Language Resources – a Part of World Cultural Heritage

Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Science, Sofia, Bulgaria
ludmila@cc.bas.bg

Abstract. This article briefly reviews multilingual language resources for Bulgarian, developed in the frame of some international projects: the first-ever annotated Bulgarian MTE digital lexical resources, Bulgarian-Polish corpus, Bulgarian-Slovak parallel and aligned corpus, and Bulgarian-Polish-Lithuanian corpus. These resources are valuable multilingual dataset for language engineering research and development for Bulgarian language. The multilingual corpora are large repositories of language data with an important role in preserving and supporting the world's cultural heritage, because the natural language is an outstanding part of the human cultural values and collective memory, and a bridge between cultures.

Keywords: natural language, multilingual corpus, parallel corpus, aligned corpus, comparable corpus, annotation

1 Introduction

The multilingual digital resources are valuable multilingual dataset for language engineering research and development; they contribute to preserving and supporting the multilingual and multicultural world heritage, of which language is an outstanding part.

The multilingual corpora – great repositories of natural language data – are very useful for preservation and support of language heritage.

The main area of application of the corpora is the translation. The aligned parallel corpora are useful for many natural language processing (NLP) applications: in systems for machine-aided human translation, or for training of software for machine translation. They are prerequisite for contrastive studies or other linguistics research, and can also be used for retrieval of linguistic information, for producing concordances.

The digital resources with Bulgarian were developed in the frame of some international projects: the first Bulgarian digital lexical resources were developed in the frame of the EC project MULTEXT-East[1], Bulgarian-Polish and Bulgarian-Polish-Lithuanian corpora – in the frame of the joint research project[2] between

[1] EU COP Project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages*

[2] Semantics and Contrastive Linguistics with a Focus on a Bilingual Electronic Dictionary

Bulgarian Academy of Sciences and Polish Academy of Sciences, Bulgarian-Slovak parallel and aligned corpus – in the frame of the joint research project[3] between Bulgarian Academy of Sciences and Slovak Academy of Sciences.

2 MULTEXT-East Digital Resources

Bulgarian digital resources were developed for the first time under multilingual research projects MULTEXT-East (MTE). The MTE project is a continuation of MULTEXT project that produced the language resources for six western European languages (Dutch, English, French, German, Italian, and Spanish) and a freely available set of extensible, coherent, and language independent tools for natural language processing (NLP) [7].

MTE project developed significant language resources for six Central and Eastern European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as for English [1]. Three of these languages belong to the Slavic language group: Bulgarian, Czech, and Slovene. The MTE digital lexical resources include a dataset of language specific resources and multilingual MTE corpus, produced as a well-structured and lemmatized CES-corpus [6].

The MTE project has succeeded in providing foundational resources for work in language engineering in Bulgarian, for morphological, grammatical, semantic or other research, or as the basis for development of new applications in NLP [5]

2.1 MTE language-specific resources

The MTE language-specific resources contain two datasets:

1. Morphosyntactic specifications (coded as MorphoSyntactic Descriptions – MSDs) for the six CEE languages, as well as for English,
2. Data for use with the various annotation tools:

   - **Segmentation rules** include rules describing the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc.
   - **Special tokens** are data includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types, required by the segmentation tools.
   - **Morphological rules** are rules for the MTE languages, needed by the morphological tools. The rules provide exhaustive treatment of inflection and minimal derivation. Each lemma in the lexical lists used by the project is associated with its part(s) of speech and morphological rules.
   - **Lexicon** (**language-specific word-form lexical lists** covering at least the words appearing in the corpus) were developed by MTE consortium for the purposes of the corpus morpho-lexical processing. Each lexical list contains at

[3] Electronic Corpora – Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources

least 15,000 lemmas for use with the morphological analyser. Each lexicon entry includes information about the: inflected-form, lemma, part-of-speech (POS), and morphological specifications. A mapping from the morphosyntactic information contained in the lexicon to a set of corpus tags (used by the POS disambiguator) is also provided, according to the MULTEXT tagging model.

A lexicon entry has the following structure:

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; and comments, which are optional.

The example shows a **Bulgarian Lexicon** excerpt:

Word-Form	**Lemma**	**MSD**
обява	=	Ncfs-n
обяви	обява	Ncfp-n
обяви	обявя	Vmia2s
обяви	обявя	Vmia3s
обяви	обявя	Vmip3s
обяви	обявя	Vmm-2s
обявил	обявя	Vmpa-sma-n
обявя	=	Vmip1s
обявяват	обявявам	Vmip3p
обявят	обявя	Vmip3p

Here it is important to mention that the number of attributes, for example, for a **POS noun** is 10, with values for these attributes being 54; correspondingly, for a **POS verb** there are 13 attributes with 53 values. In Bulgarian lexicon there are, correspondingly, for nouns 47969 word-forms with 9891 lemmas, and for verbs – 266666 word-forms with 4140 lemmas. In total the Bulgarian word-forms are 29543 with 17567 lemmas.

2.2 MTE corpus

The MTE corpus is composed of three major parts:

(1) Multilingual Comparable Corpus

For each of the six CEE languages, the comparable corpus included two subsets of at least 100,000 words each: fiction (comprising a single novel or excerpts from several novels) and newspapers.

The data was comparable across the six languages, in terms of the number and text size. The entire multilingual comparable corpus was prepared in Ces format (Ces: Corpus Encoding Standard), manually or using ad-hoc tools, and was automatically annotated for tokenization, sentence boundaries, and part-of-speech annotation using the project tools.

(2) Multilingual Speech Corpus

MTE records a small corpus of spoken texts in each of the six languages.

(3) Multilingual Parallel Corpus

A parallel text is a text placed alongside its translation or translations. Large collections of bi- or multilingual parallel texts are parallel corpora.

The parallel MULTEXT-East corpus consists of six integral translations of George Orwell's "1984": besides the original English version, the corpus contains translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, and includes approximately 100,000 words per language. For each language, the corpus was marked and validated for paragraph and sentence boundaries. Each text is marked up for **structural data** (divisions, heads, footnotes, paragraphs, etc.) and for several sub-paragraph markups e.g. abbreviations, names, quotes, highlighted material, etc.

Structure of MTE parallel corpus:
There are four versions (different encoded documents) of MTE parallel corpus, corresponding to four different levels of annotation. For Bulgarian these versions are:

Original text – the Bulgarian translation of Orwell's novel 1984, includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations);

CesDOC-encoding of the Bulgarian text of the novel (SGML mark-up of the text up to the sentence-level), includes 1322 paragraphs, 6682 sentences;

CesANA-encoding, containing word-level morpho-syntactic mark-up (undisambiguated lexical information for 156002 words, 156002 occurrences of MSD, and disambiguated lexical information for the 86020 words of the novel);

CesAlign-encoding: Bulgarian-English aligned texts, containing links to the aligned sentences.

2.3 Aligned MTE corpus

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 alignment is not a trivial task because of the role of the translator: some sentences can be split, merged, deleted, inserted or reordered during the translation. 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.

The MTE multilingual aligned corpus consists of six bi-texts: all six translations of George Orwell's novel "1984" aligned with the English original. The alignment at the sentence level was performed automatically by means of some software packages and the obtained results were manually checked. For Bulgarian language the first parallel

aligned corpus has been produced as a part of the MTE aligned corpus by means of the Vanilla Aligner software. The aligned MTE corpus was used in many applications. The following examples (Table 1) illustrate some consecutive pairs of aligned sentences from the Bulgarian-English bi-text.

Table 1. Bulgarian-English aligned sentenses

1-1 aligned sentences	<Obg.1.1.3.4>Звукът от апарата (наричаше се телекран) можеше да бъде намален, но нямаше начин да се изключи напълно. <Oen.1.1.3.4>The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely.
1-2 aligned sentences	<Obg.1.1.23.16>Не беше много вероятно и въпреки това винаги, когато тя бе наоколо, той изпитваше странно чувство на неудобство, примесено със страх, дори враждебност. <Oen.1.1.24.16>That, it was true, was very unlikely.<Oen.1.1.24.17>Still, he continued to feel a peculiar uneasiness, which had fear mixed up in it as well as hostility, whenever she was anywhere near him.
2-1 aligned sentences	<Obg.1.1.39.1>Това ставаше винаги през нощта. <Obg.1.1.39.2> Арестуваха неизменно през нощта. <Oen.1.1.41.1>It was always at night -- the arrests invariably happened at night.

3 Description and Current Development of Slavic Language Bilingual Corpora

Multilingual parallel corpora are a basic resource for contrastive and terminology studies, for research and development of machine and human translation systems, language analysis, automatic term extraction, semantic analysis, supervised and unsupervised NLP tools training, etc. Three parallel corpora, Bulgarian-Slovak, Bulgarian-Polish, and Bulgarian-Polish-Lithuanian, are currently developed as language resources for such activities. All collected texts in these corpora are texts published in and distributed over the Internet, so copyright issues for the texts are not a concern.

The aligned at the paragraph and sentence level parallel corpora give more correct approach – we are not comparing "word" with "word", we compare word-forms in a broader context, which allows us to obtain the word's meaning. The aligned corpora will be very valuable resources for machine translation research.

3.1 Specific Features of Bulgarian, Polish and Slovak

These languages exhibit some specific features, occurring repeatedly in several categories. At first, different orthography traditions – the corpora are dataset of written languages and the orthography forms an inseparable part of language analysis. Another significant feature is the analytic character of Bulgarian, and the synthetic character of Polish and Slovak. Old-Bulgarian had an elaborate case system but in the process of evolution from a synthetic (inflectional language) to an analytic (flectional)

language Bulgarian has lost most of the traditional old Slavic case system. Bulgarian case forms were replaced with combinations of different prepositions with a common case form. Bulgarian has a grammatical structure closer to English or the Neo-Latin languages than other Slavic languages. Bulgarian indicates also some innovations such as a rich system of verbal forms, and a definite article that is morphological indicator of the grammatical category determination (definiteness) – one of the most important grammatical characteristics of the modern Bulgarian language, whereas the other Slavic languages lack the definiteness attributes altogether.

3.2 Bulgarian-Slovak Corpus – Parallel and Aligned

For many of the pair languages, so called low and medium density languages, there are no multilingual or bilingual resources easily available for scientific community. The parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus is currently developed in the framework of the research project between IMI–BAS and ĽŠIL–SAS, coordinated by L. Dimitrova and R. Garabík.

The corpus currently contains translations of fiction in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak. The main part of parallel corpus contains texts in other languages translated into both Bulgarian and Slovak. The corpus consists of two subcorpora: direct and translated. The direct Bulgarian–Slovak parallel subcorpus consists of original texts in Bulgarian, such as novels and short stories by Bulgarian writers and their translation in Slovak, and original texts in Slovak, such as literary works by Slovak writers and their translation in Bulgarian. The translated Bulgarian–Slovak parallel subcorpus consists of Bulgarian and Slovak translations of literary works in third language [2].

The Bulgarian–Slovak corpus contains parallel texts, aligned at the sentence level. To align the text on sentence level the Hunalign software was used. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of the alignment; however, no such dictionary has been available for the use with the corpus. The corpus contains 376 200 words in parallel texts, aligned at the paragraph level and at the sentence level. The set of aligned texts includes Bulgarian novels: Dimitar Dimov's Doomed Souls (*Осъдени души*) and Pavel Vezhinov's The Barrier (*Бариерата*) and their Slovak translations, the novel of Slovak writer Klára Jarunková The silent wolf's brother (*Brat mlčanlivého vlka*) and its Bulgarian translation, the Slovak and Bulgarian translations of Jaroslav Hašek's The Good Soldier Švejk. Some aligned sentences of the Slovak and Bulgarian translations of Hašek's novel (without a dictionary) follow:

„ Má ich tam byť dvanásť , " povedal Švejk , keď si upil.
- Трябваше да са дванадесет - рече Швейк , като отпи.
1.21626
„ Prečo myslíte dvanásť ? " opýtal sa Bretschneider.
- Защо пък дванадесет ? - запита Бретшнайдер.
1.75385

„ Aby to išlo do počtu , do tucta , lepšie sa to ráta a na tucty je to vždy lacnejšie , “ odpovedal Švejk.
- За по - лесно , като са дузина , по - лесно се броят , пък и на дузини всичко е по - евтино - отговори Швейк.
1.29277

A dialogue box and some concordances of Slovak noun follow in the next figures:

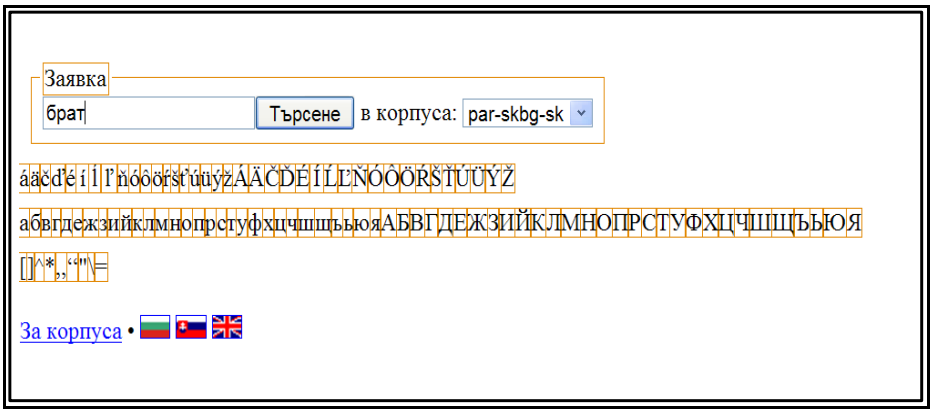

Fig. 1. Web search interface – a dialogue box in Bulgarian

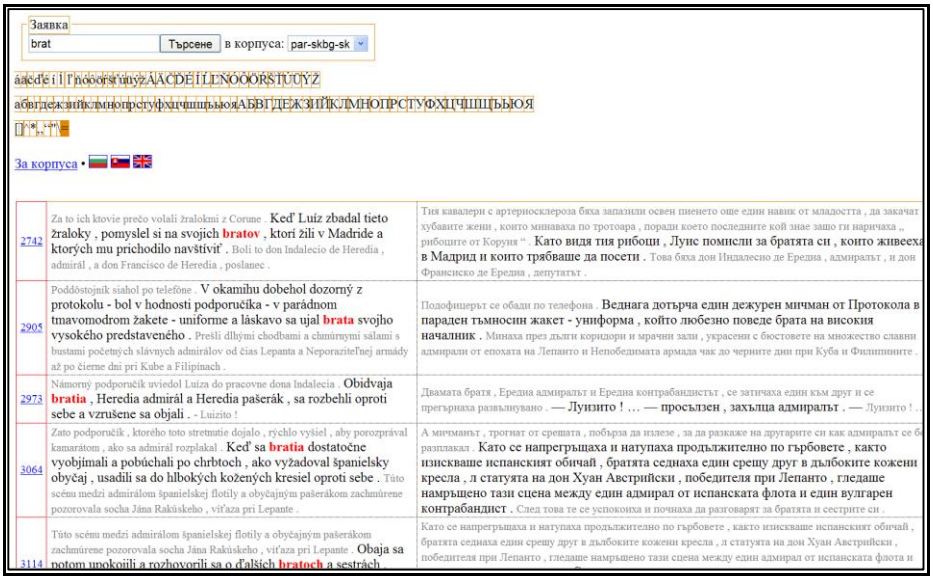

Fig. 2. Some concordances of the Slovak noun brat in the corpus

3.3 Bulgarian-Polish Corpus – Parallel, Aligned and Comparable

The first parallel Bulgarian-Polish corpus is currently developed under the joint research project "Semantics and Contrastive linguistics with the focus on a bilingual electronic dictionary" between IMI-BAS and ISS-PAS, coordinated by L. Dimitrova and V. Koseska. The corpus is built with the main purpose to ensure the selection of the entries for the first experimental electronic Bulgarian-Polish dictionary. The texts were collected concurrently and do not have a connection with national monolingual or other corpora. The corpus contains more than 3 million words [3].

A part of the parallel texts is annotated at paragraph level. A small part of the corpus is currently aligned at sentence level and forms so-called aligned Bulgarian-Polish corpus. The free available TextAlign software package was used.

The aligned corpus includes texts of 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. An example from aligned at sentence level Lem's novel "Return from the Stars" follows:

```
<tu tuid="0000000163">
  <tuv xml:lang="Polish">
     <seg>Robot rozłączył się i nie zdążyłem go spytać, gdzie mam szukać tego kalstera.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
     <seg>Роботът се изключи и аз не успях да го запитам къде да търся калстера.</seg>
  </tuv>
</tu>
<tu tuid="0000000164">
  <tuv xml:lang="Polish">
     <seg>Nie miałem zielonego wyobrażenia, jak wygląda.</seg>
  </tuv>
  <tuv xml:lang="Bulgarian">
     <seg>Нямах никакво понятие как изглежда.</seg>
  </tuv>
</tu>
```

3.4 Bulgarian-Polish-Lithuanian Parallel Corpus

The first Bulgarian-Polish-Lithuanian corpus is currently collected under the abovementioned research project between IMI-BAS and ISS-PAS for research purposes [4]. It is interesting to note that two Slavic languages are compared to a Baltic one – Lithuanian. Furthermore, the three languages are marginally present in the EU because of the later accession of these three countries to the EU.

The corpus contains more than 3 million words so far. We note that a big problem arose when we started to compile the corpus due to the mismatch in proportion of translated texts in the three languages. It turned out that it is extremely difficult to find electronic texts of translations from Bulgarian to Lithuanian or vice versa. That's why

we assumed that the Polish language would build “a bridge” between Bulgarian and Lithuanian.

The recent result of the joint work is a small, aligned Bulgarian-Polish-Lithuanian corpus. We used the TextAlign software package. The TextAlign have applications in computer-assisted translation: it aligns bilingual texts without bilingual dictionaries, but the human editing is obligatory. At the first stage we used this tool to align the original text in Polish, for example Stanislav Lem’s Solaris, and its Bulgarian translation. At the second stage the procedure is repeated with the input pair being the original Polish text and its Lithuanian translation. At the third stage, after a comparison of the two output bitexts, Polish-Bulgarian and Polish-Lithuanian, we end up with a sequence of triples: a sentence in Polish and its translations in Bulgarian and Lithuanian.

The following example presents an excerpt from the aligned at the sentence level texts of Stanislav Lem’s Solaris:

```
<tu tuid="0000000010">
        <tuv xml:lang="polish">
          <seg>Wzrok przywykał do ciemności.</seg>
        </tuv>
        <tuv xml:lang="bulgarian">
          <seg>Очите ми свикваха с тъмнината.</seg>
        </tuv>
            <tuv xml:lang="lithuanian">
          <seg>Akys priprato prie tamsos.</seg>
        </tuv>
      </tu>
      <tu tuid="0000000011">
        <tuv xml:lang="polish">
          <seg>Widziałem już seledynowy kontur jedynego wskaźnika.</seg>
        </tuv>
        <tuv xml:lang="bulgarian">
          <seg>Вече различавах светлозелените контури на универсалния указател.</seg>
        </tuv>
            <tuv xml:lang="lithuanian">
          <seg>Jau išskyriau žalsvus universalaus indikatoriaus kontūrus.</seg>
        </tuv>
      </tu>
```

4 Applications of digital language resources

The web-presented language resources are oriented both to human and machine users and are available for a wide area of applications.

The main area of application of the corpora is translation. The new developments are mainly in the direction of computer means to support the translators in their activities.

The digital bilingual resources are widely applicable to the contrastive studies of Slavic languages. They are valuable resources for machine translation research, and can be also used as a translation database and language learning materials for training of translators – human and programming tools. The uploaded on the web bilingual

corpora, aligned at the paragraph or the sentence level, could be used successfully in education.

In addition, the aligned corpora are the best resource for the development of bi- and multilingual lexical or terminological databases and different kinds of digital dictionaries. There one could find and extract many examples of a word's usage because the corpus provides samples of the word's meaning and usage in a wide context. Recently, the aligned corpora serve as a basis for development of new applications in multilingual digital libraries.

5 Conclusion

Parallel corpora are the most effective means for the creation of bi- and multilingual dictionaries and contrastive grammars. This is of great importance not only for language confrontation, but also for the typology of the studied languages. One has to remember that parallel corpora comprise direct material for the evaluation of translations and their analysis will bring out the improvement of the quality of both traditional, human translation, and machine translation. Besides, texts extracted from parallel or aligned corpora prove the necessity of evaluating translations: it is common that in translation words get omitted or word meanings get changed.

The parallel and aligned corpora are successfully used as language materials for the training of translators, as well as in education – for language learning in schools and universities. That is why online free-use parallel texts are also useful educational resource.

References

1. 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: COLING-ACL '98. Montréal, Québec, Canada, pp. 315–319. (1998)
2. Dimitrova, L., Garabík, R.: Bulgarian-Slovak Parallel Corpus. In: 6th International Conference NLP, Multilinguality. Bratislava. (2011) (be appear)
3. Dimitrova, L., Koseska, V.: Bulgarian-Polish Corpus. J. Cognitive Studies/Études Cognitives. v. 9, SOW, Warsaw, pages 133–141. (2009)
4. Dimitrova, L., Koseska, V., Roszko, D., Roszko, R.: Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus. J. Cognitive Studies/Études Cognitives. v. 10, SOW, Warsaw, pages 217–240. (2010)
5. Dimitrova, L., Pavlov, R., Simov, K., Sinapova, L.: Bulgarian MULTEXT-East Corpus – Structure and Content. J. Cybernetics and Information Technologies. v. 5, n. 1, BAS, Sofia, pages 67–73. (2005)
6. Ide, N., Bonhomme, P., and Romary, L.: XCES: An XMLbased Encoding Standard for Linguistic Corpora. In: 2nd International Language Resources and Evaluation Conference. Paris: ELRA, pages 825–830. (2000)
7. Ide, N., Veronis, J.: Multext (multilingual tools and corpora). In: COLING'94. Kyoto, Japan, pp. 90–96 (1994)