Bulgarian-Polish-Lithuanian Corpus – Current Development

Ludmila Dimitrova
IMI-BAS
Acad. G. Bonchev St bl. 8
1113 Sofia, Bulgaria
ludmila@cc.bas.bg

Violetta Koseska
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
amaz@inetia.pl

Danuta Roszko
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
danuta.roszko@ispan.waw.pl

Roman Roszko
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
roman.roszko@ispan.waw.pl

Abstract

This paper discusses the building of the first Bulgarian–Polish–Lithuanian (for short, BG–PL–LT) experimental corpus. The BG–PL–LT corpus (currently under development only for research) contains more than 3 million words and comprises two corpora: parallel and comparable. The BG–PL–LT parallel corpus contains more than 1 million words. A small part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of official documents of the European Union available through the Internet. The texts (fiction) in other languages translated into Bulgarian, Polish, and Lithuanian form the main part of the parallel corpus. The comparable BG–PL–LT corpus includes: (1) texts in Bulgarian, Polish and Lithuanian with the text sizes being comparable across the three languages, mainly fiction, and (2) excerpts from E-media newspapers, distributed via Internet and with the same thematic content. Some of the texts have been annotated at paragraph level. This allows texts in all three languages and in pairs BG–PL, PL–LT, BG–LT, and *vice versa* to be aligned at paragraph level in order to produces aligned three- and bi-lingual corpora. The authors focused their attention on the morphosyntactic annotation of the parallel trilingual corpus, according to the Corpus Encoding Standard (CES). The tagsets for corpora annotation are briefly discussed from the point of view of possible unification in future. Some examples are presented.

Keywords

Bilingual and multilingual corpora, parallel and comparable corpora, corpus annotation, lexical database, bilingual dictionaries.

1. Introduction

Due to the recent development of information and communication technologies and the increased mobility of people around the globe, the number of electronic dictionaries has increased extraordinarily. This concerns, in particular, bilingual dictionaries, in which one of the languages is English. An Internet search shows that no electronic dictionaries exist at all for pairs of languages such as Bulgarian-Polish or Bulgarian-Lithuanian. Traditional printed paper dictionaries are either an antiquarian rarity (the most recent Bulgarian-Polish and Polish-Bulgarian dictionaries were published more than 20 years ago) or have never been published at all (Bulgarian-Lithuanian). It can not be expected however that all people know English to communicate with each other, especially if their native languages (Bulgarian and Polish) belong to the same language family. For the creation of a bilingual electronic or online dictionary for Bulgarian, Polish and Lithuanian an electronic corpus is necessary which will provide the material for lexical database, supporting the dictionary and its subsequent expansion and update. In the recent decades many multilingual corpora were created in the field of corpus linguistics, such as MULTEXT corpus [6], one of the largest EU projects in the domain of language technologies, the MULTEXT-East corpus (MTE for short, annotated parallel and comparable), an extension of the project MULTEXT for Central and Eastern European (CEE) languages [2], Hong Kong bilingual parallel English-Chinese corpus of legal and documentary texts [5], etc.

2. From Bilingual to Trilingual corpus

The MTE project has developed a multilingual corpus, in which three languages: Bulgarian, Czech and Slovene, belong to the Slavic group. The MTE model is being used in the design of the first Bulgarian-Polish corpus, currently under development in the framework of the joint research project “Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary” between Institute of Mathematics and Informatics—Bulgarian Academy of Sciences and Institute of Slavic Studies—Polish Academy of Sciences, coordinated by L. Dimitrova and V. Koseska. This bilingual corpus supports the lexical database (LDB) of the first experimental online Bulgarian-Polish dictionary [3].

2.1 Bulgarian-Polish corpus

The Bulgarian–Polish corpus consists of two parts: a parallel and a comparable corpus [4]. All texts in the corpus are texts published in and distributed over the Internet. Some texts in the ongoing version of the corpus are annotated at paragraph level. The **Bulgarian–Polish parallel corpus** includes two parallel sub-corpora:

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

2) a *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 literary works in third language (mainly English).

The **Bulgarian–Polish comparable corpus** includes texts in Bulgarian and Polish: excerpts from newspapers and textual documents, shown in internet, excerpts from several original fiction, novels or short stories, with the text sizes being comparable across the two languages. Some of the Bulgarian texts are annotated at "paragraph" and "sentence" levels, according to CES [7].

2.2 Bulgarian–Polish–Lithuanian corpus

The first Bulgarian–Polish–Lithuanian (for short, BG–PL–LT) corpus (currently under development only for research) contains more than 3 million words and comprises two corpora: parallel and comparable. The **BG–PL–LT parallel corpus** contains more than 1 million words. A small part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of official documents of the European Union available through the Internet. The texts (fiction) in other languages translated into Bulgarian, Polish, and Lithuanian form the main part of the parallel corpus.

It turned out that it is extremely difficult to find electronic texts of translations from Bulgarian to Lithuanian or *vice versa* – the two languages are spoken by small nations in comparison to other languages of the EU and are distributed in remote areas of Europe. It can be assumed (provisionally of course) that the Polish language 'builds a bridge' between them: for the pairs of languages Bulgarian-Polish and Polish-Lithuanian one can find freely available translations on the Internet.

The comparable BG–PL–LT corpus includes: (1) texts in Bulgarian, Polish and Lithuanian with the text sizes being comparable across the three languages, mainly fiction, and (2) excerpts from E-media newspapers, distributed on the Internet and with the same thematic content.

Some of the texts have been annotated at paragraph level. This allows texts in all three languages and in pairs BG–PL, PL–LT, BG–LT, and *vice versa* to be aligned at paragraph level in order to produces aligned three- and bi-lingual corpora. "Alignment" means the process of relating pairs of words, phrases, sentences or paragraphs in texts in different languages which are translation equivalent. One may say that "alignment" is a type of annotation performed over parallel corpora. Excerpts of texts of the 3-languages parallel corpus, marked at paragraph level follow:

Bulgarian:

<p>Вместо отговор Гандалф гръмогласно подвикна на коня си:</p>

<p>- Напред, Сенкогрив! Трябва да бързаме. Няма време. Виж! Сигналните клади на Гондор горят, зоват за помощ. Войната е избухнала. Виж, огън бушува над Амон Дин, пламък покрива Ейленах, сигналът бърза на запад: Нардол, Ерелас, Мин-Римон, Каленхад и Халифириен на роханската граница.</p>

Polish:

<p>Zamiast odpowiedzieć hobbitowi, Gandalf krzyknął głośno do swego wierzchowca:</p>

<p>- Naprzód, Gryfie! Trzeba się spieszyć. Czas nagli. Patrz! W Gondorze zapalono wojenne sygnały, wzywają pomocy. Wojna już wybuchła. Patrz, płoną ogniska na Amon Din, na Eilenach, zapalają się coraz dalej na zachodzie! Rozbłyska Nardol, Erelas, Min-Rimmon, Kalenhad, a także Halifirien na granicy Rohanu.</p>

Lithuanian:

<p>Užuot atsakęs Gendalfas garsiai riktelėjo žirgui:</p>

<p>- Pirmyn, Žvaigždiki! Reikia skubėti. Laiko nebeliko. Žiūrėk! Jau dega Gondoro laužai, prašo pagalbos. Karo kibirkštis įžiebta. Matai, ant Amon Dino dega ugnis, liepsnoja ir Eilenachas, dar toliau vakaruose - Nardolas, Erelasas, Minas Rimonas, Kalenhadas ir Halifirienas prie Rohano sienos.</p>

//EN: For answer Gandalf cried aloud to his horse. 'On, Shadowfax! We must hasten. Time is short. See! The beacons of Gondor are alight, calling for aid. War is kindled. See, there is the fire on Amon Dîn, and flame on Eilenach; and there they go speeding west: Nardol, Erelas, Min-Rimmon, Calenhad, and the Halifirien on the borders of Rohan. (Part 3, Book 5 of *The Return of the King* of Tolkien's *The Lord of the Rings*)//

The BG-PL-LT corpus will be annotated according to the standards for morphosyntactic annotation of digital language resources. The main goal in collecting the trilingual corpus is the design and development of a BG–LT digital dictionary based on the BG-PL digital online dictionary.

The corpus will provide a sample of the vocabulary, which is to be included in an initial experimental versions of BG–LT digital dictionary.

We attempt to perform a comparison of the morphosyntactic characteristics of the words of parallel texts across the three languages from the point of view of a possible future unification.

3. Corpus annotation

Corpus annotation is the process of adding linguistic information in an electronic form to a text corpus [7], [8]. We would like to mention the following two most common types of corpus annotation: ***morphosyntactic annotation*** (also called *grammatical tagging* or *part of speech* ***(POS)*** *tagging*) and ***lemma annotation*** (where each word in the text is associated with the corresponding lemma). Lemma annotation is closely related to morphosyntactic annotation. Morphosyntactic annotation (POS tagging, where each word in the text is associated with its grammatical classification) is the task of labeling each word in a sequence of words with its appropriate part-of-speech. Words are often ambiguous with respect to their POS; for example, in Bulgarian the neuter singular forms of most adjectives serve double duty as adverbs, for example, BG: *внимателно* //EN: attentive/careful (neuter), attentively/carefully //:

(1) *внимателно* → POS specifications: adjective, Gender: neuter, Number: singular, Definiteness: no.

MTE MorphoSyntactic Descriptor (MSD) for this adjective is A--ns-n.

(2) *внимателно* → POS: adverb, Type: adjectival.

MTE MSD for this adverb is Ra.

The set of POS tags is called tagset. The size and choice of the tagsets vary across languages. The classical POS tagging system is based on a set of parts of speech including noun, adjective, numeral, pronoun, verb, participle, adverb, preposition, conjunction, interjection, particle, and often (depending on the language) article, etc. Of course, morphologically rich languages need more detailed tagsets that reflect to various inflectional categories.

The applications of the morphosyntactic annotation include lexicography, parsing, language models in speech recognition, disambiguation clues for ambiguous words (machine translation), information retrieval, spelling correction, etc.

4. Problems related to POS classification

The POS classification varies across different languages. Often there is more than one possible POS classification for a given language.

Here we would like to show that one cannot formally go about a direct use of the morphosyntactic annotation of a multilingual corpus. An in-depth contrastive study of specific phenomena in the respective languages is necessary. Next we will briefly review the POS classification of the *participle* (one of the important verbal forms) in the three languages, in comparison to another POS, the *adjective*.

4.1 Functions of the participle

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

(1) Examples of predicative function of the participle

BG: *украсен* // PL: *ozdobiony* // LT: *papuošta* [neuter], *papuoštas* [masculine] //EN: decorated//:
BG: *Коридорът е хубаво украсен*.
PL: *Korytarz jest ładnie ozdobiony*.
LT: *Koridorius gerai papuošta. / Koridorius gerai papuoštas.*
EN: *The corridor is beautifully decorated*.

(2) Examples of attributive function of the participle:

BG: *пишещ* // PL: *piszący* // LT: *rašantis //* EN: one who wrote //, in the sentences:
BG: *Пишещият тези писма старец беше осемдесетгодишен.*
PL: *Piszący te listy starzec był osiemdziesięciolatkiem.*
LT: *Rašantis tuos laiškus senelis buvo aštuoniasdešimtmetis.*
EN: *The old man who wrote these letters was eighty years old.*

(3) Examples of the semi-predicative function:

BG: *пишейки* // PL: *pisząc* // LT: *rašydamas* // EN: while writing //, in the sentences:
BG: *Пишейки, гледах през прозореца*.
PL: *Pisząc patrzyłem w okno*.
LT: *Rašydamas žiūrėjau per langą*.
EN: *While writing, I was looking out of the window*.

4.2 Participle and verb

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

a) Present active participle: BG: *говорещ* // PL: *mówiący* // LT: *kalbąs* / *kalbantis* // EN: *speaking* // (preserved active voice).

b) Present passive participle: BG: *любим*[1] //PL: *kochany* // LT: *mylimas* // EN: *beloved* // (preserved passive voice with information about present tense).

c) Past passive participle: BG: *написан* // PL: *napisany* //LT: *parašytas* // EN: *written* // (preserved passive voice with information about past tense and perfect aspect of the verbal form).

An interesting fact is that participles preserve the valency properties of the respective verbal form, for instance in Polish and Lithuanian:

PL: *Ten mężczyzna zajmuje się drobnym handlem*. – *Zajmujący sie drobnym handlem mężczyzna*. // LT: *Tas vyras užsiima mažmenine prekyba. – Mažmenine prekyba užsiimantis vyras*. // EN: *This man deals in retail. – A man dealing in retail.*

[1] Colloquial Bulgarian has lost this grammatical category. Such forms occur mostly in scientific writing, being literary loans from Russian or Church Slavonic. Because of their grammatical unproductiveness, they are classified as adjectives, corresponding to the Latin-derived adjectives in -*able*/-*ible* in English: (*не*)*допустим – (in)admissible, недосегаем – intangible, съвместим – compatible, etc.*

The phrase 'deals in what? / dealing in what?' requires the instrumental case in Polish and Lithuanian[2]. The valence of the Polish and Lithuanian participle is the same as the valence of the finite verb form.

A comparison of the three languages shows that in Bulgarian a subordinate clause in past perfect tense corresponds to a participle construction in Polish and Lithuanian:

BG: *След като си беше написал домашното, той започна да чете книга*. // PL: *Odrobiwszy lekcje zaczął czytać książkę*. // LT: *Paruošęs pamokas pradėjo skaityti knygą*. // EN: *Having written his homework, he started reading a book*.

Polish has a more modest stock of verbal forms with temporal meaning than Bulgarian or Lithuanian. In any case when the lexical means modifying the temporal meanings are taken into account, the participles, and verbal nouns, it is clear that Polish can express also the same temporal meanings.

4.3 Features of the adjective

Adjectives in Polish and Lithuanian can be declined for gender, number and case (in Bulgarian only for gender and number), but do not express a temporal or aspect relation on their own, unlike the participle. These arguments show that participles deserve a separate treatment from adjectives.

5. Towards development of annotated trilingual electronic resources

Morphosyntactic descriptions for Bulgarian have been developed in several projects, the first of which are for the purposes of corpora processing at the morpho-lexical level in MTE project of EC. The MTE consortium developed morphosyntactic specifications and word-form lexical lists (so called lexicons) covering at least the words appearing in the MTE corpus. For each of the six MTE languages, a lexical list containing at least 15,000 lemmata was developed for use with the morphological analyzer. Each lexicon entry includes information about the inflected-form, lemma, POS, and morphosyntactic specifications. A mapping from the morphosyntactic information contained in the lexicon to a set of corpus tags (used by the POS disambiguator) was also provided, according to the MULTEXT tagging model. The structure of the lexicon entry is the following:

word-form ‹TAB› **lemma** ‹TAB› **MSD** ‹TAB› **comments**

where **word-form** represents an inflected form of the lemma, characterised by a combination of feature values encoded by **MSD**-code (**MSD**: **M**orpho**S**yntactic **D**escription); the fourth (optional) column, comments, is currently ignored and may contain either comments or information processable by other tools. Here is an excerpt from the Bulgarian Lexicon:

обяснение	=	Ncns-n
обяснението	обяснение	Ncns-y
обяснения	обяснение	Ncnp-n
обясненията	обяснение	Ncnp-y

(обяснение 'explanation').

The **MSDs** are provided as strings, using a linear encoding; an efficient and compact way for the representation of the flat attribute-value matrices. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way: the character at position 0 encodes part-of-speech; each character at position 1, 2, …, *n*, encodes the value of one attribute (person, gender, number, etc.), using the one-character code; if an attribute does not apply, the corresponding position in the string contains the special marker "-" (hyphen). By convention, trailing hyphens are not included in the **MSD**s. Such specifications provide a simple and compact encoding, and are similar to feature-structure encoding used in unification-based grammar formalisms. When the word form is the very lemma, then the equal sign is written in the lemma field of the entry ("=").

For Bulgarian the morphosyntactic descriptions were designed on the basis of the traditional POS classification according to the traditional Bulgarian grammar (Bulgarian Grammar 1993). Each word form is assigned a label encoding the major category (POS), type where applicable (e.g., proper *versus* common noun) and inflectional features. Punctuation is also included, as are abbreviations, numbers written in digits, and unidentified objects (residuals). A further non-standard category contains markers of degrees of comparison. Those are formed in Bulgarian with the particles *по* (comparative) and *най* (superlative), preposed to the adjective or adverb but separated from it by a hyphen (*лек* 'light', *по-лек* 'lighter', *най-лек* 'lightest'; *леко* 'easy', *по-леко* 'more easily', *най-леко* 'most easily'). These particles are annotated as separate words:

по → POS: Particle, Type: comparative, Formation: simple,
най → POS: Particle, Type: superlative, Formation: simple.

The morphosyntactic descriptions for Polish: the description of Polish by Saloni [15] serves as a basis for the morphosyntactic descriptions for Polish and has been adapted to a large degree to the MTE MSD format in [14].

The system of morphosyntactic tags developed for the Polish at the Institute of Computer Science, Polish Academy of Sciences (IPI PAN), is based on a sound methodological foundation comprising linguistic work by authors such as J.S. Bień, Z. Saloni, M.Świdziński. It is

[2] This does not apply to Bulgarian which lacks a case paradigm for nouns.

thanks to this foundation that the IPI PAN's tagset goes beyond the fossilised traditional framework dating back to Aristotle. On the other hand, the MTE tagset, which serves as a point of reference here, is based on the traditional subdivision into parts of speech (this is why, among others, pronouns have been singled out as a part of speech).

Consequently, the aim of our work is neither to revise the good and highly refined IPI PAN tagset nor to replace it with a new tagset for Polish. The issue in question is what kind of compromise should be sought when developing a joint tagset to be used for simultaneous description of the three languages in the BG-PL-LT parallel corpus. For some reasons the MTE tagset (developed previously for many languages) has been selected as the leading one for this corpus. Therefore, the aim of our work is to provide a theoretical study of various categories of Polish (and Lithuanian), to set priorities (e.g. morphological, semantic, syntactic) in identifying various meanings and to provide a classification of morphosyntactic phenomena which does not contradict the MTE standard and does not deviate too strongly from the IPI PAN tagset.

It cannot be excluded that due to the obvious difficulties in achieving consistency of the intertagset the BG-PL-LT corpus will use the IPI PAN tagset for Polish and its modification for Lithuanian. This solution would certainly necessitate a list of more or less close equivalents for the two tagsets: a tagset for Bulgarian on the one hand, and the IPI PAN tagset on the other (for Polish and an extended version for Lithuanian).

It is important to emphasise that only a coherent tagset for a parallel multilingual corpus 1) allows complete linguistic confrontation, 2) enables identification of linguistic facts, 3) enables a search based on pre-defined unambiguous morphosyntactic characteristics.

The morphosyntactic descriptions for Lithuanian: as a basis for morphosyntactic descriptions of Lithuanian serve the Academic grammar of the Lithuanian language [10] and the Functional grammar of Lithuanian [16]. A tool for morphosyntactic annotation for Lithuanian - *MorfoLema* - has been created by Vytautas Zinkevičius in Centre of Computational Linguistics of Vytautas Magnus University (Lithuania) [18]. The program *MorfoLema* can perform a morphosyntactic analysis and generate forms of Lithuanian words based on user's morphosyntactic characteristic. For disambiguation the *MorfoLema* uses „Two-level morphology" method of Kimmo Koskenniemi [9].

The next step of the development of a system for morphological annotation (Morfologinis anotatorius [20]) has been realised by Vidas Daudaravičius and Erika Rimkutė. Vidas Daudaravičius has created disambiguation tools for the *Morfologinis anotatorius*. More information about the *Morfologinis anotatorius* and used set of tags we can find on http://donelaitis.vdu.lt/main.php?id=4&nr=7_1 (in Lithuanian). (The names of tags are in Lithuanian, because the authors of the *Morfologinis anotatorius* didn't use English terms.) It is possible to perform online a morphosyntactic analysis through the web-page http://donelaitis.vdu.lt/main.php?id=4&nr=7_2. The results are visualized on the screen, and it is possible to receive the result as a file.

The tag list for Polish and Lithuanian, based on [11], [12], [13], [17], and used in the example below, follows:

subst - noun	nwok - nonvocal
sg – singular	adj - adjective
pl – plurale	verb - verb
nom – nominative	praes - present
gen – genitive	nonpraet - nonpraeteritum
acc - accusative	ter - 3rd person
loc - locative	bezosobnik - non person form of verb
m - masculine	perf - perfective
f - feminine	imperf - imperfective
-hum – nonhuman	particle - particle
-ani – nonanimate	prep – preposition

A comparison between experimental annotations of the following sentence "*The beacons of Gondor are alight, calling for aid.*[3]" of the parallel corpus was performed:

BG: Сигналните клади на Гондор горят, зоват за помощ.

PL: W Gondorze zapalono wojenne sygnały, wzywają pomocy.

LT: Jau dega Gondoro laužai, prašo pagalbos.

The annotation of the Bulgarian text is done with MTE MSDs, and ISSCO TAGGER [19] is used for disambiguation. For manual annotation of the Polish and Lithuanian text the above-mentioned descriptors are used, because these languages lack developed MTE language specifications. Establishing a 1-1-correspondence between the tags used and the MTE tagset does not present an insurmountable difficulty. The result follows:

Bulgarian (MTE annotation)

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok type=WORD>
<orth>Сигналните</orth>
<disamb><base>сигнален</base><ctag>AP</ctag></disamb>
<lex><base>сигнален</base><msd>A---p-
y</msd><ctag>AP</ctag></lex>
</tok>
<tok type=WORD>
<orth> клади </orth>
<disamb><base>клада</base><ctag>NCFP-N</ctag></disamb>
<lex><base>клада</base><msd>Ncfp-
n</msd><ctag>NCFPN</ctag></lex></tok>
<tok type=WORD>
<orth>на</orth>
<disamb><base>на</base><ctag>SP</ctag></disamb>
```

[3] Tolkien, J.R.R. The Lord of the Rings. Boston : Houghton Mifflin, 1994, p. 731.

```
<lex><base>на</base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base>на</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD>
<orth>Гондор</orth>
<disamb><base>Гондор</base><ctag>NPMS-
N</ctag></disamb>
<lex><base>Гондор</base><msd>Npms-n</msd><ctag>NPMS-
N</ctag></lex>
</tok>
<tok type=WORD >
<orth>горят</orth>
<disamb><base>горя</base><ctag>VMIP3P</ctag></disamb>
<lex><base> горя
</base><msd>Vmia3p</msd><ctag>VMIA3P</ctag></lex>
<lex><base> горя
</base><msd>Vmip3p</msd><ctag>VMIP3P</ctag></lex>
</tok>
<tok type=PUNCT >
<orth>,</orth>
<ctag>COMMA</ctag>
</tok>
<tok type=WORD >
<orth>зоват</orth>
<disamb><base>зова</base><ctag>VMIP3P</ctag></disamb>
<lex><base>зова</base><msd>Vmia3p</msd><ctag>VMIA3P</
ctag></lex>
<lex><base>зова</base><msd>Vmip3p</msd><ctag>VMIP3P</
ctag></lex>
</tok>
<tok type=WORD>
<orth>за</orth>
<disamb><base>за</base><ctag>SP</ctag></disamb>
<lex><base>за</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD>
<orth> помощ </orth>
<disamb><base> помощ </base><ctag>NCFS-
N</ctag></disamb>
<lex><base> помощ </base><msd>Ncfs-n</msd><ctag>NCFS-
N</ctag></lex>
</tok>
<tok type=PUNCT>
<orth>.</orth>
<ctag>PERIOD</ctag>
</tok>
</chunk>
</chunkList>
</cesAna>
```

Polish [11]

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok>
<orth>W</orth>
<lex><base>w</base><ctag>prep:loc:nwok</ctag></lex>
</tok>
<tok>
<orth>Gondorze</orth>
<lex><base>Gondora</base><ctag>subst:sg:loc:f</ctag></lex>
</tok>
<tok>
<orth>zapalono</orth>
<lex><base>zapalić</base><ctag>verb:bezosobnik:perf</ctag></l
ex>
</tok>
<tok>
<orth>wojenne</orth>
<lex><base>wojenny</base><ctag>adj:pl:acc:-hum</ctag></lex>
</tok>
<tok>
<orth>sygnały</orth>
<lex><base>sygnał</base><ctag>subst:pl:acc:-hum</ctag></lex>
</tok>
<ns/>
<tok>
<orth>,</orth>
<lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
<tok>
<orth>wzywają</orth>
<lex disamb="1"><base>wzywać</base><ctag>
verb:nonpraet:pl:ter:imperf</ctag></lex>
</tok>
<tok>
<orth>pomocy</orth>
<lex><base>pomoc</base><ctag>subst:sg:gen:f</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
</chunk></chunkList></cesAna>
```

Lithuanian

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok>
<orth>Jau</orth>
<lex><base>jau</base><ctag>particle</ctag></lex>
</tok>
<tok>
<orth>dega</orth>
<lex><base>degti</base><ctag> verb:praes.ter</ctag></lex>
</tok>
<tok>
<orth>Gondoro</orth>
<lex><base>Gondoras</base><ctag>subst:sg:gen:m</ctag></lex
>
</tok>
<tok>
<orth>laužai</orth>
<lex><base>laužas</base><ctag>subst:pl:nom:m</ctag></lex>
</tok>
<ns/>
<tok>
<orth>,</orth>
<lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
```

```
<tok>
<orth>prašo</orth>
<lex disamb="1"><base>prašyti</base><ctag>
verb:praes.ter</ctag></lex>
</tok>
<tok>
<orth>pagalbos</orth>
<lex><base>pagalba</base><ctag>subst:sg:gen:f</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
</chunk>
</chunkList>
</cesAna>
```

6. Annotation of parallel corpus – problems and progress

A parallel corpus of two Slavic languages and one Baltic language is of great interest from the viewpoint of describing the similarities and differences of the formal means of these three languages. Bulgarian belongs to the South subgroup, Polish – to the West subgroup of the Slavic languages. Lithuanian belongs to the Eastern Baltic group. All three languages preserve the special features for each corresponding group.

A significant feature is the analytic character of Bulgarian, and the synthetic character of Lithuanian (with some analytic character, like word order in absolute constructions) and Polish. Bulgarian exhibits several linguistic innovations in comparison to the other Slavic languages (a rich system of verbal forms, a definite article), and has a grammatical structure closer to English, Modern Greek, or the Neo-Latin languages than Polish. The definite article in Bulgarian is postpositive, whereas in Lithuanian a similar function is served by qualitative adjectives and adjectival participial forms, both with pronominal declension. Bulgarian preserves some vestiges of case forms in the pronoun system. Polish and Lithuanian exhibit all features of synthetic languages (a very rich case paradigm for nouns). Although Lithuanian has lost the neuter gender of nouns, its case system is richer than the Polish one. Bulgarian and Lithuanian have a high number of verbal forms, but Polish has reduced most of the forms for past tense. Both Polish and Bulgarian have a strongly developed category of verbal aspect. In Lithuanian the verb can have more than one aspect depending on the usage of a base stem for present, past and future tense.

7. Conclusion

One of the main problems in human communication is the presence of a huge variety of written and spoken languages in the world. Finding ways to support the connection of people from different ethnical parts of the world is becoming more and more important. The advantage of processing a trilingual parallel corpus is to obtain context specific information about syntactic and semantic structures and usage of words in given language(s). The parallel BG–PL–LT corpus will enrich and uncover some unstudied features of the three languages. Furthermore, a trilingual corpus can find applications into the design and development of LDB of future bilingual dictionaries, for example, of a LDB supporting a BG–LT dictionary, based on a LDB that supports a BG–PL online dictionary.

Finally we note that the trilingual corpus can be used in education, in schools as well as universities; it will be useful to students, instructors, and linguists-researchers alike.

8. References

[1] Bulgarian Grammar. (1993). Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

[2] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and 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, pp. 315-319.

[3] Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 36-47. ISBN 978-5-9900813-6-9.

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

[5] May Fan, Xu Xunfeng. (2002). An evaluation of an online bilingual corpus for the self-learning of legal English. http://langbank.engl.polyu.edu.hk/corpus/bili_legal.html

[6] Ide, N., and Véronis, J. (1994). Multext (multilingual tools and corpora). In *COLING'94*, pages 90-96, Kyoto, Japan.

[7] Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*,Granada, Spain, 463-70.

[8] Geoffrey Leech. (2004). Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm

[9] Kimmo Koskenniemi. (1983) Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. Helsinki: University of Helsinki, Department of General Linguistics.

[10] Lithuanian Grammar. (1997). Ed. Vytautas Ambrazas, Baltos lankos, Vilnius, pp.802.

[11] Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Constru-ction and Optimisation. Task Quarterly. 11, p. 151-167

[12] Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN, Polonica, XXII-XXIII, p. 57-76 (In Polish)

[13] Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Warszawa: Akademicka Oficyna Wydawnicza EXIT (In Polish)

[14] Roszko, R. (2009). Morphosyntactic Specifications for Polish. Theoretical foundations. In: Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava*. 140–150. ISBN 978-80-7399-745-8.

[15] Saloni, Z., W. Gruszczyński, M. Woliński, R.Wołosz (2007). Słownik gramatyczny języka polskiego, Wiedza Powszechna, Warszawa, CD + 177 s. (In Polish)

[16] Valeckienė, A. (1998). Funkcinė lietuvių kalbos gramatika, Mokslo ir enciklopedijų leidybos institutas, Vilnius, pp.415. (In Lithuanian)

[17] Woliński, M. (2003). *System znaczników morfosyntaktycznych w korpusie IPI PAN*, Polonica, XXII-XXIII, p. 39-55 (In Polish)

[18] Zinkevičius, V. (2000). Lemuoklis - morfologinei analizei. *Darbai ir dienos*, 24, Vytauto Didžiojo universitetas, p. 245-274 (In Lithuanian).

[19] ISSCO TAGGER: http://www.issco.unige.ch/staff/robert/tatoo/tagger.html#design

[20] Morfologinis anotatorius (tagger for Lithuanian): http://donelaitis.vdu.lt/main.php?id=4&nr=7_1