uniqueness of a set — unique quantification of states*:

PL *Głos z wewnątrz wypytywał nas, cośmy za jedni.*

BG *Един глас отвътре разпитва' що за хора сме.*

We emphasize that the aligned corpus enriches the theory with language material and corrects some theoretical setups, left unnoticed by scholars; it also demonstrates the important role of textual structural annotation. Special attention has been given to enabling further distribution of the corpora by encoding them in a standard format. The web-presented language resources are oriented both to human and machine users and are available for a wide area of applications. Every web-based corpus is a resource combining a number of features that together make it a unique and useful tool not only for language studies, but also for researchers of many fields. These features include: rich linguistic content, annotation (mark-up at two levels: paragraph and sentence, POS-tagging, etc.), search query (advanced possibilities for combining many search criteria), display of the search results in

an intuitive and simple interface, advanced results handling (concordances, collocations, etc.). The Bulgarian-Polish aligned corpus will soon be available on the Internet by the Web-based software tool with a wide spectrum of features for practical applications. The corpus will be freely available for research and education on the web with an appropriate trilingual interface in Bulgarian, Polish, and English. For more detailed description of the web-application software for the presentation of bilingual aligned corpora we refer to (Dimitrova & Dutsova, 2013a, 2013b; Dutsova 2013, 2014).

3.2. Concordances

One of the major developments in linguistic research has come from the possibility of studying vast amounts of text through software tools, namely through text retrieval and concordancing programs. The basic investigation procedure for querying text corpora consists in producing multiple concordance lines, for a specified string of characters — a word, a lemma or a phrase. The extracted citations can be sorted to reveal recurring clusters of words. The analysis of these recurring patterns highlights the behavior of actual language in context, and complements and sometimes challenges the information provided by standard reference tools such as dictionaries and grammars. This means that the opportunity exists — thanks to the broader context — to study more precisely the meanings of word-forms in each language. The Figure 1 shows a concordance with the Bulgarian word *смятам* [I *count*; II *consider*, *figure*, *guess*, *think*] from a part of the Bulgarian-Polish aligned corpus — Antoine de Saint-Exupery's "*The Little Prince*":

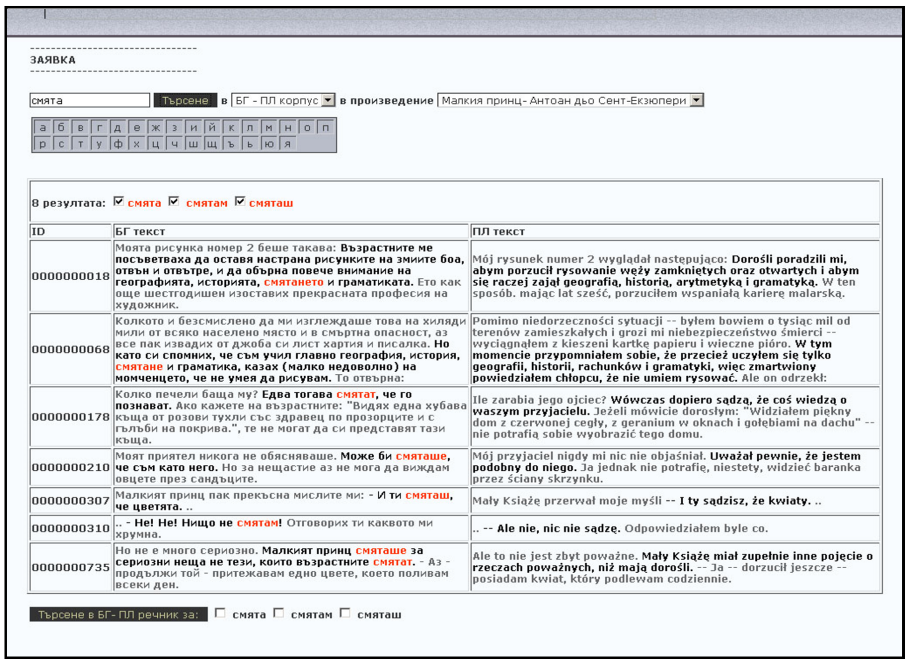

Figure 1: A concordance with the Bulgarian word *смятам*.

3.3. Development of multilingual lexical databases and digital dictionaries

Parallel and aligned corpora are the best resource for the development of bi- and multilingual lexical databases and different kinds of digital dictionaries.

Multilingual parallel corpora represent a good base of data for the creation of bilingual dictionaries. There are many research projects for automatic extraction of bilingual lexical knowledge from parallel corpora in the field of information retrieval from large scale text corpora. Thus parallel corpora are successfully used for automatic lexicon extraction.

There one could find and extract many examples of the usage of the words from a corpus in a wide context.

4. The First Bulgarian-Polish Digital Dictionary

Computer lexicography encompasses computer methods and resources for the automation of lexicographic activities. Such activities include: setting up of basic principles for development, creation and maintenance of dictionaries, recording of linguistic information in databases, creation of electronic indices, etc. Commonly, the dictionary is a list of words, so called headwords (or main words), arranged in alphabetical order, and their meanings. Depending of the aims of a usage of the dictionary, information is given about the pronunciation, grammar, derivative words, history or etymology of the main word, as well as recommendations for its usage, examples, phraseological expressions, illustrations. The classification of the dictionaries is based on multiple criteria, for example, according to:

- **the type of carrier** the dictionaries are *traditional dictionary* — this is developed with a human/computer, but in the final form, it reaches the user in a paper form; and *digital/electronic dictionary* — this exists in a digital format and can be referred to one of the following categories: online (web-based) or local (desktop) dictionary.
- **the number of languages** the dictionaries are *monolingual* — dictionary in which headwords are defined in the same language; *bilingual* — dictionary which contains texts in two languages, a source language in which the headwords are presented (and their characteristics are also described in this language) and a target (exactly one) language in which the words and their senses are translated; and *multilingual* — dictionary which contain translations of the headwords and their senses in more than one languages.

4.1. Advantages of the Digital Dictionaries

The printed (or paper) dictionary is a static collection of dictionary entries. The digital dictionary is a dynamic collection of dictionary entries which provides a dynamic structure of the dictionary entry per se. The dynamic structure gives **the basic advantages of the digital** vs paper dictionary:

- the collection of words can be continuously expanded because new dictionary entries can be added;
- the dictionary entry content can be enriched by addition of supplementary information about the headword (grammatical, etymological), of examples (for clarification of usage), of phrases and combinations, etc.;

- a relatively easy refinement of the system of classifiers, used for structuring the dictionary entry in order to describe optimally the headword;
- the digital content of the dictionary entries can with time serve the purposes of not only one, but multiple dictionaries, e.g. its usage for the creation of a new (or different type of), for example, digital dictionary of synonyms, antonyms, word-forming, etc., based on the main digital dictionary; or two monolingual digital dictionaries (explanatory or terminological) in two different languages can be used to produce a new bilingual dictionary (although in practice that is non-trivial);
- when necessary — last but not least — correction of various mistakes.

The bi- and multilingual digital dictionaries have more limitations and require even more so that the description of language specifications of the headword in each entry of the dictionary be simple and simultaneously more comprehensive. The fact that the lexical form in every language may have several meanings that do not overlap across the respective compared languages also has to be addressed: foe example, a word from the source language has more than one meaning, while in the target language different words correspond to the different meanings. For this reason, we propose the headword form in the dictionary entry of the digital dictionary to be indexed according to the number of meanings, and each different meaning to be related unambiguously to the form. In this manner most meanings of the form can be encompassed. Such a description might require more classifiers but it is obvious that the greater number of classifiers provides a more adequate translation correspondence (Dimitrova & Koseska-Toszewa, 2008).

4.2. Design of Structure and Content of the Dictionary Entry

Every dictionary entry is a structured object which uses different abbreviations and structural units in order to present succinctly the whole information about the headword and its specifications. The structure of dictionary entries varies a lot within the dictionary as well as between separate dictionaries. In spite of these variations some strict and constant structural rules exist so that the dictionaries can be understood by their readers. The external structure (presentation of text) does not completely determine the internal structure (information content in the dictionary database). There is a great diversity of hierarchical structures: in some dictionary entries the hierarchal organization of their structure may be deeply embedded, whereas in other cases some structural elements may be missing. The problems of representation of the specific features (lexical, syntactic or semantic) of the headword in a bilingual dictionary entry had been discussed in parallel with the development of the Bulgarian-Polish digital dictionary (Dimitrova & Koseska-Toszewa, 2014). The build-up of electronic dictionaries is a complex and strenuous process, associated with overcoming various difficulties (Dimitrova & Koseska, 2009b; Dimitrova, Koseska, Dutsova, Panova, 2009; Dimitrova et al., 2010; Koseska-Toszewa, 2009).

Why a digital dictionary was chosen?

The advantages of the digital bilingual dictionaries provided by the advanced IT determined our choice. In addition, the information collected, updated and stored

in digital archives (dictionary entries in our case), allows the production of two other dictionaries: an online (web-based) dictionary and a hard-copy dictionary (if printed on paper).

4.3. Formal model for the Bulgarian-Polish online dictionary encoding
The Bulgarian-Polish digital dictionary was developed in line with the design and development of a *bilingual lexical database* (LDB) for support of the first Bulgarian-Polish online dictionary (forthcoming at www.math.bas.bg). The structural units and content of the designed Bulgarian-Polish LDB should fully meet international standards so that the LDB and the digital dictionaries should be compatible with other TEI-conformant language resources. The formal model of the bilingual LDB follows the model for monolingual dictionary encoding of EU project CONCEDE *Consortium for Central European Dictionary Encoding* but some expansions were made (Erjavec, Evans, Ide, & Kilgarriff, 2003).

We started the development of a bilingual LDB having in mind the future development of the Bulgarian-Polish digital resources (Dimitrova, 2009, 2010). The structure of developed Bulgarian-Polish LDB is described in Dimitrova, Panova, and Dutsova (2009); Dimitrova, Dutsova, and Panova (2011a, 2011b). The brief description of the structural and content tags used in the Bulgarian-Polish LDB, including new content tags we suggested, follows.

The structural tags are three: **alt** (alternation, though generally for use in quite different contexts), **entry** (indicates the text that describes a dictionary entry), and **struc** (indicates a separate independent part in a dictionary entry).

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

The **hw** tag contains the headword and is used for alphabetization and indexing, and 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 **gram** tag contains *grammatical information* relating to a word other than *gender*, *number*, *case*, *person*, *tense*, *mood*, 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, like *countable*/*uncountable* uses of the nouns.

The suggested *new syntactic classifier for verbs* plays the significant role of *a filter* — it releases all verbs that can be "*transitive*". The syntactic classifier, which value is "*transitive*", means that Bulgarian transitive verbs are always followed by the direct object, but Polish transitive verbs are always followed either by the accusative case of adjectives, nouns and pronouns, or by a direct object. In the entry of a transitive Bulgarian verb such value is presented as <subc>**transitive**</subc>. The value(s) of such syntactic classifier of the corresponding Polish verb(s) is/are also presented by **subc** tag, but translations are divided in some groups (if it's necessary) according to attribute values (**transitive**/**intransitive**) and/or according to groups of synonyms, for example:

```
<hw>знай</hw>...<subc>transitive</subc>...

<struc type="Sense" n="1">
```

```
<trans>wiedzieć</trans><subc>transitive</subc>
</struc>

<struc type="Sense" n="2">
<trans>umieć</trans><subc>transitive</subc>
</struc>

<struc type="Sense" n="3">
<trans>znać</trans><subc>transitive</subc>
</struc>...
```

The **semantic tag**: a new tag is added to represent explicitly semantic information of semantic classifier about *state*/*event* of the verb. In the entry of a Bulgarian verb, indicated *state*, such value is presented as <semantic><orth>**state**</orth> <type>1</type></semantic>.

The **trans tag**, a new tag contains translation text and related information. Everything under trans relates to the target language: <trans>**znać**</trans><subc> **transitive**</subc>.

Hereby discussed classifiers in our digital dictionary differ from the classifiers in traditional dictionaries! We attach classifiers not only to the Bulgarian headwords (as a source language words) but also to their Polish translation equivalences. This description ensures the possibility to obtain, for example, automatically a Polish-Bulgarian entry from a Bulgarian-Polish entry, using the well-structured LDB and an extension aiming at the deeper study of linguistic problems in both languages.

The Bulgarian-Polish aligned corpus provided data for the selection of the Bulgarian lemmata that were included as headwords in the dictionary entries of Bulgarian-Polish digital dictionary (Dimitrova & Dutsova, 2012). The words distribution according to POS-classification follows the procedure for selecting of headwords included in the six monolingual lexical databases of the project CONCEDE. The main forms (lemmata) of the most frequent word forms in the corpus were selected. Lemmata were chosen for the relevant ten grammatical categories, identified in the Bulgarian-Polish aligned corpus, according to the frequency of their occurrence in corpus:

- open classes POS — noun, verb, adjective, adverb — no more than 90%,
- closed classes POS — numeral, pronoun, conjunction, preposition, particle, interjection — minimum 10% of the whole set of lemmata chosen.

In order to keep all different meanings we suggest the option where each meaning is shown with the same form but enumerated. In other words, the **homonymous entries** are indexed according to the number of meanings with I, II, etc., and appears in the list as many times as its different meanings:

I ми|н|а, -и *noun feminine*; kopalnia *noun feminine*; **каменовъ|глена ~а** kopalnia węgla kamiennego

II ми|н|а, -и *noun feminine military*; mina *noun feminine*

III ми|н|а, -еш *verb perfective*, *event*, *intransitive*; przejść *event*, *intransitive*; **ня|ма да ∼е** *colloq.* nic z tego nie będzie; **∼а ми през ума|** *figur.* przypomniałem sobie; **от ме|не да ∼е!** *figur.* niech ci będzie!

5. Conclusion

In conclusion, with the following examples we show how valuable are the links between the Bulgarian-Polish aligned corpus and the Bulgarian-Polish dictionary (Dimitrova &, Dutsova, 2013b; Dutsova, 2013, 2014):

напи|ш|а, -еш *verb perfective*, *event*, *transitive*; napisać *event*, *transitive*; **сло|жих лист на маши|ната и се опи|тах да ∼а телегра|ма до Варша|ва** wkręciłem papier w maszynę i próbowałem napisać depeszę do Warszawy [Bulgarian-Polish corpus]

I напо|мня|м, -ш *verb imperfective*, *state*, *intransitive*; przypominać *state*, *intransitive*; **Вътрешността| на у|лдера ∼ше доня|къде на експеримента|лната раке|та "Те|рмо Факс", коя|то ня|кога пилоти|рах...** Wnętrze uldera przypominało trochę eksperymentalną rakietę Termo-Fax, którą kiedyś prowadziłem...; ... **а ста|ртовите площа|дки върху| тръ|бните опо|ри напо|мняха на етаже|рки** ... a lądowiska, które wystawały z nich na tle nieba, wysunięte w powietrze na rurowych przęsłach, przypominały etażerki. [Bulgarian-Polish corpus]

Finally, we note that the Chapter VIII of the book "Semantics Properties of Selected Universal Language Categories in Digital Bilingual Resources" presents a dictionary of the most frequently used verbs in the Bulgarian-Polish aligned corpus (more than 2000 Bulgarian and Polish verbs) (Dimitrova & Koseska-Toszewa, 2014).

References

Burnard, L. (1995). The Text Encoding Initiative: An overview. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcript, mark-up and application* (pp. 69–81). New York: Longman.

CONCEDE. (n.d.). Retrieved from `http://www.itri.brighton.ac.uk/projects/concede/`

Dimitrova, L. & Koseska-Toszewa, V. (2014). *Semantics properties of selected universal language categories in digital bilingual resources.* Sofia: Demetra Ltd. Publisher.

Dimitrova, L., Koseska-Toszewa, V., Dutsova, R., & Panova, R. (2009). Bulgarian-Polish online dictionary — Design and development. In *Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June – 1 July 2009* (pp. 76–88). Warsaw: SOW.

Dimitrova, L. (2009). From electronic corpora to online dictionaries (on the example of Bulgarian Language Resources). In J. Levická & R. Garabík (Eds.), *Proceedings of the Fifth International Conference NLP, Corpus Linguistics, Corpus Based Grammar Research, Smolenice, Slovakia, 25–27 November 2009* (pp. 78–92). Brno: Tribun.

Dimitrova, L. (2010). Multilingual digital resources with Bulgarian language. *Cognitive Studies | Études cognitives*, *10*, 241–252.

Dimitrova, L. & Dutsova, R. (2012). Implementation of the Bulgarian-Polish online dictionary. *Cognitive Studies | Études cognitives*, *12*, 219–229.

Dimitrova, L. & Dutsova, R. (2013a). Web-application for the presentation of bilingual corpora (Focusing on Bulgarian as one of the two paired languages). *Cognitive Studies | Études cognitives*, *13*, 183–193. http://doi.org/10.11649/cs.2013.012

Dimitrova, L. & Dutsova, R. (2013b). A software package for processing Bulgarian digital resources: Parallel corpora and a bilingual dictionary. In *Proceedings of the Seventh International Conference NLP, Corpus Linguistics, E-Learning SLOVKO'2013, 13–15 November 2011, Bratislava, Slovakia* (pp. 40–50). Lüdenscheid: RAM-Verlag.

Dimitrova, L. & Koseska, V. (2009a). Bulgarian-Polish Corpus. *Cognitive Studies | Études cognitives*, *9*, 133–141.

Dimitrova, L. & Koseska, V. (2009b). Classifiers and digital dictionaries. *Cognitive Studies | Études cognitives*, *9*, 117–131.

Dimitrova, L. & Koseska, V. (2012). Bulgarian-Polish parallel digital corpus and quantification of time. *Cognitive Studies | Études cognitives*, *12*, 199–207.

Dimitrova, L. & Koseska-Toszewa, V. (2008). Some problems in multilingual digital dictionaries. *Cognitive Studies | Études Cognitives*, *8*, 237–254.

Dimitrova, L., Dutsova, R., & Panova, R. (2011a). Information technologies for the preservation of language heritage. In *Proc. of the International Conference Digital Presentation and Preservation of Cultural and Scientific Heritage DiPP 2011, 11–14 September 2011, Veliko Tarnovo, Bulgaria* (pp. 140–150).

Dimitrova, L., Dutsova, R., & Panova, R. (2011b). Survey on current state of Bulgarian-Polish online dictionary. In *Proceedings of the International Workshop "Language Technology for Digital Humanities and Cultural Heritage" within International Conference RANLP'2011, 16 September 2011, Hissar, Bulgaria* (pp. 43–50). Shoumen: INCOMA, Association for Computational Linguistics. http://aclweb.org/anthology-new/W/W11/W11-41.pdf

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H. J., Petkevic, V., & Tufis, D. (1998). Multext-East: Parallel and comparable corpora and lexicons for six Central and Eastern European languages. In *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada (Vol. 1, pp. 315–319). Stroudsburg, PA: Association for Computational Linguistics. http://doi.org/10.3115/980845.980897

Dimitrova, L., Koseska, V., Garabík, R., Erjavec, T., Iomdin, L., & Shyrokov, V. (2010). MONDILEX — Towards the research infrastructure for digital resources in Slavic lexicography. *Cognitive Studies | Études cognitives*, *10*, 147–162.

Dimitrova, L., Panova, R., & Dutsova, R. (2009). Lexical database of the experimental Bulgarian-Polish online dictionary. In R. Garabík (Ed.), *Metalanguage and encoding scheme design for digital lexicography: Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009* (pp. 36–47). Bratislava: L'. Štúr Institute of Linguistic, Slovak Academy of Sciences

Dutsova, R. (2013). Web-application for presentation of Bulgarian language heritage: Bilingual digital corpora and dictionaries. In *Proc. of the International Conference Digital Presentation and Preservation of Cultural and Scientific Heritage DiPP'2013, 18–21 September 2013, Veliko Tarnovo, Bulgaria* (pp. 99–108).

Dutsova, R. (2014). Web-based software system for processing bilingual digital resources. *Cognitive Studies | Études cognitives*, *14*, 33–43. http://doi.org/10.11649/cs.2014.004

Erjavec, T., Evans, R., Ide, N., & Kilgarriff, A. (2003). From machine readable dictionaries to lexical databases: The concede experience. In *Proceedings of the 7th*

International Conference on Computational Lexicography, COMPLEX'03, Budapest, Hungary, 2003.

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

Ide, N. M. & Sperberg-McQueen, C. M. (1995). The TEI: History, goals, and future. *Computers and the Humanities*, *29*(1), 5–15. http://doi.org/10.1007/BF01830313

Koseska, V. & Mazurkiewicz, A. (2010). *Time flow and tenses.* Warsaw: SOW.

Koseska-Toszewa, V. (2009). Form, its meaning, and dictionary entries. In *Metalanguage and encoding scheme design for digital lexicography: Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009* (pp. 105–111). Bratislava: L'. Štúr Institute of Linguistic, Slovak Academy of Sciences.

MULTEXT-East Home Page [MTE]. (n.d.). Retrieved 1 October 2015, from http://nl.ijs.si/ME

Sperberg-McQueen, C. M. & Burnard, L. (Eds.). (2002). *TEI P4: Guidelines for electronic text encoding and interchange.* Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen. http://www.tei-c.org/P4X/

Acknowledgment

This work was supported equally by the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences and a core funding for statutory activities form the Polish Ministry of Science and Higher Education.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study, data analyses, the writing: Ludmila Dimitrova, Violetta Koseska-Toszewa.

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation