COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 13–20
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.002

LUDMILA DIMITROVA[1,A], **VIOLETTA KOSESKA**[2,B], **DANUTA ROSZKO**[2,C] & **ROMAN ROSZKO**[2,C]

[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[2]Institute of Slavic Studies, Polish Academy of Sciences Warszawa, Poland
[A]ludmila@cc.bas.bg ; [B]amaz@inetia.pl ; [C]{danuta,roman}.roszko@ispan.waw.pl

TRILINGUAL ALIGNED CORPUS — CURRENT STATE AND NEW APPLICATIONS

Abstract

This article describes current state of a trilingual parallel corpus consisted of texts in two Slavic (Bulgarian and Polish) and one Baltic language (Lithuanian). The corpus contains original literary texts (fiction, novels, and short stories) in one of the three languages with translations to the other two, and texts in other languages translated into Bulgarian, Polish, and Lithuanian. A part of the texts are aligned at the sentence level. The authors propose a semantic annotation of verbs appearing in these aligned texts that will facilitate contrastive studies of natural languages. A theoretical background for the proposed semantic annotation is briefly also discussed.

Keywords: aligned trilingual corpus, digital resources, event, Petri net theory, semantic annotation, state.

1 Introduction

It is interesting to compare two Slavic languages and one Baltic language by describing the similarities and differences between the formal means of the three languages. Every language has specific features which make it unique within the respective language group: Bulgarian belongs to the South Slavic subgroup, Polish to the West Slavic subgroup of the Slavic language family, whereas Lithuanian belongs to the Eastern Baltic group. A significant characteristic is the analytic character of Bulgarian, and the synthetic character of Polish and Lithuanian (with some analytic character, for example, word order in absolute constructions). In addition, Bulgarian possess several linguistic innovations in comparison with the other Slavic languages (a rich system of verbal forms, a definite article), which render its grammatical structure closer to that of English, the Neo-Latin languages, or Modern Greek than to the other Slavic languages. Other essential features include: high number of verbal forms in Bulgarian and Lithuanian, reduced number

of past tense forms in Polish; a strongly developed category of verbal aspect in Polish and Bulgarian, multiple aspect depending on the usage of a base stem for present, past and future tense in Lithuanian; postpositive definite article in Bulgarian. 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 (very rich case paradigm for nouns). Although Lithuanian has lost the neuter gender of nouns, its case system is richer than the Polish one. Such specific features can be illustrated with examples, extracted from a trilingual parallel corpus, and be studied and compared by means of a corpus.

The first presentation of the trilingual Bulgarian-Polish-Lithuanian corpus was announced in (Dimitrova, Koseska, Roszko, D. & Roszko R., 2009a), an extension of the corpus is presented in (Dimitrova, Koseska, Roszko, D. & Roszko, R., 2011).

The paper describes new important application in contrastive studies of a Bulgarian-Polish-Lithuanian parallel corpus, partially aligned at the paragraph or sentence level.

2 Description of the Bulgarian-Polish-Lithuanian Parallel Corpus

The parallel corpus consists of original texts in one of the three languages with translations in the other two, and texts in other languages translated into Bulgarian, Polish, and Lithuanian: several novels found in Internet or other available origins, texts of brochures and documents of the EC, EU, etc. Literary works include Stanisław Lem's "Solaris" with Bulgarian and Lithuanian translations, Vladas Braziūnas's "Yesterday is Tomorrow" with Bulgarian and Polish translations, translations of A. de Saint-Exupéry's "Le Petit Prince", M. Bulgakov's "Master and Margarita", Tolkien's "The Lord of the Rings", A. A. Milne's "Winnie-the-Pooh".

Some of the collected parallel texts are aligned manually at the paragraph level; others are aligned at the sentence level by software package TextAlign. Aligned text example from Lem's "Solaris" follows:

```
<tu tuid="0000000196">
   <tuv xml:lang="Polish">
      <seg>— Z Ziemi — odparłem wściekły.</seg>
   </tuv>
   <tuv xml:lang="Bulgarian">
      <seg>— От Земята — отговорих ядосано.</seg>
   </tuv>
   <tuv xml:lang="Lithuanian">
      <seg>— Iš Žemės, — atrėžiau įsiutęs.</seg>
   </tuv>
</tu>
<tu tuid="0000000197">
   <tuv xml:lang="Polish">
      <seg>— Słyszałeś może o niej?</seg>
   </tuv>
   <tuv xml:lang="Bulgarian">
      <seg>— Чувал ли си за нея?</seg>
```

```
    </tuv>
    <tuv xml:lang="Lithuanian">
      <seg>— Gal girdėjai tokią?</seg>
    </tuv>
</tu>
```

The results of alignment have been evaluated correspondingly, and errors detected in the aligned process have been corrected. The corpus is prepared not only for different applications in linguistics studies but also for preparation of a trilingual lexical database.

3 Applications of Trilingual Aligned Corpus in Contrastive Studies

Comparative and contrastive studies of Polish and Bulgarian, and of Polish and Lithuanian have been already conducted (Koseska, 2006), (Roszko, D., 2006), (Roszko, R., 1993, 2004), but no such studies exist for Bulgarian and Lithuanian at a similar high level.

The problems of the morphosyntactic annotation of the Bulgarian-Polish-Lithuanian corpus have already discussed (Dimitrova, Koseska, Roszko, D. & Roszko R., 2009b), and some of its applications in contrastive studies have presented in (Dimitrova, Koseska, Roszko, D. & Roszko, R., 2010).

Next, we want to point out other type of annotation, namely semantic annotation of a multilingual corpus.

4 Semantic Annotation of a Trilingual Aligned Corpus

The semantic annotation of a trilingual parallel corpus is a challenging research problem due to the lack of a uniform system of a semantic annotation (mark-up) for Bulgarian, Polish, and Lithuanian. The semantic annotation, presented and analyzed in this paper, is contained in the second approach of semantic theories (direct approach semantics) used in the Bulgarian-Polish Contrastive Grammar (Koseska, 2006).

Our semantic mark-up distinguishes quantificational meanings of names and predicates, and indicates aspectual and temporal meanings of verbs (Koseska & Mazurkiewicz, 2010). It relies on the formal net theory of processes, known as "Petri nets theory" and its application in a natural language (Mazurkiewicz, 1986), (Koseska & Mazurkiewicz, 1988).

A group of classifiers — **semantic classifiers**, which were not elaborated earlier, will be discussed here. These semantic classifiers were elaborated thanks of long standing work on the ***Contrastive Polish and Bulgarian Grammar with a semantic interlanguage***, the first in the world completed semantic contrast between these languages. The authors are aware that elaboration of semantic classifiers is not an easy task, and in order to achieve such a goal, one should **consistently distinguish between the form and the meaning**.

The notions of events, states and their configurations are understood here as in the **network-based description of time and aspect**, i.e. so that an event does not last in time (it begins, ends or interrupts states), while a state lasts and is begun or ended by an event.

For example, the notions of "*state*", "*event*" and their configurations are distinguished there as units of the interlanguage, as defined in the net theory. The meanings described using a formal logic theory are not only strictly defined, but can also be expressed in a formal way, and readily used in contrasting multiple languages. The meanings chosen for semantic annotation in this trilingual parallel corpus are based on those theories alone. It should be stressed that imperfective forms of a verb can express both "*a state*", and "*sequence(s) of states and events terminating with a state*", while perfective forms of a verb can similarly express both "*an event*", and "*sequence(s) of events and states terminating with an event*". As a result a precise distinction between a language form and its contents can be made, e.g. a perfective verb form has two meanings: "*an event*" or "*a sequence of events and states terminating with an event*". An imperfective verbal form also has two meanings: "*a state*" or "*a sequence of states and events terminating with a state*".

This is an innovation in elaborating sentence-level semantics in parallel corpora. For this reason, our semantic mark-up is manual. Hopefully it will raise the interest of computer scientists working on automatic methods for processing natural languages.

The proposed semantic annotation will facilitate contrastive studies of natural languages, and this in turn will verify the results of those studies, and will certainly facilitate human and machine translations.

The Table 1 shows examples from the trilingual corpus with the proposed semantic mark-up of verbs. In these examples the distinction between two meanings, both "*state*" (1. "*state*" or 2. "*sequence of states and events terminating with a state*"), and of an "*event*" (1. "*event*" or 2. "*sequence of events and states terminating with an event*"), is possible due to additional information in the sentence.

For instance, (see Table 1, example 1) Lithuanian *girdėjau* /I heard/ can express "*sequence of events and states terminating with an event*" thanks to its combination with a form of "past non-frequentative" *sušvilpė* /to whistle/ (in Polish corresponding to the noun *świst* /whiz/, in Bulgarian — the verbal noun *избръмчаване* /buzzing/), in Lithuanian, in addition, multiply quantified *aštuonis kartus* /eight times/.

In contrast, the Lithuanian *bėga* appearing in example 2 can express either "*a state*" or "*a sequence of events and states*". However, another verb, the verb *atvyksta*, cooperating with it, decides whether in a particular situation it is "*a sequence of events and states, finally ended with a state*". The theme of the present tense form of the verb *atvykti* /to arrive/ is perfective. In the Lithuanian language, the present tense forms with the perfective theme express meanings of general quantification. The collocation *kas naujas* (containing the interrogative-relative determiner *kas* /who/) cooperates with the mentioned Lithuanian forms. This determiner is exactly a typical exponent of habitual general quantification (habitual universality) in the Lithuanian language. The Polish and Bulgarian equivalents of the Lithuanian *kas najas*, i.e. *ktoś nowy* and *някой нов* respectively are ambiguous, and their concrete quantification meaning can be established only in context.

The example 3 shows that the compound forms: the Bulgarian *беше арестувана* and the Lithuanian *buvo suimta* unlike the Polish impersonal verb form

Table 1 Examples from the trilingual corpus with the proposed semantic mark-up of verbs

	Polish	Bulgarian	Lithuanian
	Stanislav Lem's "Solaris"		
1	**Usłyszałem** (*sequence of states and events terminating with an* EVENT) ośmiokrotnie powtórzony świst motorów elektrycznych, które **dociągały** (*sequence of states and events terminating with a* STATE) śruby.	**Чух** (*sequence of events and states terminated with an* EVENT) осмократното избръмчаване на електромоторите, които **дозатягаха** (*sequence of states and events terminating with a* STATE) болтовете.	**Girdėjau** (*sequence of events and states terminated with an* EVENT), kaip aštuonis kartus sušvilpė (*sequence of events and states terminating with an event*) elektriniai motorai, kurie **baigė veržti** (*sequence of events and states terminating with a* STATE) sraigtus.
2	Normalnie, kto żyw **biegnie** (*sequence of events and states terminating with a* STATE) na lotnisko, kiedy **przybywa** (*sequence of events and states terminating with a* STATE) ktoś nowy, i do tego jeszcze prosto z Ziemi.	Обикновено всичко живо **тича** (*sequence of events and states terminating with a* STATE) към летището, когато **пристига** (*sequence of events and states terminating with a* STATE) някой нов, и то направо от Земята.	Normaliai visi kas gyvas **bėga** (*sequence of events and states terminating with a* STATE) į nutūpimo aikštelę, kai **atvyksta** (*sequence of events and states terminating with a* STATE) kas naujas, be to, dar tiesiai iš Žemės.
	Michael Bulgakov's "Master and Margarita"		
3	Annuszkę **aresztowano** (EVENT) w chwili, kiedy **usiłowała wręczyć** (*sequence of states and events terminating with a* STATE) kasjerce w domu towarowym na Arbacie dziesięciodolarowy banknot.	Анушка **беше арестувана** (EVENT) в момента, когато **правеше опит да пробута** (*sequence of states and events terminating with a* STATE) на касиерката в универсалния магазин на Арбат банкнота от десет долара.	Anuška **buvo suimta** (EVENT) tuo metu, kai Arbato universalinėje parduotuvėje **mėgino įbrukti** (*sequence of states and events terminating with a* STATE) kasininkei dešimties dolerių banknotą.
	Paulo Coelho's "The Alchemist"		
4	Tu **postanowił spędzić** (EVENT) noc.	**Реши да пренощува** (EVENT) тук.	Vaikinas **nusprendžia** čia **praleisti** (STATE) naktį.

5	**Wprowadził** (EVENT) swoje owce przez rozpadającą się bramę i **zagrodził** (EVENT) wejście deskami tak, by w nocy **nie mogły się wymknąć** (sequence of states and events terminating with a EVENT).	**Вкара** (EVENT) овцете през разнебитената порта и я **залости** (EVENT) с няколко дъски така, че **да не могат да избягат** (sequence of states and events terminating with a EVENT).	**Suvaro** (STATE) avis į griuvėsius ir, kad šios per naktį **neišsilakstytų** (*sequence of states and events terminating with a event*), iš kelių lentų **padaro** (STATE) užtvarą.

aresztowano unambiguously specify a previous event in regard to the state of the statement. The Polish form *aresztowano* can express a "*state*", "*an event*" or "*a sequence of events and states*", and therefore it requires a context and/or a situation, in which one of the potential meanings becomes relevant, comp. *aresztowano* (event), kiedy *usiłowala wręczyć*, /(she was) arrested when (she was) trying to give/ and *aresztowano* ("*sequence of states and events, finally ended with and a state*"), kiedy *usiłowała wręczać* /(she was) arrested every time she tried to give/. An aspectual difference of the Polish verbs *wręczyć* and *wręczać* determines the meaning of the Polish aresztowano.

The following examples 4–5 show, the incompatibility of the used verbal forms is significant in Polish and Bulgarian on one hand and in Lithuanian on the other. And so the Polish *postanowił* (spędzić), *wprowadził* and *zagrodził*, and the Bulgarian *реши* (да пренощува), *вкара* and *залости* out of context unambiguously describe past events. The Lithuanian equivalents *nusprendžia* (*praleisti*), *suvaro*, *padaro* are forms of the present tense. The consequence of the poetic effort applied by the Lithuanian translator is a different network interpretation resulting from using the mentioned forms. Formal differences across the three languages can be underlined by the semantic annotation in the corpus, helping "trace" identical content. It is important only how THE SEQUENCES OF EVENTS AND STATES terminate: with an event or a state, and not how they began.

5 Training of Software Tools

As practice shows, software tools (programs) for certain automatic procedures such as automatic alignment, translations (human and machine), trained only with texts from different restricted thematic areas (newspaper articles, laws, medical and pharmacological literature, tourist brochures and guides, etc.), are not applicable enough for work with literary texts.

However, usage of literary texts from aligned multilingual corpora leads to improved performance, demonstrating that aligned multilingual corpora are very useful and valuable resources for such activity. These corpora comprise direct material for the evaluation of translations and their analysis helps improve the quality of both traditional human translation, and machine translation.

Furthermore, aligned corpora are successfully used as language materials for the training of translators. This is the advantage of parallel corpora in comparison with monolingual corpora.

6 Conclusion and further work

The described semantics mark-up of a trilingual corpus is still a work in progress. Our materials demonstrate well the connection between semantics and language confrontation in linguistics studies, which is impossible with monolingual corpora. We must remember that same or similar verbal forms in different languages may present different temporal situations.

Hence, the descriptions of temporal situations can be useful not only for comparison, analyzing, processing, or translating phrases in different languages, containing temporal dependencies, but also to distinguish verbal forms from temporal meaning in different languages (Koseska and Mazurkiewicz, 2010). Without understanding the meaning of temporal statements in various languages it is not possible to compare them or to create an adequate translated correspondence between them.

Various multilingual corpora are available via the Internet, for example, ParaSol corpus, OPUS corpus, a set of subcorpora (EMEA corpus, Europarl3, MultiUN, etc.) (Tiedemann, 2009). However, these multilingual corpora comprise only pairs of parallel text as bitexts, and most of these texts are texts of administrative documents of European Medicine Agency, European Parliament, United Nations, etc.

The above-represented trilingual corpus connects two Slavic languages with a Baltic language, serving as a valuable digital resource for linguists.

Aligned corpora are the most effective tools for the creation of contrastive grammars and bi- and multilingual dictionaries. The volume of parallel and aligned texts in the Bulgarian-Polish-Lithuanian trilingual corpus will increase. It is envisaged to make the trilingual aligned corpus available for a free access via Internet. The freely available online parallel and aligned texts are useful language materials not only for the training of translators, but for language learning in schools and universities.

References

Dimitrova, L., Koseska, V., Roszko, D., & Roszko, R. (2009a). Bulgarian-Polish-Lithuanian Corpus — Current Development. In C. Vertan, S. Piperidis, E. Paskaleva, & M. Slavcheva (Eds.), *Multilingual resources, technologies and evaluation for Central and Eastern European languages. Proc. of the International Workshop in conjunction with International Conference RANPL — 2009. Borovec, Bulgaria, 17 September 2009* (pp. 1–8). Bulgaria, Shoumen: INCOMA Ltd.

Dimitrova, L., Koseska, V., Roszko, D., & Roszko, R. (2009b). Bulgarian-Polish-Lithuanian Corpus — Problems of Development and Annotation. In T. Erjavec (Ed.), *Research Infrastructure for Digital Lexicography. Proc. of the MONDILEX Fifth Open Workshop within International Conference Information Society'2009, 14–15 October, 2009, Ljubljana* (pp. 72–86). Ljubljana: Informacijska družba.

Dimitrova, L., Koseska, V., Roszko, D., & Roszko, R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies | Études Cognitives*, *10*, 217–240.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2011). Bulgarian-Polish-Lithuanian Corpus — Recent Progress and Application. In D. Majchráková, & R. Garabík (Eds.), *NLP, Multilinguality. Proc. of the 6th International Conference SLOVKO'2011, Modra, Slovakia, 20–21 October 2011* (pp. 44–50). Brno: Tribun EU.

EMEA. (n.d.) Retrieved from `http://opus.lingfil.uu.se/EMEA.php`

Koseska-Toszewa, V. (2006). *Semantyczna kategoria czasu, Gramatyka konfrontatywna bułgarsko-polska* (Vol. 7). Warszawa.

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

Koseska V., & Mazurkiewicz A. (2010). *Time flow and tenses.* Warszawa: Slawistyczny Ośrodek Wydawniczy.

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

MultiUN (n.d.). Retrieved from `http://opus.lingfil.uu.se/MultiUN.php`

OPUS corpus (n.d.). Retrieved from `http://opus.lingfil.uu.se/`

ParaSol corpus (n.d.). Retrieved from `http://parasol.unibe.ch/`

Roszko, D. (2006). *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*, Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, R. (1993). *Wykładniki modalności imperceptywnej w języku polskim i litewskim.* Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, R. (2004). *Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim).* Warszawa: Slawistyczny Ośrodek Wydawniczy.

TextAlign (n.d.). Retrieved from `http://mt2007-cat.ru/index.html`

Tiedemann, J. (2009). News from OPUS — A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (Eds.) *Recent Advances in Natural Language Processing* (Vol. V: Proceedings, pp. 237–248). Amsterdam/Philadelphia: John Benjamins.

Publisher: Institute of Slavic Studies PAS & University of Silesia in Katowice