COGNITIVE STUDIES | ÉTUDES COGNITIVES, 10
SOW Publishing House, Warsaw 2010

LUDMILA DIMITROVA[1], **VIOLETTA KOSESKA-TOSZEWA**[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

APPLICATION OF MULTILINGUAL CORPUS IN CONTRASTIVE STUDIES (ON THE EXAMPLE OF THE BULGARIAN-POLISH-LITHUANIAN PARALLEL CORPUS)

Abstract. In this paper we present applications of a trilingual corpus in language research. 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 one hand, it is interesting to note that two Slavic languages are compared to a Baltic language (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. The paper shortly describes the first electronic Bulgarian–Polish–Lithuanian experimental corpus, currently under development only for research. We also focus our attention on the morphosyntactic annotation of the parallel trilingual corpus according to the Corpus Encoding Standard: we present a review of the Part-of-Speech (POS) classification of the *participle* in the three languages — Bulgarian, Polish, and Lithuanian in comparison to another POS, the *adjective*. We briefly discuss tagsets for corpus annotation from the point of view of possible unification in the future with some examples.

Keywords: multilingual electronic corpora, parallel and comparable corpora, corpus annotation, lexical databases, multilingual electronic dictionaries.

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 (for example, 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 dictionaries and their subsequent expansion and update. Furthermore, it is interesting to note that two Slavic languages are compared to a Baltic language (Lithuanian). Thus we expect a new and interesting scientific problem in front of us and hope that our studies will find a wider application.

2 Multilingual Corpora — Brief Overview

In recent decades many multilingual corpora were created in the field of corpus linguistics, such as the MULTEXT corpus; the MULTEXT-East corpus, annotated parallel and comparable, an extension of the corpus MULTEXT; the ECI/MCI corpus; Oslo Multilingual Corpus; 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 [10]; Hong Kong bilingual parallel English-Chinese corpus of legal and documentary texts [7], etc.

MULTEXT corpus
Project MULTEXT ***Multilingual Tools and Corpora*** [8], is one of the largest EU projects in the domain of language engineering, whose goals are to develop standards and specifications for the encoding and processing of linguistic corpora, and to develop tools, corpora and linguistic resources embodying these standards. MULTEXT develops tools, corpora, and linguistic resources for a wide variety of languages, initially for seven West European languages Dutch, English, French, German, Italian, Spanish and Swedish, with more in later editions, including Bambara, Catalan, Kikongo, Occitan and Swahili. All Multext results are **made freely and publicly available** for non-commercial, non-military purposes.

European Corpus Initiative Multilingual Corpus I
The first release of the European Corpus Initiative, the Multilingual Corpus 1 (ECI/MCI: [6]), has 46 subcorpora in 27 (mainly European) languages. The total size of these is circa 92 million (lexical) words. The corpus has been available in digital form for scientific research **at a low a cost as possible** on CD-ROM since 1994, and is being distributed by ELSNET.
Contents: German newspaper texts (approximately 34 million words) from the Frankfurter Rundschau from July 1992 – March 1993; French newspaper texts (approximately 4.1 million words) from Le Monde, consisting of material from

September 1989, October 1989, and January 1990; extracts from the Leiden Corpus of Dutch, consisting of newspapers, transcribed speech, etc. (approximately 5.5 million words); parallel texts in English, French and Spanish from International Labor Organisation (ILO) "Official Bulletin, B Series" (approximately 5 million words); texts in Lithuanian (approximately 20 thousand words); scientific papers from Bulgarian journal "Science" (about 5 thousand words); etc.

MULTEXT-East annotated parallel, comparable, and speech corpora
The MULTEXT-East, a **freely available** standardised multilingual dataset for language engineering research and development, first developed in the scope of the EU MULTEXT-East project [2], an extension of the project MULTEXT. MULTEXT-East covers a large number of mainly Central and Eastern European languages, three languages of which: Bulgarian, Czech and Slovene, belong to the Slavic group. It includes the morphosyntactic specifications (EAGLES-based), defining the features that describe word-level syntactic annotations; medium scale morphosyntactic lexicons; and annotated parallel, comparable, and small speech corpora. The most important component of this dataset is the linguistically annotated parallel corpus consisting of Orwell's novel "1984" in the English original and translations.

Oslo Multilingual Corpus
Oslo Multilingual Corpus [21], which is an extension of the English-Norwegian Parallel Corpus (ENPC). The ENPC consists of text excerpts of approximately 10,000 to 15,000 words from fictional and non-fictional Norwegian and English original texts and their translations, amounting to a total of 200 texts, or 2.6 million words. German, Dutch and Portuguese translations were added for some of the texts. The texts are SGML-encoded and aligned at sentence level. The corpus is being extended on the German and French, to ensure equal representation of texts in Dutch, English, French, German, Norwegian and Portuguese. Due **to copyright restrictions**, the corpus is only available to researchers and graduate students at the universities in Oslo and Bergen.

Bulgarian–Polish corpus
The first Bulgarian–Polish corpus [3], currently under development only for research 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, contains approximately 5 million words. It consists of two parts: a parallel and a comparable corpus. This bilingual corpus supports the lexical database (LDB) of the first experimental online Bulgarian-Polish dictionary [4]. Some texts in the ongoing version of the parallel corpus are annotated at paragraph level.
Some texts of the Bulgarian comparable corpus are annotated at "paragraph" and "sentence" levels, according to Corpus Encoding Standard (CES) [9].

3 Trilingual 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 so far.

All collected texts in the corpus are texts published in and distributed over the Internet. The trilingual corpus comprises two corpora: parallel and comparable.

3.1 Bulgarian–Polish–Lithuanian parallel corpus

The BG–PL–LT parallel corpus contains more than 1 million words up to now. A part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of brochures of the European Commission, official documents of the European Union and the European Parliament, 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.

The development of methods allowing the construction of a multilingual parallel electronic corpus is a continuous process. We must stress that the parallel corpus of any three languages cannot be a sum of the individual corpora. It is obligatory to meet the condition of simultaneous accumulation of equivalent texts for all chosen languages. In other words, we cannot use ready monolingual corpora because the language material in them is accumulated to show the diversity and different levels (synchronic and diachronic) of a language system's development. Our aim should be to collect equivalent (nonetheless translated) language material, i.e. stylistically unambiguous, and contemporary. The diachronic level in the development of a language should not be taken into account. This level requires a different approach to the annonation of the material and is useless for the creation of multilingual dictionaries or electronic translation.

Another problem is the proportion of translated texts in the languages. 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 spoken relatively far from one another. 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. For example, Polish literature is more frequently translated to Bulgarian or Lithuanian than Bulgarian or Lithuanian to Polish. However, the translated texts in the three languages must be of comparable size.

We plan to annotate the BG-PL-LT parallel corpus according to the standards for morphosyntactic annotation of digital language resources. Due to typological differences (Bulgarian is analitical, Polish and Lithuanian synthetical) work during annotation of the parallel corpus will be difficult. Therefore, a condition that must necessaily be met is strict differentiation between form and content in the sentence of the natural language.

3.2 Bulgarian–Polish–Lithuanian comparable 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 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, *Соларис.* 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 trilingual parallel corpus, marked at paragraph level follow:

Bulgarian:
<p>Контейнерът се разтърси един-два пъти и завибрира непоносимо. Това трептене премина през всички изолирни обвивки, през въздушните възглавници и проникна в тялото ми. Зелените очертания на указателя се размазаха. Не усещах страх. Не бях долетял от толкова далеко, за да загина точно пред целта.</p>
<p>— Станция Соларис — обадих се аз, — станция Соларис. Станция Соларис! Направете нещо! Струва ми се, че губя равновесие! Станция Соларис, тук новият. Приеми!</p>

(**Лем — Соларис.** Преведе от полски Андреана Радева, Издателство Отечество София, 1980, c/o Jusautor, Sofia)

Polish:
<p>Zasobnik zadygotał raz i drugi, zawibrował nieznośnie, drżenie to przeszło przez wszystkie powłoki izolacyjne, przez powietrzne poduszki i wtargnęło w głąb mego ciała — seledynowy kontur wskaźnika rozmazał się. Patrzałem na to bez strachu. Nie przyleciałem z tak daleka, aby zginąć u celu.</p>

<p>— Stacja Solaris — powiedziałem. — Stacja Solaris, Stacja Solaris! Zróbcie coś. Zdaje się, że tracę stabilizację. Stacja Solaris, tu przybysz. Odbiór.</p>

(**Stanisław Lem, Solaris**, www.bookswarez.prv.pl)

***Lithuanian**:*
<p>Kapsulė suvirpėjo kartą, kitą, paskui ėmė vibruoti, šis nepakenčiamas virpulys perėjo per visas izoliacines plėveles, pripučiamas pagalves ir giliai įsismelkė į mano kūną. Žalsvas indikatoriaus kontūras išskydo. Aš nejutau baimės. Atskridau iš taip toli ne tam, kad žūčiau, pasiekęs tikslą.</p>
<p>— Stotis Soliaris, — ištariau. — Stotis Soliaris, stotis Soliaris! Darykite ką nors. Man rodos, aš netenku stabilizacijos. Stotis Soliaris, priimkite.</p>

(Soliaris, Stanislavas Lemas. Iš lenkų kalbos vertė Giedrė Juodvalkytė VILNIUS 1978
Stanislaw Lem, Solaris Wydawnictwo Literackie, Kraków, 1968. Vertimas į lietuvių kalbą, leidykla "Vaga", 1978)

//**ENG:** <p>The capsule was shaken by a sudden jolt, then another. The whole vehicle began to vibrate. Filtered through the insulating layers of the outer skins, penetrating my pneumatic cocoon, the vibration reached me, and ran through my entire body. The image of the dial shivered and multiplied, and its phosphorescence spread out in all directions. I felt no fear. I had not undertaken this long voyage only to overshoot my target!</p>
<p>I called into the microphone:</p>
<p>"Station Solaris! Station Solaris! Station Solaris! I think I am leaving the flight-path, correct my course! Station Solaris, this is the *Prometheus* capsule. Over."</p>

(http://www.lem.pl/cyberiadinfo/english/dziela/solaris/solarispl.htm, *Translated by Joanna Kilmartin and Steve Fox, Harcourt Brace*) //

4 Corpus annotation, POS classification and problems related to contrastive studies

Corpus annotation is the process of adding linguistic information in an electronic form to a text corpus [9], [11]. 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: непоносимо //EN adverb: unbearably, intolerably; EN adjective: unbearable, unendurable, intolerable, insufferable, insupportable, past/beyond endurance, not to be endured, beyond all bearing//:

(1) ***непоносимо*** //unbearable, unendurable, intolerable, insufferable, insupportable// → POS specifications: adjective, Gender: neuter, Number: singular, Definiteness: no.

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

(2) ***непоносимо*** //unbearably, intolerably// → 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, 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*.

The syntactic functions of the participle and the adjective cause their confusion as POS. In a sentence both participle and adjective have attributive and predicative function. One overlooks the fact that their meaning is quite different: a good illustration is the comparison of Polish and Bulgarian adjectives and participle taken from the electronic Bulgarian-Polish dictionary in working.

I паˈр — ен, -на, -но *adi.* parowy; **~на машиˈна** maszyna parowa

II паˈрен *part.* poparzony, sparzony

I повтоˈr | en, -na, -no *adi.* powtórny

II повтоˈрен *part.* powtórzony

I познаˈт *part. adi.* znany

II познаˈт, -и *m* znajomy m

These examples illustrate well the difference in meaning between Bulgarian adjectives and participles and prove that syntactic criteria are not sufficient to classify POS. Of big importance is the semantic perspective differentiating the meanings of participle and adjective in both languages although the forms I and II in Bulgarian are equal. The comparison of Bulgarian participle and adjective with their Polish correspondences underlines the role of language confrontation in solving theoretical problems in a natural language. A description concerning only Bulgarian or Polish would not be able to solve decisively the question of differentiation of the chosen POS. The language comparison in Bulgarian and Polish shows that the lack of differentiation of the two POS types is a sign of incompetence.

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 language use, distribution, quantity of forms, properties and functions are different. In contrast to English, for instance, where the participle are invariable, in the Slavic languages the forms of the participles are inflected (only adjectival participles). Participles are associated with verbal stem and contain information about the aspect, tense and valency of the finite forms of the respective verb. 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 writing 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.*)

A short explanation of the last example: the participles, used in the sentences, are related to the past tense forms to express simultaneity of the two states of the same agent.

Description:
The agent is speaking.
State 1: The speaker is watching.
State 2: The speaker is writing.
State1(Agent1) || State2(Agent1)

	BG	PL	LT
Agent	Говорещ /The speaker/	Mówiący /The speaker/	Kalbantysis /The speaker/
State 1	гледах	patrzyłem	žiūrėjau
State 2	*Пишейки* (**participle**)	*Pisząc* (**participle**)	*Rašydamas* (**participle**)

4.2 Participle and verb

It is important to emphasize that participles preserve some properties of the finite 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[1]. The valence of the Polish and Lithuanian participle is

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

the same as the valence of the finite verb form. The valency of passive participles changes according to the grammatical rules of the passive transformation, for instance:

PL: Jerzy czyta książkę. — Czytana przez Jurka książka.
LT: Jurgis skaito knygą. — Jurgio skaitoma knyga.
(EN: George is reading a book. — A book read by George.)

A comparison of the three languages shows that in Bulgarian a subordinate clause in past perfect tense corresponds to a participle construction in Lithuanian and Polish (only in the case when the events described in two parts of the subordinate clause refer to the one and the same agent):

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.*)

We stress that in Lithuanian a variant using past perfect tense is also possible:

LT: Jis *buvo paruošęs* (past perfect) pamokas, kaip pradėjo skaityti knygą.

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, verbal nouns, adverbs, and other lexical means it is clear that Polish can express also the same temporal meanings. In Lithuanian the quantity of finite verbal forms and participles is great. Lithuanian participles are distinguished by their ability to replace subordinate clauses in Polish and Bulgarian, for example (in **A** and **B**):

A. Case of expressing simultaneity of two states (or states and events), referring to two separate agents, for instance:

BG: И така — докато ти се *разхождаше* напред-назад с отворен чадър и *се тюхкаше*, че май ще вали — Пух пееше тази песничка...
PL: No i wtedy, kiedyś ty *przechadzał się* tam i z powrotem i *myślał*, czy będzie deszcz, czy nie będzie, Kubuś Puchatek zaśpiewał taką piosenkę...
LT: Taigi tau *vaikštinėjant* pirmyn ir atgal ir *svarstant*, ar lis, ar ne, Mikė Pūkuotukas traukė tokią dainelę...

	BG	PL	LT
Agent 1	ти /You/	ty /You/	tau /You/ (dat.sg)
State 1a	*докато* (conj) + *се разхождаше* (imperfect)	*kiedyś* (= pron *kiedy* + aggl. -ś) + *przechadzał się* (past)	*vaikštinėjant* (**participle**)
State 1b	*докато* (conj) + *се тюхкаше* (imperfect)	*kiedyś* (= pron *kiedy* + aggl. -ś) + *myślał* (past)	*svarstant* (**participle**)
Agent 2	Пух	Kubuś Puchatek	Mikė Pūkuotukas
State 2	*пееше* (imperfect)	zaśpiewał (past)	traukė (past)

Description:
Agent 1: You (Krzyś)
State 1: Krzyś przechadza się i myśli
Agent 2: Pooh
State 2: Pooh śpiewa (valid for BG and PL)
Event 2: Pooh zaśpiewał (valid for PL)
State1(Agent1) || State2(Agent2)

As we can see, participles are used only in the Lithuanian example: *vaikštinėjant* i *svarstant* (part of dative absolute (dativus cum participio) construction: *tau* (dat) + *participle* [*vaikštinėjant* and *svarstant*]). In Bulgarian and Polish constructions of subordinate clauses and lexical means such as *докато* (bg) and *kiedy* (pol) are correspondingly used.

B. Case of expressing "sequentiality-causality" of two states (or states and events), referring to two different agents:

BG: А когато някоя муха му *кацнеше* на носа, трябваше да я духа: "Пух!"
PL: Ile razy mucha *siadła* mu na nosie, nie mógł odpędzić jej łapką, tylko zdmuchiwał ją, ot tak: "puch, puch, puch!"
LT: Net musei *atsitūpus* ant nosies, jis negalėdavo jos nuvyti letena. O tiktai pūsdavo: "Pūk, pūk, pūk"

	BG	PL	LT
Agent 1	*муха*	*mucha*	*musei* (dat.sg.)
Event 1	кацнеше (imperfect)	siadła (past)	*atsitūpus* (**participle**)
Agens 2	Пух	Kubuś Puchatek	Mikė Pūkuotukas
State 2	*трябваше да духа* (imperfect+ infinitive construction)	*zdmuchiwał* (past)	*pūsdavo* (past iterative)

Description:
Agent 1: fly
Event 1: the fly landed
Agent 2: Pooh
State 1: Pooh blew
Event1(Agent1) → State1(Agent2)

Lithuanian uses the participle *atsitūpus*, which is part of dative absolute (dativus cum participio) construction: *musei* (dat.sg.) + *participle* [*atsitūpus*]). In Bulgarian and Polish to such constructions correspond subordinate clauses where the relation "sequentiality-causality" is based on context and knowledge (of speaker and listener) about reality. Furthermore, the statement's content is complicated by the contained repetition expressed by the form of past iterative: *pūsdavo.* Bulgarian and Polish use other means, for instance, PL: *ile razy*, BG: conjunction *а когато* or imperfect tense.

C. Case of expressing "sequentiality-causality" of two states (or states and events), referring to the one and the same agent, for instance:

BG: — Някой я е взел! — каза Ийори — … Колко Им Прилича Това! — прибави той след дълго мълчание.
PL: — Ktoś musiał mi go zabrać — powiedział Kłapouchy. — I jak tu mieć dla nich serce? — dodał po dłuższej chwili milczenia.
LT: — Kas nors bus pasiėmęs,- pasakė Nulėpausis. — Va kokie, — pridūrė ilgokai patylėjęs.

	BG	PL	LT
Agent	Ийори	Kłapouchy	Nulėpausis
State 1	след дълго мълчание	po dłuższej chwili milczenia	patylėjęs (**participle**)
Event 1	прибави (past: aorist)	dodał (past)	pridūrė (past)

Description:
Agent 1: Eeyore
State 1: Eeyore is silent
Event 1: Eeyore adds
State1(Agent1) → Event1(Agent1)

In this example to the Lithuanian participle correspond the following constructions: BG: *след дълго мълчание*, PL: *po dłuższej chwili milczenia* (preposition + noun/gerund).

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. The main 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[2]. (rarely: Namas yra mažas.) //*The house is small.* //

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

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

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

PL: Smakuje (verb) mi (to).
LT: Man skanu (adjective, neuter).
(EN:*I find it delicious.*)

BG: На село се живее добре (adverb).
PL: Dobrze (adverb) się mieszka na wsi.
LT: Gera (adjective, neuter) gyventi kaime.
(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 languages. Unification of adjectives and participles might be allowed 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.

4.4 Participle and adjective

Some participles possess adjectival properties, like gender, number and case (only valid for Polish and Lithuanian participles):

Singular forms:

Masculine forms: // BG: четящ // PL: czytający // LT: skaitantis //EN: reading//
Feminine forms: // BG: четяща // PL: czytająca // LT: skaitanti //EN: reading//
Neutral forms: // BG: четящо // PL: czytające // LT: *missing* //EN: reading//

Plural forms:

// BG: четящи // PL: czytający, czytające // LT: skaitantys, skaitančios //EN: reading//

Case forms:

// BG: *not valid for Bulgarian* //
// PL: czytający (nom.sg, nom.pl, acc.sg.-HUM), czytającego (gen.sg, acc.sg.+HUM), czytającemu (dat.sg) //
// LT: skaitantis (nom.sg), skaitančio (gen.sg), skaitančiam (dat.sg) //

Lithuanian has neuter participles (for example, sakoma "*they say*", palydavę "*it used to rain*", etc.), but they do not possess the category *case* or *number*. Since Lithuanian does not have neuter nouns, the neuter participles have received a new, predicative function in impersonal sentences: sakoma, naktį palydavę "*they say it used to rain at night*".

In Bulgarian and Lithuanian, participles, like adjectives, can also possess definite forms :

// BG: четящият (sg. m. full form), четящия (sg. m. short form) //
// PL: *not valid for Polish* //

// LT: skaitantysis (sg. m), skaitančioji (sg. f.) //
// EN: the reading//

The close relationship between participles and adjectives is only on a formal level. On a semantic level there are differences, see the list in 4.1. i 4.2.

The following examples show that the usage of Lithuanian participles has specific characteristics, which are not characteristic of adjectives:

(1) Attributive usage of present passive participle: *turimus* (acc. pl):

LT: Jis skaičiuoja *turimus* pinigus. "He counts the money owned."

(2) Attributive usage of future passive participle: *turėsimus* (acc.pl):

LT: Jis skaičiuoja *turėsimus* pinigus. "He counts foreseeable money (money that will be owned in the future)."

(3) Attributive usage of past active participle: *pasirodęs* (nom. sg)

LT: Vilko žvilgsnį patraukė tolumoj *pasirodęs* avynas. "The wolf's attention was attracted by the ram which (had) appeared in the distance. "

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 morpholexical 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:

въображение	=	Ncns-n
въображението	въображение	Ncns-y
въображения	въображение	Ncnp-n
въображенията	въображение	Ncnp-y

(въображение: *imagination, fancy*)

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 [1]. 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 of Bulgarian participles are discussed in detail in [5].

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.

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 [20] 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 [21]. 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 [20]) 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.

	POS	LT term	LT acronym
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.

A comparison between experimental annotations of the following sentence *"I felt no fear."* of the parallel corpus was performed:

BG: Не усещах страх.
PL: Patrzałem na to bez strachu.
LT: Aš nejutau baimės.

The tagsets for Polish (based on [13], [14], [18]) and Lithuanian ([20], [21]), used in the corresponding examples in the **Appendix**, follow:

For Polish:

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

The annotation of the Bulgarian text is done with MTE MSDs. 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**.

6 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*/.

The table shows the structure of the entry with headword ***spya*** /sleep/ in **BG-PL** LDB and a possible structure of entry with the same headword in **BG-LT** LDB:

The structure in **BG-PL** LDB:
<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:

```
<entry>
<hw>сп|яˈ</hw>
<pos>verb</pos>
<gram>imperfect</gram>
<conjugation><orth>-иˈш</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>
<alt><trans> apima mane miegas </trans></alt>
 </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, for machine translation, cross-lingual information retrieval, multilingual lexicon extraction, sense disambiguation, etc.

References

[1] Bulgarian Grammar. (1993). Д. Тилков, Ст. Стоянов, К. Попов. (Eds) *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, pages 315–319.

[3] Dimitrova, L., Koseska-Toszewa, V. (2009). Bulgarian-Polish Corpus. *International Journal Cognitive Studies | Études Cognitives.* 9, SOW, pages 133–141. ISSN 2080-7147.

[4] Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In Garabík (Ed.), *Metalanguage and*

Encoding scheme Design for Digital Lexicography. Bratislava, pages 36–47. ISBN 978-5-9900813-6-9.

[5] Dimitrova, L., Rashkov, P. (2009). A New Version for Bulgarian MULTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources.* Kyiv, Dovira Publ. House, 2009, pages 30–37. ISBN 978-966-507-252-2.

[6] ECI/MCI: `http://www.elsnet.org/ecilisting.html`

[7] 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`

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

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

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

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

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

[13] Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Constru-ction and Optimisation. Task Quarterly. 11, pages 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.* pages 140–150. ISBN 978-80-7399-745-8.

[16] Saloni, Z., Gruszczyński, W., Woliński, M. Wołosz, R. (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, pages 415. (In Lithuanian)

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

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

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

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

[22] Oslo Multilingual Corpus:
`http://www.tei-c.org/Activities/Projects/os01.xml`,
`http://www.tei-c.org/Activities/Projects/os01.xml`

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

Appendix

Bulgarian (MTE annotation):
BG: **Не усещах страх**.

```
<cesAna version=“1.0” type=“lex disamb”>
<chunkList>
<chunk type=“s”>
    <tok type=WORD>
      <orth> Не </orth>
      <disamb><base>не</base><ctag>QZS</ctag></disamb>
      <lex><base>не</base><msd>Qgs</msd><ctag>QGS</ctag></lex>
      <lex><base>не</base><msd>Qzs</msd><ctag>QZS</ctag></lex>
    </tok>
    <tok type=WORD>
      <orth>усещах</orth>
      <disamb><base>усещам</base><ctag>VMII1S</ctag></disamb>
    <lex><base>усещам</base><msd>Vmia1s</msd><ctag>VMIA1S</ctag></lex>
    <lex><base>усещам</base><msd>Vmii1s</msd><ctag>VMII1S</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>страх</orth>
      <disamb><base>страх</base><ctag>NCMS-S</ctag></disamb>
    <lex><base>страх</base><msd>Ncms-s</msd><ctag>NCMS-S</ctag></lex>
    </tok>
<tok type=PUNCT><orth>.</orth><ctag>PERIOD</ctag></tok>
</chunk>
</chunkList>
</cesAna>
```

Polish
PL: **Patrzałem na to bez strachu.**
PL version [13]:

```
<?xml version=“1.0” encoding=“UTF-8”?>
<!DOCTYPE cesAna SYSTEM “xcesAnaIPI.dtd”>
<cesAna version=“1.0” type=“lex disamb”>
<chunkList>
<chunk type=“s”>
<tok>
<orth>Patrzał</orth>
<lex disamb=“1”><base>patrzeć</base><ctag>praet:sg:m1:imperf</ctag></lex>
<lex><base>patrzeć</base><ctag>praet:sg:m2:imperf</ctag></lex>
<lex><base>patrzeć</base><ctag>praet:sg:m3:imperf</ctag></lex>
</tok>
<ns/>
<tok>
<orth>em</orth>
<lex disamb=“1”><base>być</base><ctag>aglt:sg:pri:imperf:wok</ctag></lex>
</tok>
<tok>
<orth>na</orth>
```

```
<lex><base>na</base><ctag>prep:loc</ctag></lex>
<lex disamb=“1”><base>na</base><ctag>prep:acc</ctag></lex>
</tok>
<tok>
<orth>to</orth>
<lex><base>to</base><ctag>subst:sg:nom:n</ctag></lex>
<lex disamb=“1”><base>to</base><ctag>subst:sg:acc:n</ctag></lex>
<lex><base>ten</base><ctag>adj:sg:nom:n:pos</ctag></lex>
<lex><base>ten</base><ctag>adj:sg:acc:n:pos</ctag></lex>
<lex><base>to</base><ctag>pred</ctag></lex>
<lex><base>to</base><ctag>conj</ctag></lex>
<lex><base>to</base><ctag>qub</ctag></lex>
</tok>
<tok>
<orth>bez</orth>
<lex><base>beza</base><ctag>subst:pl:gen:f</ctag></lex>
<lex><base>bez</base><ctag>subst:sg:nom:m3</ctag></lex>
<lex><base>bez</base><ctag>subst:sg:acc:m3</ctag></lex>
<lex disamb=“1”><base>bez</base><ctag>prep:gen:nwok</ctag></lex>
</tok>
<tok>
<orth>strachu</orth>
<lex disamb=“1”><base>strach</base><ctag>subst:sg:gen:m3</ctag></lex>
<lex><base>strach</base><ctag>subst:sg:loc:m3</ctag></lex>
<lex><base>strach</base><ctag>subst:sg:voc:m3</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb=“1”><base>.</base><ctag>interp</ctag></lex>
</tok>
</chunk>
</chunkList>
</cesAna>
```

Lithuanian
LT: **Aš nejutau baimės.**
LT version [21]:

```
<word=“Aš” lemma=“aš” type=“įv., vns., V.”>
<space>
<word=“nejutau” lemma=“nejusti(-nta,-to)” type=“vksm., neig., nesngr., tiesiog. n., būt. k. l., vns., 1 asm.”>
<space>
<word=“baimės” lemma=“baimė” type=“dkt., mot. g., vns., K.”>
<sep=“.”>
<p>
```