COGNITIVE STUDIES | ÉTUDES COGNITIVES, 12
SOW Publishing House, Warsaw 2012

LUDMILA DIMITROVA[1], **VIOLETTA KOSESKA-TOSZEWA**[2]

[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
ludmila@cc.bas.bg
[2]Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland
amaz@inetia.pl

BULGARIAN-POLISH PARALLEL DIGITAL CORPUS AND QUANTIFICATION OF TIME

Abstract

The paper presents the current state of the first Bulgarian-Polish parallel and aligned corpus, prepared in the frame of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" between the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences and the Institute of Slavic Studies, Polish Academy of Sciences, coordinated by L. Dimitrova and V. Koseska-Toszewa. In particular, problems related to tense quantification are also discussed.
Keywords: Bulgarian, Polish, digital language resources, parallel and aligned corpora, quantification of time.

1. Introduction

At the start of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" no bilingual Bulgarian-Polish and Polish-Bulgarian digital resources, corpora or dictionaries existed. Both languages belong to the Slavic language family: Bulgarian belongs to the South-Slavic and Polish — to the West-Slavic language family, and linguistic and contrastive studies of both languages can be carried out based on bilingual digital resources (corpora and dictionaries). To realize the goals of the project we started to gather texts in order to create a bilingual parallel corpus. Furthermore, the first Bulgarian-Polish parallel corpus serves as a main source of vocabulary for the Bulgarian-Polish digital dictionary.

2. Corpus annotation

Corpus annotation is the process of adding linguistic or structural information to a text corpus (Ide 1998), (Leech 2004)). One common type of annotation is the addition of labels — **tags** — that indicate the word class for the words in the text. This is the so called **part-of-speech tagging** (or POS tagging).

2.1. Morphosyntactic Annotation Systems for Bulgarian and Polish

The first **Bulgarian** digital resources (corpus, lexica, and morphosyntactic descriptions) are developed for MULTEXT-EAST[1] (MTE) project. The MTE corpus and the specific language resources are developed for all six languages of the project, Bulgarian, Czech, Estonian, Hungarian, Rumanian, Slovene, and English as a "hub-language" (Dimitrova et al. 1998). In addition to the multilingual MTE corpus other language-specific digital resources were developed; these include lexicons /lexica/, sets of language-specific rules and data (needed by the MTE software tools — segmenter, morphological analyser, POS-disambiguator, aligner), and sets of morpho-syntactic descriptors (MSD) for the six MTE languages. The MTE corpus serves as a model for corpus design and development: this model is being used in the design of the first Bulgarian-Polish corpus (Dimitrova, Koseska 2009).

The first Bulgarian digital corpora, developed for MTE corpus, include:

- Bulgarian translation of "1984" in CesANA-encoding — **word-level morphosyntactic mark-up** (undisambiguated lexical information for 156002 words, 156002 occurrences of MSD, and disambiguated lexical information for the 86020 words of the novel);
- Bulgarian-English **aligned corpus** (Bulgarian translation of "1984" aligned at sentence level with the English original),
- **comparable corpus** in two parts (Fiction.bg and News.bg), **sub-paragraph markups** (abbreviations, names, quotes, highlighted material, etc.).

The Bulgarian texts, containing about 300 000 wordforms, marked with SGML or their morpho-syntactic descriptions, in CES-format, were manually validated for paragraph and sentence boundaries (Dimitrova et al. 2005).

For the Polish language a morphosyntactic tagset, called the IPI PAN Tagset, was used for annotation of the IPI PAN Corpus, the first linguistically annotated monolingual corpus (Przepiórkowski 2006). A comparison of two morphosyntactic tagsets of Polish could be found in (Przepiórkowski 2009).

2.2 Structural Annotation of Corpora

Apart from POS tagging, there are other types of annotation, for example, structural annotation, which corresponds to different structural levels of a corpus or text. Written texts contain a number of different structural forms — divisions. Novels have a complex hierarchy and are divided into parts and chapters, newspapers are divided into sections, reference works — into articles, etc. The most common division in this hierarchy is the paragraph.

Some texts in the ongoing version of the Bulgarian-Polish corpus are annotated at paragraph level, other — are aligned and therefore annotated at the segment level. We use the standard markers: and for the paragraph's boundaries, <seg> and </seg> for the segment's boundaries. This annotation allows

[1]The EU COP Project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages,* (1995–1998), http://nl.ijs.si/ME/

texts in both languages (Bulgarian/Polish and vice versa) to be aligned at paragraph (<p> level) or at segment level in order to produces aligned bilingual corpora.

The <p> level allows us to draw a broader context in both languages. This means that we have the opportunity — thanks to the broader context — to study more precisely the meanings of word-forms in both languages.

This approach is more correct — we are not comparing "word" with "word", we compare word-forms in a broader context (level paragraph, segment, or sentence), which allows us to obtain a more adequate meaning of the word.

3. Aligned Bulgarian-Polish corpus

The Bulgarian-Polish corpus consists of two parts: a parallel and a comparable corpus. The parallel corpus contains literary texts and texts of documents in both languages, whereby the translation correspondence is one-to-one. Recently, a new sub-corpus of the Bulgarian-Polish corpus was prepared — the Bulgarian-Polish **aligned** corpus. The alignment is not a trivial task because of the role of the human translator: some sentences can be split, merged, deleted, inserted or reordered during the translation.

An aligned corpus is a multilingual (at least bilingual) parallel corpus. It is a result of the process of parallel text alignment that aims to produce a set of corresponding sentences (original and its translation(s)) in both or more parts of the parallel text (one of the most well-known example of parallel text alignment is inscribed on the famous Rosetta Stone). The result of the alignment of two parallel texts is a merged document, called bi-text, composed of both source- and target-language versions of a given text that retains the original sentence order. The software tools, generating bi-texts, are called alignment tools, or bi-text tools, which automatically align the original and translated versions of the same text. The tools generally match these two texts sentence by sentence.

A part of parallel Bulgarian and Polish texts were aligned by free available TextAlign software package and so the Bulgarian-Polish **aligned** corpus was produced. The TextAlign have applications in computer-assisted translation: it aligns bilingual texts without bilingual dictionaries, but the human editing is obligatory. The aligned bilingual texts are **annotated** for segment boundaries (here we use tags <seg> and </seg> for language pairs with corresponding number in the sequence of the aligned segments).

The aligned corpus includes texts of five Polish novels: Stanisław Lem's ***Solaris*** and ***Return from the Stars***, Ryszard Kapuściński's ***The Shadow of the Sun*** and ***Another Day of Life***, and Stefan Żeromski's ***Ashes,*** and their Bulgarian translations.

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

4. Linguistics studies and Bulgarian-Polish aligned corpus

The comparison of the Bulgarian and Polish material requires some explanations, which are very important for the future usage of the bilingual digital corpus.

A corpus itself proposes to the researchers more language material than examples presented in theoretical studies and articles.

The usages of given wordform in a wide context, like a set of sentences from bilingual aligned texts, show specific features of this wordform, such as gender and number for nouns; tense, aspect, and mode for verbs, etc. Tense is a meaning of the form, but has not been fully defined, see **the examples** about aorist (*аорист* in Bulgarian) and imperfectum (*имперфект*).

The parallel digital corpora are good tools with wide applications in contrast of semantics problems. In the current paper we will mainly analyze the quantification of time in both languages.

The reason we have chosen this problem is that the quantification is category of the sentence in logical sense, and is category of semantic structure of the sentence in linguistic sense.

The aligned Bulgarian-Polish corpus annotated at sentence-level and therefore represents the formal structure of the text, is an appropriate tool to contrast problems, typical for the semantic structure of sentence in both languages.

The corpus gives us a possibility to contrast such semantic categories like the different kinds of modality, the semantics relation "***antecedent — descendant of the time***", the semantic category of the time and especially quantification of the time, that is also category of the sentence, not only category of the verbal phrase (Koseska 1982), because of the obligatory confrontation of the state and the event (the basic/main semantic elements of the time) with the state of the utterance.

Logical quantification (in other words, quantification of the scope) can refer to names in the 1^{st} order logic as well as to predicates of the 2^{nd} order logic. Quantification of the time is closely related with the quantification of the predicates (Grzegorczyk 1972), (Koseska 1982). It is well-known that the quantifier converts every predicate into a sentence (in the logical sense).

Linguists still pose the question whether quantification refers to the semantics of aspect or to the semantics of tense. The book (Koseska, Gargov 1991) has examined the importance of aspect and tense as an entity and it won't be considered here. In our opinion the quantitative quantification is related to aspect and quantification of scope is related to tense (Koseska 2006).

The distribution of the aorist form of perfective and imperfective aspects in Bulgarian language shows that within the semantic structure of the sentence these forms are found only in the vicinity of the so-called unique quantifier (jota operator). They never appear next to existential or universal quantifiers, i.e. in the Bulgarian language there are **no** sentences such as:

* Той ходѝ там понякога (винаги).
* Той винаги (понякога) замѝна за София.
* Той винаги (понякога) се лекува'.

On the contrary, the imperfect form of imperfective aspect is found both with a universal and an existential quantifier, as well as with a unique quantifier. As the aorist form of both verb types may express only uniqueness we consider it a distinct and context-independent carrier of this quantificative meaning, while imperfect of

imperfective verbs is not a form with distinct quantification. This form may serve as a placeholder for a universal, existential quantification and can be found in contexts with singly-quantified temporal information, compare: *Тя седеше пред прозореца,* where the form — ***imperfect*** — can express universal quantification depending on further elaboration of quantification:

Тя винаги седеше пред прозореца
Ona zawsze siedziała przy oknie

with meaning: $(\forall s)\ P\ (s)$.

It can also express existential quantification, compare:

Тя понякога седеше пред прозореца
Ona czasem siedziała przy oknie

with meaning: $(\exists s)\ P\ (s)$.

As mentioned above, the imperfect form of imperfective verbs is found exceptionally also in a meaning analogical to ***praesens*** in contexts like:

В (точно) този момент той я обичаше
W tej chwili właśnie on ją kochał

or $(\imath s)\ P\ (s)$.

In this case the unique quantification affects previous condition, continued during the situation chosen as unique ("exactly in this moment").

In the Polish language, unlike Bulgarian the above sentences correspond only to sentences with praeteritum of imperfective verbs:

On dzisiaj chodził do szpitala.
On czasem chodził do szpitala.
On zawsze chodził do szpitala.

There is no doubt that the absence of such a semantic and distributional distinction between the verbal forms, existing in Bulgarian makes it more difficult for Poles who study Bulgarian to understand the subtle difference between the use of the aorist and imperfect of imperfective verbs.

Examples of the Bulgarian-Polish corpus show that the claims of some researchers to extinction (decrease) in Bulgarian language use of aorist forms of imperfective verbs are unfounded. It should be noted that in the western Bulgarian dialects the use of imperfect forms of imperfective verbs decreases as it happened already in Serbian (Koseska 1977).

The above facts point to the greater importance of quantificative expressions and other lexical resources of the Polish language, without which it would not be possible to express separate elements of temporality: states, events and their various combinations, see Mazurkiewicz 2008, Koseska, Mazurkiewicz 1988, 2010, Koseska 2006, or the examples:

On ***przed chwilą*** *wyszedł*
Той излезе ***преди малко***
On ***dzisiaj*** *zaglądał do szpitala*
Той ***днес*** *ходи до болницата*

(aorist of perfective verbs / praeteritum of perfective verbs and unique quantification of tense expressed by the adverb ***przed chwilą*** *(преди малко)* и ***dzisiaj*** *(днес).*

In conclusion, we emphasize that the Polish language handles the temporal expression of meanings more often than Bulgarian not only by the use of verbal forms, but also by lexical resources. The Bulgarian aorist of perfective and imperfective verbs holds a place for a single quantifier while the imperfect — for all quantifiers.

Although all elements of temporality can be expressed in both languages, it is noteworthy that some temporal meanings would not be better displayed in the Polish language without comparison to Bulgarian. This is a merit of the parallel corpus. It is also interesting to note that the temporal quantification is constant throughout the passage.

Neither the Bulgarian language, nor the parallel corpus has examples such as

* Катя винаги закъсня.

or

* Тя понякога закъсня.

The most common examples of temporal meanings and verbal forms for their utterance follow:

1. Polish praeteritum of perfective verbs // Bulgarian aorist form of perfective verbs — represent unique quantification of an event:

```
<tu tuid="0000000004">
  <tuv xml:lang="Polish">
    <seg>A koszulę wywalczyłem.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>А ризата си извоювах.</seg>
  </tuv>
</tu>

<tu tuid="0000000013">
  <tuv xml :lang="Polish">
    <seg>Stewardessa poprowadziła mnie między rzędami foteli na sam przód.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>Стюардесата ме поведе напред между редицата от кресла.</seg>
  </tuv>
</tu>

<tu tuid="0000001303">
  <tuv xml:lang="Polish">
    <seg>Zjechałem na dół, chyba kilka pięter, i wyszedłszy na ulicę dolnego poziomu
         zdziwiłem się, zobaczywszy znów nad sobą niebo.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>Спуснах се надолу, може би няколко етажа, и когато излязох на улицата
         на долното ниво, учудих се, че отново виждам небе.</seg>
  </tuv>
</tu>
```

2. Polish praeteritum of imperfective verbs // Bulgarian imperfect form of imperfective verbs — represent unique quantification of a state:

```
<tu tuid="0000000011">
  <tuv xml:lang="Polish">
    <seg>Chciał jeszcze coś powiedzieć.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>Той искаше да ми каже още нещо.</seg>
  </tuv>
</tu>

<tu tuid="0000001307">
  <tuv xml:lang="Polish">
    <seg>W białej portierni, przypominającej przewróconą wannę wielkoluda, siedział
         robot, pięknie stylizowany, półprzeźroczysty, o długich, delikatnych ramio-
         nach.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>В белия хол, приличащ на преобърната гигантска вана, седеше робот,
         чудесно стилизиран, полупрозрачен, с дълги, деликатни ръце.</seg>
  </tuv>
</tu>

<tu tuid="0000001311">
  <tuv xml :lang="Polish">
    <seg>Nie mogłem go znaleźć i nawet szukać nie próbowałem.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>Не можех да го намеря и дори не се опитвах да го търся.</seg>
  </tuv>
</tu>
```

3. Polish praeteritum form of imperfective verbs // Bulgarian aorist form of imperfective verbs — represent uniqueness of a set — unique quantification of states:

```
<tu tuid="0000000103">
  <tuv xml:lang="Polish">
    <seg>Głos z wewnątrz wypytywał nas, cośmy za jedni.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
    <seg>Един глас отвътре разпитва що за хора сме.</seg>
  </tuv>
</tu>
```

Outlook and future work. Next stage in the development of our parallel corpus is the formation of children's literature texts in a separate sub-corpus. Our goal

is that sub-corpus to serve as a comparative study of translation characteristics of prose and poetry.

6. Conclusion

The paper describes the three-million-word Bulgarian-Polish parallel aligned corpora. Parallel aligned corpora are a language resource for contrastive, translation and terminology studies, for development of machine translation and other multilingual technologies, such as tools for development of lexical databases and digital dictionaries. Special attention has been given to enabling further distribution of the corpora by encoding them in a standard format.

The web-presented language resources are oriented both to human and machine users and are available for a wide area of applications. The parallel bilingual corpora, aligned at paragraph or sentence level, annotated in accordance with international standards, provide samples of the word meaning and usage in various contexts, for instance during development of digital dictionaries. The parallel texts are successfully used as language materials for translator training as well as in education — for language learning in schools and universities. That is why online free-use parallel texts can also be a useful educational resource. In addition, such corpora are useful as a language material for bilingual lexical and terminological databases and on-line dictionaries development (Dimitrova, Panova, Dutsova 2009).

In conclusion, we emphasize that the parallel corpus enriches the theory with language material and corrects some theoretical setups, left unnoticed by scholars. It demonstrates also the important role of textual structural annotation, for example, throughout the whole passage there is a single-type quantification.

The described digital aligned bilingual resources have wide application in natural language processing. They are used successfully in multilingual software systems for automatic text segmentation (so called "***segmenters"***), that analyze and divide text into portions in a process of recognizing punctuation and separate words and performing morphological analysis and automatic text-to-speech transition.

Digital bilingual resources have also applications in machine learning: they are input data not only for "self-learning" software packages, so called „***taggers***", but also for "prediction" systems for a possible tag set (a type of specific characteristics) of unknown, not encountered in the text, or missing from the lexicons words.

The current results from machine-translation research demonstrate that the aligned bilingual resources make machine translation more adequate.

References

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

Dimitrova et al. 2005: Dimitrova, L., Pavlov, R., Simov, K., Synapova, L. Bulgarian MULTEXT-East Corpus — Structure and Content. J *Cybernetics and Information Technologies*. Vol. 5, num. 1, 67–73, 2005.

Dimitrova, L., Koseska, V. 2009: *Bulgarian-Polish Corpus*. Cognitive Studies/Études Cognitives. Vol. 9, SOW, Warsaw, 2009, 133-141.

Dimitrova, L., Panova, R., Dutsova, R. 2009: Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, 36–47.

Grzegorczyk, R. 1972: Wykładniki kwantyfikacji w polskimzdaniu. In: *Z Polskich Studiów Slawistycznych*, Warszawa, 13–19.

Ide, N. 1998: Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference, Granada, Spain.* 463–470.

Koseska-Toszewa, V. 1977: System temporalny gwar bułgarskich na tle języka literackiego. Wrocław, 1977. (In Polish)

Koseska-Toszewa, V. 1982: Semantyczne aspekty kategorii określoności/nieokreśloności (na materiale z języka bułgarskiego, polskiego i rosyjskiego). Wrocław, 1982 (In Polish).

Koseska-Toszewa, V. 2006: Semanticzna kategoria czasu, Gramatyka konfrontatywna bułgarsko-polska, vol. 7 , Warszawa, SOW, 210 pages. (In Polish)

Koseska-Toszewa, V., Gargov, G. 1991: The Semantic Category of Definiteness / Indefiniteness in Bulgarian and Polish. Warszawa, SOW, 140 pages.

Koseska-Toszewa, V., Mazurkiewicz A. 1988: Net representation of sentences in natural languages, Advances in Petri Nets, 1988, LNCS 340, Springer Verlag, 249–259.

Koseska-Toszewa, V., Mazurkiewicz, A. 2010: *Time Flow and Tenses*. SOW, Warsaw, 2010, 223 pages.

Leech, G. 2004: Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm

Mazurkiewicz, A. 2008: A formal description of temporality (Petri net approach). In: Iomdin, L., Dimitrova, L. (Eds.) Lexicographic Tools and Techniques. *Proceedings of the MONDILEX First Open Workshhop, 3–4 October 2008, Moscow.* 98–108. 80–88.

Przepiórkowski, A. 2006: The Potential of the IPI PAN Corpus. In *PSiCL*. Vol. 41, Poznań, Poland, 31–48.

Przepiórkowski, A. 2009: A Comparison of Two Morphosyntactic Tagsets of Polish. In: Koseska, Dimitrova, Roszko (Eds. 2009), Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, 29 June-1 July, Warsaw.* SOW, 138–144.