# Bulgarian-Polish-Lithuanian Corpus –Problems of Development and Annotation[1]

Ludmila Dimitrova[1], Violetta Koseska[2], Danuta Roszko[2], Roman Roszko[2]

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

**Abstract.** The paper shortly describes the first Bulgarian–Polish–Lithuanian (for short, BG–PL–LT) experimental corpus, currently under development only for research. The trilingual corpus comprises two corpora: parallel and comparable. We focused our attention on the morphosyntactic annotation of the parallel trilingual corpus, according to the Corpus Encoding Standard (CES). We briefly discuss the tagsets for corpora annotation from the point of view of possible unification in the future. Next, we review the Part-of-Speech (POS) classification of the *participle* in the three languages, in comparison to another POS, the *adjective*. Some examples are presented.

## 1 Introduction

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. Due to the recent development of information and communication technologies and the increased mobility of people around the globe, the number of bilingual electronic dictionaries, in which one of the languages is English, has increased extraordinarily. One cannot expect 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. 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). 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.

On the one hand, it is interesting to note that two Slavic languages are compared to a Baltic language (Lithuanian). Comparative and contrastive studies of Polish and Bulgarian as well as Polish and Lithuanian have been already conducted, but up to the best of our knowledge no such studies exist for Bulgarian and Lithuanian. On the other hand, the three languages are marginally present in the EU because of the later ascension of the three countries to the EU. Thus we expect a new and interesting scientific problem in front of us and hope that our studies will find a wider application.

## 2 From Bilingual to Trilingual corpus

In recent decades many multilingual corpora were created in the field of corpus linguistics, such as the MULTEXT corpus [7], the MULTEXT-East corpus, annotated parallel and comparable, (MTE for short), an extension of the corpus MULTEXT with six Central and Eastern European (CEE) languages [2], ParaSol, a parallel and aligned corpus of Slavic and other languages (so-called Regensburg Parallel Corpus) [23], Italian-German parallel corpus, a collection of legal and administrative documents written in Italian and German, due to the equal status of the both languages in South Tyrol [9], Hong Kong bilingual parallel English-Chinese corpus of legal and documentary texts [6], etc.

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 [4], [5], 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

[1] 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 IMI-BAS and ISS-PAS. The study and preparation of this paper have been partly supported by the EC's Seventh Framework Programme [FP7/2007-2013] under the grant agreement 211938 MONDILEX.

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. All collected 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 [8].

### 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 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 main part of the parallel corpus comprises texts (fiction, novels, short stories) in other languages translated into Bulgarian, Polish, and Lithuanian. When we have provided the electronic text of the original literary work or its translation, we include it as well in the 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.

We plan to annotate the BG-PL-LT parallel corpus according to the standards for morphosyntactic annotation of digital language resources.

**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 electronic newspapers, distributed via Internet and with the same thematic content.

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.

The structure of the parallel corpus groups texts according to content. Every group contains three parts (respectively four if the original language is different from the languages in the corpus). A detailed description of the corpus is provided for clarification to the user.

An excerpt of the description of the trilingual parallel corpus follows:

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

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

LT Lithuanian: Stanislavas Lemas, *Soliaris*. Translated by Giedrė Juodvalkytė. Vaga, Vilnius, 1978.

// EN:*Stanislav Lem, Solaris*.//

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**:*

Вместо отговор Гандалф гръмогласно подвикна на коня си:

- 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.

**Polish**:

Zamiast odpowiedzieć hobbitowi, Gandalf krzyknął głośno do swego wierzchowca:

- 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.

**Lithuanian**:

Užuot atsakęs Gendalfas garsiai riktelėjo žirgui:

- 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.

//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*)//

## 3 Corpus annotation and problems related to POS classification

*Corpus annotation* is the process of adding linguistic information in an electronic form to a text corpus [8], [10]. 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:

BG: гръмогласно //EN: loud(-voiced), uproariously, voice of thunder //:

(1) ***гръмогласно*** //loud(-voiced)//→ POS specifications: adjective, Gender: neuter, Number: singular, Definiteness: no.

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

(2) ***гръмогласно*** // uproariously, voice of thunder, cried aloud // → 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 POS classification varies across different languages. Often there is more than one possible POS classification for a given language.

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.

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 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. 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*.

## 3.1 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 semi-predicative function or adverbial 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** е осемдесетгодишен.

PL: **Piszący** te listy **starzec** jest osiemdziesięciolatkiem.

LT: **Rašančiam** tuos laiškus **seneliui** aštuoniasdešimt metų.

(EN: *The old man who wrote these letters is 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.*)

## 3.2 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) 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.*)

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.

## 3.3 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. The main

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

grammatical meaning of the adjective is the attributive meaning. Unlike the participle, which is closely related to a verbal action (state or event in the past, present and future), the adjective denotes a constant property or quality of the object such as:

малко дете | małe dziecko | mažas vaikas // *a little child* //

The adjectives across all three languages function not only as attribute, but also as predicate. As predicate they are only a nominal part of the predicate and express neither time nor aspect. Examples:

Малка къща | Mały dom | Mažas namas // *a small house*//

Къщата е малка. | Dom jest maly | Namas mažas[3]. (rarely: Namas yra mažas.) //*The house is small. //*

The neuter forms of Lithuanian adjectives possess a semi-predicative function:

LT: Man skanu (adjective, neuter).

BG: На мен ми е вкусно (adverb)./ Вкусно (adverb) ми е.

PL: Smakuje (verb) mi (to).

(EN:*I find it delicious.*)

LT: Gera (adjective, neuter) gyventi kaime.

BG: На село се живее добре (adverb).

PL: Dobrze (adverb) się mieszka na wsi.

(EN:*Living in the village is good.*)

Our observations show that participles have to be considered apart from the adjectives, since adjectives do not carry the verbal characteristics: voice, tense, aspect and valence. Mixing adjectives and participles is a sign of insufficient knowledge of the grammatical structure of Slavic or Baltic languages. Unification of adjectives and participles might be allowd for languages without aspect and/or whose descriptive system of aspect and tense of the verbal form is simpler compared to that of Slavic or Baltic languages. That is the main reason why participles have to be classified as separate POS and not re-qualified as adjectives.

## 4. 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:

[3] In Lithuanian the word order plays a great role in distinguishing the two functions.

| | | |
|---|---|---|
| потвърждение | = | Ncns-n |
| потвърждението | потвърждение | Ncns-y |
| потвърждения | потвърждение | Ncnp-n |
| потвържденията | потвърждение | Ncnp-y |

(потвърждение: *confirmation, corroboration*).

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).

**The morphosyntactic descriptions for Polish:** the description of Polish by Saloni [16] serves as a basis for the morphosyntactic descriptions for Polish and has been adapted to a large degree to the MTE MSD format in [15].

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 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 [12] and the Functional grammar of Lithuanian [17]. A tool for morphosyntactic annotation for Lithuanian - *MorfoLema* - has been created by Vytautas Zinkevičius in Centre of Computational Linguistics of Vytautas Magnus University (Lithuania) [19]. 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 [11].

The next step of the development of a system for morphological annotation (Morfologinis anotatorius [21]) 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 [21] 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 [22]. The results are visualized on the screen, and it is possible to receive the result as a file.

The authors of the Lithuanian *Morfologinis anotatorius* (see [21]) use the traditional to Lithuanian description of POS. They add two new POS: acronym (like LR for ***L****ietuvos* ***R****espublika* 'Republic of Lithuanian') and abbreviation (like gen. for ***gen****eralinis* 'main, leading (chief)'). In practice these are not POS, but a means to denote some phenomenon specific to the written language.

The list of POS used for Lithuanian in *Morfologinis anotatorius* follows:

| | POS | LT term | LT acronym |
|---|---|---|---|
| 1. | noun | daiktavardis | dkt. |
| 2. | adjective | būdvardis | bdv. |
| 3. | numeral | skaitvardis | sktv. |
| 4. | pronuon | įvardis | įv. |
| 5. | verb | veiksmažodis | vksm. |
| 6. | adverb | prieveiksmis | prv. |
| 7. | interjections | jaustukas | jst. |
| 8. | onomatopoeic words | ištiktukas | išt. |
| 9. | paricles | dalelytė | dll. |
| 10. | prepositions | prielinsknis | prl. |
| 11. | conjustions | jungtukas | jng. |
| 12. | acronym | akronimas | akronim. |
| 13. | abbreviation | sutrumpinimas | sutr. |

Subcategories such as gender, number, case, present, past, passive, active, etc., are described as separate categories and are not related to POS. This division is in correspondence with many of the subcategories in the Lithuanian academic grammar.

There are certain differences, for example: new case illative (who into? what into? where to?), new gender: bendroji giminė (bi-gendered), new number dviskaita (dual number), new voice reikiamybės (lat. necessitatis, eng. necessity). The grammar recognizes only synthetic verb tenses and adds one form of past tense būtasis laikas (lat. praeteritum, eng. past). The authors of *Morfologinis anotatorius* deviate from the tradition and ascribe the *tense* characteristic to participles, do not distinguish the analytic tense forms (for example, present perfect, present inchoative), but describe every element of theirs separately. They also form new categories: stabiliosios frazės (phrasal expressions), romėniški skaičiai (roman number), teigiamumas, negiamumas (negation, confirmation), apibrėžtumas (definiteness/indefiniteness). The category of apibrėžtumas (definiteness/indefiniteness) has two subcategories: įvardžiuotinis (definiteness) i neįvardžiuotinis (indefiniteness).

The names of tags are in Lithuanian, because the authors of the *Morfologinis anotatorius* did not use English terms.

The tag list for Polish and Lithuanian, based on [13], [14], [18], [21], [22] and used in the example below, follows:

For Polish:

| | | | | | |
|---|---|---|---|---|---|
| acc | – | accusative | m3 | – | masculine 3 |
| adj | – | adjective | n | – | neuter |
| conj | – | conjugation | nom | – | nominative |
| dat | – | dative | pl | – | plurale |
| f | – | feminine | perf | – | perfective |
| gen | – | genitive | pos | – | positive degree |
| inf | – | infinitive | praet | – | past |
| interp | – | punctuation mark | prep | – | preposition |
| m1 | – | masculine 1 | sg | – | singular |
| m2 | – | masculine 2 | subst | – | noun |

For Lithuanian:

| | | | | | |
|---|---|---|---|---|---|
| 3 asm. | – | 3rd person | prv. | – | adverb |
| būt. k. l. | – | past | sep. | – | punctuation mark |
| dkt. | – | nomen | teig. | – | confirmation |
| dlv. | – | participle | tiesiog. n. | – | indicative mood |
| N. | – | dative | veik. r | – | active voice |
| neįvardž. | – | indefiniteness | vyr. g. | – | masculine |
| nelygin. l. | – | positive degree | vksm. | – | verb |
| nesngr. | – | non-reflexive | vns. | – | singular |
| nežinomas | – | unknown | V. | – | nominative |

A comparison between experimental annotations of the following sentence "*For answer Gandalf cried aloud to his horse*.[4]" of the parallel corpus was performed:

BG: Вместо отговор Гандалф гръмогласно подвикна на коня си:
PL: Zamiast odpowiedzieć hobbitowi, Gandalf krzyknął głośno do swego wierzchowca:
LT: Užuot atsakęs Gendalfas garsiai riktelėjo žirgui:

The annotation of the Bulgarian text is done with MTE MSDs, and ISSCO TAGGER [20] 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 could be seen in **Appendix.**

## 5. Applications of the trilingual corpus

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. Each one of the three languages however has specific traits which make it unique within the respective language group.

We studied some characteritics in the previous parts. Here we will consider some significant differences between the languages which can be illustrated by examples of texts from the trilingual corpus.

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.

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. 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 or languages. Let us consider an entry of the BG–PL LDB, whose respective dictionary entry of the BG-PL printed dictionary is:

**сп|я, -иш** *vi*. spać; **~и ми се** chce mi się spać, ogarnia mnie senność

The grammatical features of this Bulgarian verb ***спя*** /sleep/ are:

**aspect - imperfect (progressive)** /*несвършен вид*/, this verb is **intransitive** /*непреходен*/, its conjugation is a **II type** /*II спрежение*/.

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

Its structure in **BG-PL** LDB is:

<entry>
<hw>**сп|я'**</hw>
<pos>verb</pos>
<gram>imperfect</gram>
<conjugation><orth>**-и'**</orth>
<type>**II**</type>
</conjugation>
<subc>intransitive</subc>
<struc type="Sense" n="1">
<trans**> spać** </trans>
</struc>
<struc type="Derivation" n="1">
<orth>**~и ми се**</orth>
<struc type="Sense" n="1">
<trans> **chce mi się spać** </trans>
<alt><trans> **ogarnia mnie senność** </trans></alt>
</struc>
</struc>
**<**/entry>

A possible structure in a future **BG-LT** LDB should be:

<entry>
<hw>**сп|я'**</hw>
<pos>verb</pos>
<gram>imperfect</gram>
<conjugation><orth>**-и'ш**</orth>
<type>**II**</type>
</conjugation>
<subc>intransitive</subc>
<struc type="Sense" n="1">
<trans**> miegoti** </trans>
</struc>
<struc type="Derivation" n="1">
<orth>**~и ми се**</orth>
<struc type="Sense" n="1">
<trans> **(aš) noriu miego** </trans>
</struc>
</struc>
**<**/entry>

**In conclusion** we note that the parallel BG–PL–LT corpus will enrich and uncover some unstudied features of the three languages. It will be useful to linguists-researchers for research purposes alike, for instance in contrastive studies of the three languages together or in pairs.

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

## 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] Dimitrova, L., V. Koseska-Toszewa. (2009). Bulgarian-Polish Corpus. In: *International Journal Cognitive Studies / Études Cognitives*. 9, SOW, (in print).

[6] 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

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

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

[9] Gamper, Dongilli. (1999). Primary Data Encoding of a Bilingual Corpus. http://titus.uni-frankfurt.de/curric/gldv99/paper/gamper/gamperx.pdf

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

[11] 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.

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

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

[14] Przepiórkowski A. (2004), The IPI PAN Corpus: Preliminary version. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

[15] 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.

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

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

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

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

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

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

[22] http://donelaitis.vdu.lt/main.php?id=4&nr=7_2

[23] ParaSol corpus: http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/

# Appendix

## **B**ulgarian (MTE annotation):

BG: Вместо отговор Гандалф гръмогласно подвикна на коня си:

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
    <tok type=WORD>
      <orth>Вместо </orth>
      <disamb><base>вместо</base><ctag>RG</ctag></disamb>
      <lex><base>вместо</base><msd>Rg</msd><ctag>RG</ctag></lex>
      <lex><base>вместо</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>отговор</orth>
      <disamb><base>отговор</base><ctag>NCMS-N</ctag></disamb>
      <lex><base>отговор</base><msd>Ncms-n</msd><ctag>NCMS-N</ctag></lex>
    </tok>
<tok type=WORD>
<orth>Гандалф</orth>
<disamb><base>Гандалф</base><ctag>NPMS-N</ctag></disamb>
<lex><base>Гандалф</base><msd>Npms-n</msd><ctag>NPMS-N</ctag></lex>
</tok>
    <tok type=WORD>
      <orth>гръмогласно</orth>
      <disamb><base>гръмогласно</base><ctag>RA</ctag></disamb>
      <lex><base>гръмогласен</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
      <lex><base>гръмогласно</base><msd>Ra</msd><ctag>RA</ctag></lex>
    </tok>
    <tok type=WORD>
      <orth>подвикна</orth>
      <disamb><base>подвикна</base><ctag>VMIA3S</ctag></disamb>
      <lex><base>подвикна</base><msd>Vmia2s</msd><ctag>VMIA2S</ctag></lex>
      <lex><base>подвикна</base><msd>Vmia3s</msd><ctag>VMIA3S</ctag></lex>
      <lex><base>подвикна</base><msd>Vmip1s</msd><ctag>VMIP1S</ctag></lex>
    </tok>
<tok type=WORD>
<orth>на</orth>
<disamb><base>на</base><ctag>SP</ctag></disamb>
<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>коня</orth>
      <disamb><base>кон</base><ctag>NCMS-S</ctag></disamb>
      <lex><base>кон</base><msd>Ncms-s</msd><ctag>NCMS-S</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>коня</orth>
      <disamb><base>кон</base><ctag>NCMT</ctag></disamb>
      <lex><base>кон</base><msd>Ncmt</msd><ctag>NCMT</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>си</orth>
      <disamb><base>си</base><ctag>PX</ctag></disamb>
      <lex><base>си</base><msd>Ncns-n</msd><ctag>NCNS-N</ctag></lex>
      <lex><base>си</base><msd>Px------ys</msd><ctag>PX</ctag></lex>
      <lex><base>си</base><msd>Px---d--yp</msd><ctag>PX</ctag></lex>
      <lex><base>си</base><msd>Qgs</msd><ctag>QG</ctag></lex>
      <lex><base>си</base><msd>Qvs</msd><ctag>QV</ctag></lex>
      <lex><base>съм</base><msd>Vaip2s</msd><ctag>VAIP2S</ctag></lex>
    </tok>
```

```
<tok type=PUNCT><orth>:</orth><ctag>PERIOD</ctag></tok>
</chunk>
</chunkList>
</cesAna>
```

**Polish** [13]

PL: Zamiast odpowiedzieć hobbitowi, Gandalf krzyknął głośno do swego wierzchowca:

```
<!DOCTYPE cesAna SYSTEM "xcesAnaIPI.dtd">
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok>
        <orth>Zamiast</orth>
        <lex disamb="1"><base>zamiast</base><ctag>prep:gen</ctag></lex>
        <lex><base>zamiast</base><ctag>conj</ctag></lex>
</tok>
<tok>
        <orth>odpowiedzieć</orth>
        <lex disamb="1"><base>odpowiedzieć</base><ctag>inf:perf</ctag></lex>
</tok>
<tok>
        <orth>hobbitowi</orth>
        <lex disamb="1"><base>hobbit</base><ctag>subst:sg:dat:m3</ctag></lex>
        <lex><base>hobbitowy</base><ctag>adj:pl:nom:m1:pos</ctag></lex>
</tok>
<ns/>
<tok>
        <orth>,</orth>
        <lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
<tok>
        <orth>Gandalf</orth>
        <lex disamb="1"><base>gandalfa</base><ctag>subst:pl:gen:f</ctag></lex>
</tok>
<tok>
        <orth>krzyknął</orth>
        <lex disamb="1"><base>krzyknąć</base><ctag>praet:sg:m1:perf</ctag></lex>
        <lex><base>krzyknąć</base><ctag>praet:sg:m2:perf</ctag></lex>
        <lex><base>krzyknąć</base><ctag>praet:sg:m3:perf</ctag></lex>
</tok>
<tok>
        <orth>głośno</orth>
        <lex disamb="1"><base>głośno</base><ctag>adv:pos</ctag></lex>
</tok>
<tok>
        <orth>do</orth>
        <lex disamb="1"><base>do</base><ctag>prep:gen</ctag></lex>
</tok>
<tok>
        <orth>swego</orth>
        <lex><base>swój</base><ctag>adj:sg:gen:m1:pos</ctag></lex>
        <lex disamb="1"><base>swój</base><ctag>adj:sg:gen:m2:pos</ctag></lex>
            <lex><base>swój</base><ctag>adj:sg:gen:m3:pos</ctag></lex>
        <lex><base>swój</base><ctag>adj:sg:gen:n:pos</ctag></lex>
        <lex><base>swój</base><ctag>adj:sg:acc:m1:pos</ctag></lex>
        <lex><base>swój</base><ctag>adj:sg:acc:m2:pos</ctag></lex>
</tok>
<tok>
        <orth>wierzchowca</orth>
        <lex disamb="1"><base>wierzchowiec</base><ctag>subst:sg:gen:m2</ctag></lex>
        <lex><base>wierzchowiec</base><ctag>subst:sg:acc:m2</ctag></lex>
</tok>
</chunk>
</chunkList>
</cesAna>
```

### Lithuanian

LT: Užuot atsakęs Gendalfas garsiai riktelėjo žirgui:

**LT version** [22]:

<word="Užuot" lemma="užuot" type="prv., teig., nelygin. l.">
<space>
<word="atsakęs" lemma="atsakyti(-o,-ė)" type="dlv., teig., nesngr., veik. r, būt. k. l., neįvardž., vyr. g., vns., V.">
<space>
<word="Gendalfas" lemma="Gendalfas" type="nežinomas">
<space>
<word="garsiai" lemma="garsiai" type="prv., teig., nelygin. l.">
<space>
<word="riktelėjo" lemma="riktelti(-telia,-telėjo)" type="vksm., teig., nesngr., tiesiog. n., būt. k. l., vns., 3 asm.">
<space>
<word="žirgui" lemma="žirgas" type="dkt., vyr. g., vns., N.">
<sep=":">