ÉTUDES COGNITIVES, 8
SOW, Warszawa 2008

LUDMILA DIMITROVA
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Sofia

VIOLETTA KOSESKA-TOSZEWA
Institute of Slavic Studies
Polish Academy of Sciences
Warsaw

SOME PROBLEMS IN MULTILINGUAL DIGITAL DICTIONARIES

Abstract

The article discusses some observations from the joint work of Polish and Bulgarian research groups on the digital Bulgarian-Polish and Polish-Ukrainian dictionaries, as well as the projected multilingual (initially: Bulgarian-Polish-Ukrainian) dictionary. The researchers are currently working on a parallel corpus containing texts in Bulgarian and Polish, distributed over the Internet, whereby the translation correspondence is one-to-one. They are developing a comparable corpus that includes texts in Bulgarian and Polish (excerpts from newspapers, literary works, Internet textual documents) with the text sizes being comparable across the two languages. The two corpora, parallel and comparable, form the first Bulgarian-Polish corpus, that will be prepared in CES format, manually or using ad-hoc tools, and will be annotated on "paragraph" and "sentence" levels, according to the text annotation international standards. This bilingual corpus will provide a sample of the vocabulary to be included in an initial experimental version of the Bulgarian-Polish digital dictionary. 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. Great difficulties have to be addressed in order for a dictionary to satisfy the needs of a translator, a language researcher or an everyday user.

Introduction

This article discusses some observations from the joint work of Polish and Bulgarian research groups on the digital (in electronic form) Bulgarian-Polish and Polish-Ukrainian dictionaries, as well as the projected multilingual (initially: Bulgarian-Polish-Ukrainian) dictionary. The Bulgarian and Polish researchers are currently working on a parallel corpus (containing literary works and texts of documents in Bulgarian and Polish in digital form, with a one-to-one translation correspondence), and in addition developing a comparable corpus that includes texts in Bulgarian and Polish (excerpts from newspapers, literary works, Internet textual documents, with the text sizes being comparable across the two languages). These two corpora, parallel and comparable, form the first Bulgarian-Polish corpus, that will be annotated according to the digital language resource annotation standards. The bilingual Bulgarian-Polish corpus will provide a sample of the vocabulary, which is to be included in an initial experimental version of the Bulgarian-Polish dictionary.

Languages selection

The three languages, Bulgarian, Polish and Ukrainian, have been chosen for the following reasons: 1) there are no digital dictionaries for these languages, 2) there are no parallel corpora for these languages, and 3) each language represents one of the three Slavic language families, Bulgarian belongs to the South-Slavic, Polish to the West-Slavic, and Ukrainian to the East-Slavic language family. The differences between the three families, such as some phonetic systemic features, could be presented via algorithms.

Bulgarian, as well as the other South-Slavic languages employs in its vocabulary the phonetic groups ***–трат–***, ***–тлат–***, for example, in the Bulgarian nouns ***град***, ***градина***, ***брада***, ***крава***, ***глава***. The exact Polish correspondences of these phonetic groups are ***–trot–***, ***–tlot–***, later transformed into ***–trót–***, ***–tłót–***. For example, ***–trot–***, ***–tlot–*** **in the nouns *broda, krowa, głowa;*** ***–trót–***, ***–tłót–*** in the town names ***Wyszegród, Rajgród***, also in the nouns ***ogród, główny***. The Ukrainian correspondences of these phonetic groups are – ***торот–***, ***–толот–***, and are present, for example, in the nouns: ***город, борода, корова, голова***. The vocabulary of the three languages emphasizes the characteristic vicinity within the Slavic language family. Nevertheless, similar lexical forms of similar construction have not only close, but also completely different meanings. The Bulgarian word ***глава*** (*head* in English) corresponds to the Polish word ***głowa*** and the Ukrainian ***голова*** only in the sense "***head, body part***". The meaning of the Bulgarian word ***глава*** in the sense "***book chapter***" is missing in Polish and Ukrainian, but exists in Russian in the phonetic variant ***глава книги*** (not ***голова***), which shows that the meaning has been borrowed from Old Bulgarian („***глава на книга***", **"head of a book"**). The proof of this fact is the South-Slavic phonetic group –***тлат***– in ***глава***. (Stieber, 1973). If the meaning of ***глава*** (as book chapter) were not borrowed from Old Bulgarian, the word would countain the characteristic East-Slavic phonetic group –***толот***–. As is well-known (Stieber, 1966), the prototype –***tolt***– from Old Slavic becomes –***tlot***– in the West-Slavic family, –***толот***– in the East-Slavic family, and –***тлат***– in the South-Slavic family.

Semantic and contrastive studies of the Polish language

The Polish research group works on several projects:

- Semantic Bulgarian-Polish contrastive grammar;
- Parallel corpora;
- Digital dictionaries.

1. Semantic Bulgarian-Polish contrastive grammar

We should underline the fact that the study and development of a Bulgarian-Polish contrastive grammar (BPCG) started in the 1990s and the Polish team has an advantage in this area. The work on BPCG has been completed. The Grammar is a complete publication of monographs in nine volumes (the sixth one having four equal parts); the first four volumes of the publication of the BPCG are in Bulgarian, the remaining in Polish. The main problem in BPCG was the assessment of characteristics of the contrastive inventory of the two languages and their model of specification. Most often the comparison and contrast of two languages follow a direction from form to meaning, so that the contrastive study is limited to a description of the initial language's formal means. As a result the description is not only incomplete, but untrue. Many of the means for expressing a specific content in either language are not registered. For this reason the traditional grammars note that a given language does not possess a given lexical feature. It was considered for instance that Polish does not have a renarrative modal category, because the language does not possess the morphological means to express such modality. Polish, however, expresses this modality by lexical and morphological means, for example:

Тя била добра жена. — Ona była podobno dobrą kobietą.
Той бил добър лекар. — Miał to być dobry lekarz.

In this case, the speaker is not certain whether what he/she is saying is right or not (which does not imply that he/she has not been a witness of what is being talked about). This doubt (uncertainty) can be expressed in English with the following constructions: it is said, there is some talk that, it is possible, etc. (Translations in English as follows. First sentence: *There is some talk that she is a good woman. / She seems to be a good woman. / She is considered to be a good woman.* Second sentence: *There is some talk that he is a good doctor. / He seems to be a good doctor. / He is considered to be a good doctor.*)

Similar problems posed the question about the creation of a semantic interlanguage, which is initial for the description of the contrasted languages. That was a very difficult task, but it has been completed. The Bulgarian-Polish contrastive grammar, as to be seen from the essence of the problem, is the first and so far only detailed publication to present a semantic comparison with a permanently evolving interlanguage.

Restricting the description of the compared languages to semantic structure level of the sentence allowed the modelling of linguistic phenomena to be based on the contemporary theory of processes, a theory, called *Petri net* (Petri 1962). The contemporary theory of processes turned out to be very convenient for modelling phenomena in contrastive linguistics. The Petri net theory allowed the compared languages' description to be limited to the semantic sentence structure and to the language situation, which through a Petri net description can be considered as a history reflected in the net (see Mazurkiewicz 1986, Koseska, Mazurkiewicz 1988, 1994, 2004, Koseska 2006). The Petri net description of the tense and the modal phenomena in a natural language are related to the so-called Russell's direct reference semantics (Russell 1967), and to situational semantics. The situational semantics treats the semantic structure of the sentence as a set of abstract situations (see Barwise, Perry 1983), Cooper 1996). The Petri net theory was introduced in the language studies by A. Mazurkiewicz (Mazurkiewicz 1986). This theory is a language-independent and language-indifferent tool, set up on just three basic terms: state, event and implication relation. That is why these terms can form the basis of a semantic interlanguage for the description of two or more contrastive languages. The Petri net theory and quantification model were used to describe the semantic category definiteness/indefiniteness in the BPCG (Koseska, Gargov 1990).

The study method from meaning to form allowed us to discover many unexplored and so far not described linguistic facts in both Polish and Bulgarian. For example, for the first time a full list of the semantic modal categories such as imperceptivity, conditionality, and irrealis, and of other semantic categories such as tense, aspect, definiteness, quantity, communicant for Polish was presented. We must emphasize once more that the contrastive studies so far have proceeded in the direction from form to meaning. A single form in any natural language can be polysemantic, while the same semantic category can be implemented by different lexical forms. The implementation of a description in the direction from content to form is more difficult, because it is necessary to define precisely the content that is described in the contrasted languages. For example, we must point out that by tense we mean state, event and their mutual relations that happen before, during and after the statement of expression. Also we have to specify the meaning of definiteness, which is a uniqueness of an element or a set, which fulfil a given predicate (Koseska, Gargov 1990).

2. Synthesis of the Semantic Bulgarian-Polish contrastive grammar

At the current work stage it turns out that the Bulgarian-Polish contrastive grammar needed Synthesis for the Polish reader (like "Polish-Bulgarian contrastive grammar"). This Synthesis is not a summary of the different problems, studied in the volumes of the BPCG. The Synthesis is fully based on the semantic-logical sentence structure, and on the order it determines.

The traditional contrastive grammars present the description of the compared languages from form to form, with an *initial* and a *target* language, which does not lead to a parallel and equivalent description of the languages. The BPCG for the first time uses a semantic interlanguage, which provides an equivalent description of the two compared languages. At first the modal categories are described as the outermost "modal operator" in the semantic sentence structure, and are followed by the description of the tense, aspect, the meaning of the quantor expressions, and last, the predicate-argument positions, corresponding to selected semantic cases.

The work on Synthesis of the BPCG has been completed and will be published by the end of 2007 (Koseska, Korytkowska, Roszko 2007).

3. Corpora and dictionaries in Polish

The Polish research group from ISS-PAS is currently working on three parallel corpora, one of which will be in collaboration with the Bulgarian researchers.

The Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences (IPI PAN for short) has been developing for several years the *IPI PAN Corpus*, currently consisting of annotated texts exclusively by Polish authors: contemporary and older literary works, science, newspapers, parliamentary proceedings, etc. (Przepiórkowski 2006).

The Department of Semantics of ISS-PAS is however developing parallel corpora, specifically designed for bilingual and multilingual digital dictionaries. The dictionaries are a scientific novelty for the Polish team. The studies will be conducted in four directions that aim at the development of

1. Digital Dictionary of Concepts describing main semantic categories in the natural language;
2. Polish-Ukrainian digital dictionary - in collaboration with an Ukrainian research group from the ULIF-NANU;
3. Bulgarian-Polish digital dictionary - in collaboration with the Bulgarian research group;
4. Polish-Lithuanian digital dictionary.

The *Digital* Dictionary of Concepts (DDC), describing main semantic categories in the natural language, plays the role of an interlanguage in the contrastive studies of several languages: Bulgarian, Polish, Ukrainian and Lithuanian.

The first requirement for the DDC was that those concepts, which are its “entries”, be related to mutually non-contradictory theories. For instance, when we determine the concepts, describing the semantic category *definiteness/indefiniteness*, we could take advantage of P. Strawson’s Reference theory or of B. Russell’s Description theory and logical quantification. The simultaneous use of both theories, which is most common in the linguistic studies of the category of interest, leads to internal contradictions and the lack of discrimination of definiteness from egocentricity (Russell 1948, 1967, 1970), (Strawson 1967). DDC comprises the following semantic categories: [1] modality: imperceptive, hypothetical, irrealis, imperative, interrogative; [2] tense; [3] aspect; [4] definiteness/indefiniteness; [5] quantity; [6] degree; [7] communicant; [8] case, presented by selected types of predicate-argument sentence structure.

A *Polish-Ukrainian digital dictionary* is currently being created in collaboration with a Ukrainian research group from the ULIF-NANU. The principal source for developing the dictionary is a frequency list based on the *IPI PAN Corpus*. 30 000 most frequently used words were selected from the *IPI PAN Corpus* as dictionary entries for the public pilot version of the dictionary. Please, refer to `http://www.ulif.org.ua`, `http://www.ispan.waw.pl/`.

Corpora and dictionaries in Bulgarian

1. International standards for the applications to language engineering

In 1995 the international project Text Encoding Initiative (TEI) (Ide et al. 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 and Burnard 1994), proposed the usage of standards for text presentation. The TEI group chose Standard Generalized Markup Language (SGML), a metalanguage 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 annotation of the lexical data in accordance with international standards is usually based on morpho-syntactic descriptions. Thus it provides efficient exchange of digital language resources and language technology between researchers in linguistics, informatics, the humanities and social sciences.

2. Annotated digital resources in Bulgarian language

The Bulgarian team is well experienced in creating annotated corpora, monolingual digital dictionaries and lexical databases (LDBs). Research and development in this field started in 1995, when the Department of Mathematical Linguistics at the IMI-BAS began collaborating on behalf of the Bulgarian side in the multilingual EC-project, **MULTEXT-East**: *Multilingual Text Tools and Corpora for Central and Eastern European Languages.* After the successful completion of the work the Department participated with great success in another multilingual EC-project, **CONCEDE**: *Consortium for Central European Dictionary Encoding.* Both projects developed lexial resources, corpora, lexicons, lexical databases for six East-European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, and English being a “hub” language.

In the framework of MULTEXT-East project the Department of Mathematical Linguistics at IMI-BAS developed the standards for Bulgarian language resources, (Dimitrova 1998), the Bulgarian MULTEXT-East corpus, (Dimitrova et al. 1998). The Bulgarian-specific language resources that are developed for the MULTEXT-East project are segmentation and mapping rules, lists of special tokens/symbols, morphological rules and a dictionary or a lexical list. The rules describe the assignment of sentence boundaries, word splitting (cliticised form decomposition), word compounding, quotations, numbers, dates, punctuation, abbreviations, capitalization, etc. The lists of special tokens contain the most frequent abbreviations and names, titles, patterns for proper names, surnames, etc. together with their types.

The morphological rules and the morphological list provide information for the morphological analyzer: for complete processing of the inflection and minimal derivation. Each lemma in the lexical list, containing at least 17,000 lemmas, is associated with its part(s) of speech and morpho-syntactic specifications, which form its morpho-syntactic description (MSD). Each entry of the lexical list includes the triple < **inflected-form** (word-form), **lemma, MSD** >. Whenever the word-form is the very lemma, then the equal sign is written in the lemma-field of the entry ('=').

Bulgarian digital texts are collected not only in IMI-BAN, but in other institutes off BAS as well, for instance Institute for Parallel Processing and Institute for Bulgarian Language. However, the above-mentioned Bulgarian-Polish corpus is created for the first time ever.

Special features of working on digital dictionaries

1. Special features of working on bilingual and multilingual digital dictionaries

We are fully aware of the great difficulties to be addressed, in order for a dictionary to satisfy the needs of a translator, a language researcher or an everyday user. The dictionary has to comprise, simply put, "whatever the Pole is in regular contact with". According to A. Bogusławski, speaking about the creation of a new Polish dictionary, a dictionary "cannot include randomly appearing and short-lived words, but cannot overlook and leave empty spots where new words have appeared and remained in the language" (Bogusławski 1988).

The bi- and multilingual dictionaries have more limitations and require even more so that the description of lexical specifications of the headword in each entry of the dictionary be simple and simultaneously more comprehensive. The digital form of the dictionary requires the word-forms of the languages compared to be bilaterally classified, not unilaterally, only according to the source language, as is with standard bilingual dictionaries. The digital dictionaries are created with more care and work, but have the advantage that they can be continuously corrected, and with time can serve the purposes of not only one, but many dictionaries (e.g. of synonyms, antonyms, word-forming, etc, based on the main digital dictionary).

2. About the bilingual dictionaries

A bilingual dictionary usually means an alphabetical list of words or phrases from a source language **A**, and their translated correspondences from a target language **B**. The greatest possible difficulty during the creation of any bilingual dictionary is that a word or a lexeme from language **A** has more than one meaning, while in language **B** different words correspond to the different meanings.

For example:

(1) The *Bulgarian word* ***бал*** (***ball***/***grade***) has several meanings:

> **бал**[1] *м.* Тържествена танцувална вечеринка, забава. *Абитуриентски бал. Маскен бал. Бал на журналистите.* (*ball* in English) [фр.]
> **бал**[2] *м.* 1. Цифрова оценка на успех в училище, постижение в спорта и др. *Кандидатите са подредени по общия бал от изпитите. Балът му е нисък за влизане в университета.* (*examination mark*, *grade* in English) 2. *спец.* Единица за оценка по определена скала на силата, интензивността на някакво явление. *Силата на вятъра е шест бала.* (unit of measurement of intensity, e.g. *wind force*) [фр.]

In Polish different words correspond to these different meanings, depending on the context: *bal* (bal maskowy – *ball* in English), *bal* (towaru – *pack* in English), *stopień* (stopień równy – *degree of comparison* in English, cyfrowa ocena – *drade, rating, evaluation* in English).

(2) The *Bulgarian word* ***мир*** (**peace**) has several meanings:

> **мир**[1]*м.* 1. Покой, спокойствие, тишина. 2. Липса на вражди; сговор, съгласие, разбирателство. 3. Липса на война.
> **мир**[2] *м. старин. поет.* Свят.

In Polish different words correspond to these different meanings, depending on the context: *spokój*; *pokój*; *zgoda*; *świat*.

3. Features of the printed dictionaries available to us

We would like to note that in the past 10-15 years neither Bulgarian-Polish nor Polish-Bulgarian dictionaries have been published. The market in both countries is saturated with English-Polish and English-Bulgarian dictionaries, resulting in an absurd situation, where two languages belonging to the same language family, communicate via a language from another language family.

The wide spreading information technologies in many areas of science, industry, and lifestyle started a new era in the development of modern lexicography, in particular, the creation of monolingual, bi- and multilingual digital dictionaries. The digital dictionaries have a very useful property, allowing us to check the correctness of a translation. We can look up the appropriate meaning of word, translated using a **A-B** bilingual dictionary by looking it up in a **B-A** bilingual dictionary.

One of the sources for a Polish-Bulgarian dictionary could be the printed dictionaries. However, we must take a critical stand towards them. The first printed Polish-Bulgarian dictionary has been prepared by Ivan Lekov. The second one (Lekov and Sławski, 1961) has been published 47 years ago – in 1961. Both aforementioned dictionaries are of bibliographic rarity. Two printed dictionaries are available for our purposes: Bulgarian-Polish dictionary by Franciszek Sławski (Sławski 1987) and Polish-Bulgarian dictionary by Sabina Radeva (Radeva 1988). Their volume is approximately 60 000 words, and they are more or less equivalent in terms of lexical content. Both dictionaries do nonetheless have several disadvantages. First, they are somewhat outdated, as they were published 20 years ago. Second, they do not always contain the translated correspondences, i.e. sometimes instead of the Polish word, there is a definition, which interprets the meaning of the Bulgarian word. Both dictionaries contain also many outdated words and expressions, which are no longer used, e.g. dialects or loanwords from Turkish, for example:

погълчавам, -ш *vi.* v. **погълча**
подвъргвам, -ш *vi. arch.* poddawać kogoś np. mękom, torturom
подвъргна, -еш *vp.* v. **подвъргвам**
пущ, -ове *m pot.* lump *m*, rozpustnik *m*
спрасен, -на, -но *adi. lud.* v. **прасен**
спреварвам, -ш *vi. lud.* prześcigać
спреваря, -иш *vp.* v. **спреварвам**
спрепвам се, -ш *vi.* v. **спрепна се**
спрепна се, -еш *vp. lud.* potknąć się
фудул *adi. indecl. pot.* nadęty, pyszałkowaty
фудулин, **-и** *m pot.* nadęty, pyszałkowaty człowiek
фудуллук *m* pot. nadętość *f*, zarozumiałość *f*
фуквам, -ш *vi.* v. **фукна**
фукна, -еш *vp. lud.* puścić się biegiem, fuknąć (przestań.)
фурда *f lud.* resztki *pl*, pozostałości *pl*, odpadki *pl*; lichy, wybrakowany towar

There are also many synthetic words, which could formally belong to a certain word-formation group, for example: *по*гостя, *по*гощавам; *поза*бързам, *под*взема, *под*вземам; *под*големея се, *под*големя се, *под*големявам се; *под*гордея се, *под*гордявам се, and others.

While in Bulgarian it is allowed to add prefixes to create new words, these words may not have been in use 20 years ago. That is why we would use the above-mentioned dictionaries not as a primary source, but as reference.

From an digital corpus to an digital dictionary

The starting point of our collaborative investigation is the first Bulgarian-Polish digital corpus, developed in the framework of the cooperation between the Polish and the Bulgarian Academies of Sciences – the project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary".

1. Annotated Bulgarian–Polish corpus

The Bulgarian–Polish corpus consists of two parts: a parallel and a comparable corpus. All texts in the corpus are texts published in and distributed over the Internet. The parallel corpus contains literary texts and texts of documents in both languages, whereby the translation correspondence is one-to-one. The Bulgarian–Polish corpus will be annotated according to the encoding schemes for Bulgarian language, developed in the EU-project MULTEXT-East. The entire Bulgarian–Polish corpus will be prepared in CES format, manually or using ad-hoc tools, and will be annotated for sentence (<s>, </s>) and paragraph (<p>, </p>) boundaries. The first Bulgarian–Polish corpus will be the chief source of vocabulary for the Bulgarian–Polish digital dictionary.

The Bulgarian-Polish parallel corpus includes two parallel sub-corpora:

(1) a *pure* Bulgarian-Polish corpus
(2) a *translated* Bulgarian-Polish corpus.

The *pure* Bulgarian-Polish corpus consists of

(1) original texts in Polish - excerpts of some novels by Henryk Sienkiewicz, excerpts of some fiction novels by Stanisław Lem and other Polish writers and their translation in Bulgarian,
(2) 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 in Internet; Bulgarian and Polish translations of Antoine de Saint-Exupery's "The Little Prince"; Bulgarian and Polish translations of Karl May's "Winnetou".

The Bulgarian-Polish comparable corpus includes texts in Bulgarian and Polish: excerpts from newspapers and textual documents, shown in internet, excerpts from several original fiction, novels or short stories, with the text sizes being comparable across the two languages. Some of the Bulgarian texts in the comparable Bulgarian-Polish corpus are annotated on "paragraph" and "sentence" levels, according to the text annotation international standards.

Some examples follow:

Literary text:

<**Text Pl**-1.1.1> ...
<p> Na razie rozmowa urwała się, albowiem uwagę Stasia zwróciły ptaki lecące od strony Echtum om Farag ku jezioru Menzaleh. Leciały one dość nisko i w przezroczystym powietrzu widać było wyraźnie kilka pelikanów z zagiętymi na grzbiety szyjami, poruszających z wolna ogromnymi skrzydłami. Staś począł zaraz naśladować ich lot, więc zadarł głowę i biegł kilkanaście kroków groblą, machając rozłożonymi rękoma. </p>
... </**Text Pl**-1.1.1>

<**Text Bg**-1.1.1> ...
<p> Тук разговорът се прекъсна, защото вниманието на Стас беше привлечено от птиците, които летяха откъм Ехтум ом Фараг към езерото Мензалех. Те летяха твърде ниско и в прозрачния въздух ясно се виждаха няколко пеликана с извити върху гърбовете шии, размахващи бавно огромните си криле. Стас веднага започна да подражава на полета им, навири глава и пробяга петнайсетина метра по насипа, като размахваше разперените си ръце. </p>
... </**Text Bg**-1.1.1>

Texts of documents:

<**Text Bg**-2.8.15> ...
<p> Европейският съвет дава на Съюза необходимия тласък за неговото развитие и определя неговите общи политически насоки.</p>
<p> Европейският съвет включва държавните глави или правителствените ръководители на държавите-членки и председателя на Комисията. Те се подпомагат от министрите на външните работи на държавите-членки и от член на Комисията. Европейският съвет заседава най-малко два пъти годишно под председателството на държавния глава или правителствения ръководител на държавата-членка, която председателства на Съвета. <p>
<p> Европейският съвет предоставя на Европейския парламент доклад след всяко от своите заседания, както и ежегоден писмен доклад относно постигнатия от Съюза напредък. <p>
... </**Text Bg**-2.8.15>

<**Text Pl**-2.8.15> ...
<p>Rada Europejska nadaje Unii impulsy niezbędne do jej rozwoju i określa jego ogólne kierunki polityczne.</p>
<p>W skład Rady Europejskiej wchodzą szefowie państw lub rządów

Państw Członkowskich oraz przewodniczący Komisji. Towarzyszą im ministrowie spraw zagranicznych Państw Członkowskich i członek Komisji. Rada Europejska zbiera się co najmniej dwa razy w roku pod przewodnictwem szefa państwa lub rządu Państwa Członkowskiego, które przewodniczy Radzie. <p>
<p>Rada Europejska składa Parlamentowi Europejskiemu sprawozdanie po każdym swym spotkaniu oraz roczne sprawozdanie pisemne o postępach dokonanych przez Unię. <p>
... </**Text Pl**-2.8.15>

2. Experimental bilingual digital Bulgarian-Polish dictionary

The Bulgarian-Polish corpus will provide a sample of the vocabulary to be included in an initial experimental version of the Bulgarian-Polish digital dictionary. For that purpose we shall employ the selection methodology (Tufis et al. 1999) of the most frequently used words in the MULTEXT-East corpus, which was used to select the dictionary entries for the multilingual LDB (six East-European languages and English as a "hub" language). This LDB was developed in the CONCEDE project (Dimitrova et al. 2002). Initially we plan to choose 5000 to 6000 words with the highest frequency in the corpus. In terms of parts of speech, about 2000 should be nouns, 1800 verbs, 1000 adjectives, 500 adverbs, 100 pronouns, 50-60 numerals, conjunctions, prepositions, interjections, abbreviations. Thus we will have a representative sample of the most frequent words in the language that will serve as the core of the experimental bilingual digital Bulgarian-Polish dictionary.

Some examples of translation of words by means of on-line bilingual dictionaries follow:

(1) English-Francais on-line dictionary,
`http://www.wordreference.com/`

peace

Principal Translations:
peace *n* (not war) paix *nf*
After the war ended, there were 30 years of peace before the next war.
Après la fin de la guerre, il y a eu 30 ans de paix avant la guerre suivante.
peace *n* (quiet, calm) tranquillité nf
She went to her room for some peace and quiet.
Elle est allée dans sa chambre pour avoir de la tranquillité et du calme.

Additional Translations:
peace *n* (mental calm) paix nf
After years of depression, he is finally at peace with himself.

Après des années de dépression, il est enfin en paix avec lui-même.

peace *n* (peace treaty) paix nf

The two warring countries made peace after 3 years of war.

Les deux pays ennemis ont fait la paix après trois ans de guerre.

peace *n* (absence of civil disorder) ordre public nm

He was charged with breach of the peace.

Il a été condamné pour troubles de l'ordre public.

Compound Forms:

at peace		en paix
breach of the peace		perturbation de l'ordre public
breach of the peace	*nf*	rupture de la paix
breach of the peace	*nf*	violation de la paix
bring peace to	*v*	pacifier
dove of peace	*nf*	colombe de la paix
go in peace	*v*	aller en paix
in peace time	*adv*	en temps de paix
justice of the peace	*nf*	justice de la paix
keep the peace	*v*	préserver la paix
kiss of peace	*n*	baiser de paix (*dans l'église catholique*)
live in peace	*v*	vivre en paix
make peace	*v*	faire la paix
make-peace	*v*	faire la paix
peace and quiet	*nf*	tranquillité
peace be with you (religion)	*phrase*	que la paix soit avec vous (*religion*)
peace dove	*nf*	colombe de la paix
peace of God	*nf*	paix de Dieu
peace of mind	*nf*	tranquillité d'esprit
peace offering	*nf*	offre de paix
peace offering	*nm*	gage de réconciliation
peace officer	*nf*	agent de la paix
peace officer	*nm*	gardien de la paix
peace pipe	*nm*	pipe indienne (*pipe indienne*)
peace pipe	*nm*	calumet de la paix
peace-loving	*adj*	pacifique
rest in peace	*v*	reposer en paix
seek peace	*v*	chercher la paix
treaty of peace	*nf*	traité de paix

2) English-Bulgarian (Bulgarian-English) on-line dictionary, `http://www.eurodict.koralsoft.com/`

linguistics **[ling‘wistiks]** *n pl* (=*sing*) езикознание, лингвистки.

lingo
lingua franca
lingual
linguiform
linguist
linguistic
linguistics
lingulate
linhay
liniment
lining
link

Conclusion

Looking at several bilingual dictionaries of the Polish language we can assume that the creation of a trilingual, and later on, a quadri- or quintilingual experimental digital dictionary would not be a great problem, if we set Polish as a "hub" language. The data for such a experimental multilingual dictionary would be provided by the Polish-Ukrainian, Bulgarian-Polish and Polish-Lithuanian digital dictionaries, which are being developed.

From a theoretical and cognitive perspective it should not be too hard to create a multilingual dictionary of the Slavic languages with English as a "hub" language. The ideal for a multilingual digital dictionary, from a theoretical standpoint however, would be a dictionary with a semantic initial interlanguage.

The main difficulty in the creation of a multilingual digital dictionary is the fact that in every language the lexical form has not just one, but several meanings and they do not overlap across the respective compared languages. This is the reason why in such cases we should follow the content and not the lexical form, which is exemplified by our experience with the contrastive Bulgarian-Polish grammar. For this purpose we should begin working on the creation of an interlanguage dictionary of basic language concepts. Given forms in the different languages would correspond to the given concept. Examples would be the concepts respect, love, living creatures, household objects, etc. Such a dictionary would have to include concepts, which "classify" reality, for instance, names of human beings (man, woman, child), of a living being (man, animal), of working places (office, workshop), etc. This could be an overambitious goal, but it could be realizable in a broader interdisciplinary research group.

About the authors
We would like to present briefly the Polish and Bulgarian research teams, working on the project described in this paper, because of the project's interdisciplinary nature. The Polish research group from the Department of Semantics at the Institute for Slavic Studies (ISS) of Polish Academy of Sciences (PAS) works in theoretical semantics and cognitive linguistics. The Bulgarian research group from the Department of Mathematical Linguistics at the Institute for Mathematics and Informatics (IMI) of Bulgarian Academy of Sciences (BAS) works in the field of digital lexicology and lexicography.

L. Dimitrova and V. Koseska are the heads of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual digital dictionary" between PAS and BAS.

References

Barwise J., Perry J. (1983) Situations and Attitudes. Bradford Books, MIT.

Bogusławski A. (1988) Język w Słowniku, Wrocław.

Burnard L. (1995) What is SGML and How Does It Help? *In Computers and the Humanities*, 29, pp. 41–50.

Cooper R. (1996) The Role of Situations and Generalized Quantifiers. *In The Handbook of Contemporary Semantic Theory.* Shalom Lappin (ed.). Oxford.

Dimitrova L. (1998) Lexical Resource Standards and Bulgarian Language. *In International Journal Information Theories & Applications.* Vol. 5, Nr. 1. Pages 27–34.

Dimitrova L., Erjavec T., Ide N., Kaalep H.-I., Petkevic V, Tufis D. (1998) Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *Proceedings of COLING-ACL '98.* Montréal, Québec, Canada, pp. 315–319.

Dimitrova L., Pavlov R., Simov K. (2002) The Bulgarian Dictionary in Multilungual Data Bases. *Cybernetics and Information Technologies.* Vol. 2, nr. 2, pp. 12-15.

Ide N., Sperberg-McQueen C. M. (1995) The TEI: History, Goals, and Feature. *In Computers and the Humanities*, 29, pp. 5–15.

Koseska-Toszewa V. (2006) Bułgarsko-polska gramatyka konfrontatywna, t. 7, Semanticzna kategorija czasu. Warszawa, Polska.

Koseska-Toszewa V., Gargov G. (1990) *Semantičnata kategorija opredelenost-neopredelenost, Bàlgarsko-polska sàpostavitelna gramatika*, t. II, Sofija. (in Bulgarian)

Koseska-Toszewa V., Mazurkiewicz A. (1988) Net Representation of Sentences in Natural Languages. *In Lecture Notes in Computer Science* 340. Advances in Petri Nets. Springer—Verlag, pp. 249-266.

Koseska-Toszewa V., Mazurkiewicz A. (1994) Description a l'aide de reseaux de la temporalite et modalite dans la phrases dans la langue naturelle. *In Studia kognitywne.* Vol. 1, Warszawa, pp. 89-112.

Koseska-Toszewa V., Mazurkiewicz A. (2004) Once More about Net Representation of the Semantic Category of Tense. *In Etudes Cognitives.* Vol. 6, Warszawa, pp. 63-81.

Koseska-Toszewa V., Korytkowska M., Roszko R. (2007) *Polsko-bułgarska gramatyka konfrontatywna*, Warszawa, Dialog.

Lekov I., Sławski F. (ed.) (1961) Polish-Bulgarian dictionary. BAS Publishing House, Sofia, Bulgaria.

Mazurkiewicz A. (1986) Zdarzenia i stany: elementy temporalności. *In Studia gramatyczne bułgarsko-polskie.* Vol. I, Temporalność. Wrocław, pp. 7-21.

Petri C. A. (1962) Fundamentals of the Theory of Asynchronious Information Flow. *In Proceedings of IFIP'62 Congress.* North Holland Publ. Comp., Amsterdam.

Przepiórkowski A. (2006) The Potential of the IPI PAN Corpus. *In* PSiCL. Vol. 41, Poznań, Poland, pp. 31-48.

Radewa S. (1988) Podręczny słownik Polsko-Bułgarski z suplementem. Warszawa, Polska.

Russell B. (1948) Human Knowledge: Its Scope and Limits. London, UK.

Russell 1967) Denotowanie, Deskrypcje. *In Logika i język.* Warszawa, pp. 259-293.

Russell 1970) Mój rozwój filozoficzny. PWN, (tłumaczenie dzieła z roku 1959), Warszawa, pp. 276-279.

Sławski F. (1987) Podręczny słownik bułgarsko-polski z suplementem. Warszawa, Polska.

Sperberg-McQueen C. M., Burnard L. (ed.) (1994) *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford.

Stieber Z. (1966) Historyczna i współczesna fonologia języka polskieg. Warszawa, PWN, pp. 85-86.

Stieber Z. (1973) Zarys gramatyki porównawczej języków słowiańskich. Fleksja werbalna. Warszawa.

Strawson P. F. (1967) Odnoszenie się ujęcia wyrażeń do przedmiotów. *In Logika i język. Studia z semiotyki logicznej.* Warszawa, pp. 277-293.

Tufis D., Rotariu G., Barbu A.-M. (1999) Data sampling, lemma selection and a core explanatory dictionary of Romanian. *In COMPLEX'99*, Pecs, Hungary, pp. 219-228.