Bulgarian–Slovak Parallel Corpus

Ludmila Dimitrova[1] and Radovan Garabík[2]

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
ludmila@cc.bas.bg
[2] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
garabik@kassiopeia.juls.savba.sk

Abstract. The paper presents a currently collected parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus, developed in the framework of the joint research project between IMI–BAS and ĽŠIL–SAS. The texts in the ongoing version of the corpus are automatically aligned at the sentence level. The whole corpus is oriented towards emphasizing the applicability of the digital bilingual data for computerized natural language processing, but also as a source of human readable information. The paper describes the structure of the corpus. Other existing bilingual resources for Slovak and Bulgarian in combination with other languages are briefly mentioned as well.

1 Introduction

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

2 Short comparison of the Bulgarian and Slovak languages

We will mention here only the differences that result from different orthography traditions, because we are primarily dealing with the written language, where the orthography forms an inseparable part of language analysis. The most visible (but from linguistic point of view rather superficial) difference is the writing system – Bulgarian uses the Cyrillic alphabet, Slovak – Latin with additional letters with diacritical marks.

There are some features specific for both languages, which have an influence on the language grammar repeatedly in several different phenomena, and which we will summarise here.

- First significant feature is the analytic character of Bulgarian, and the synthetic character of Slovak.
- In the process of evolution of Bulgarian from a synthetic, inflectional language, to an analytic language, case forms were replaced by combinations of different prepositions with a common case form. Bulgarian has lost most of the traditional Slavic case system.

- Bulgarian exhibits several linguistic innovations in comparison to the other Slavic languages, namely a rich system of verbal forms, and has a grammatical structure closer to English or the Neo-Latin languages than Slovak. The definite article is one of the most important grammatical characteristics of the Bulgarian language which sets it off the rest of the Slavic languages. The definite article is a morphological indicator of the grammatical category of determination (definiteness). It is not just a simple suffix, nor is it a particle, but a word-forming morpheme, which is placed at the end of words in order to express definiteness, knowledgeability, familiarity, conversance. In Bulgarian, nouns, adjectives, numerals, and full-forms of the possessive pronouns and participles can acquire an article. Slovak, as most of the other Slavic languages, lacks the definiteness altogether.
- For Slovak, a special feature of masculine nouns exists, the animate feature of masculine gender. This feature is variously analysed either as a subattribute of masculine gender where the words belonging to the masculine gender split into two categories, or the Slovak is sometimes said to have two different masculine genders, the animate and inanimate one. For Bulgarian there is no animate category at all.

3 Standards and models for corpora encoding

In our work on the corpus we used our own experience in the area of construction of language resources and technologies and the widely accepted international standards and models [IBR00]. We used some of the basic widespread technologies and software packages for annotation and usage of digital bilingual resources.

The problem of reusability of the digital language resources has been discussed so many times. It is well known that the development of such resources is a sustained process that is frequently repeated again and again. Some language resources are not available in a common standard usable format or their distribution is hindered due to intellectual property rights, diverse and poorly documented encodings and other reasons. We tried to use open standards in our corpus compilation process and internal data formats and to use simple textual representation of the data wherever possible.

The multilingual corpora are parallel and comparable. A parallel corpus is defined as a corpus comprising source texts and their translations. The parallel corpora (similar to the dictionaries) can be bilingual or multilingual. A comparable corpus is defined as a corpus comprising not-identical, but similar in their content texts in one or more languages. It can be defined as a set of texts, collected independently from their languages and connected on the basis of their content, domain and communicative function.

3.1 Bulgarian parallel corpora

The first multilingual parallel corpus with Bulgarian has been produced as a part of the MULTEXT-East corpus. It consists of George Orwell's *1984* in English and its translations into six Central and East European languages (CEE): Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian [DEI+98], with Slovak as a later addition. Recently, bilingual Bulgarian-Polish [DK09] and trilingual Bulgarian-Polish-Lithuanian [DKRR10] corpora have been constructed.

3.2 Slovak parallel corpora

The first bilingual parallel corpus has been the Slovak-Russian parallel corpus, produced as a common project of ĽŠIL and Saint-Petersburg State University [R. 06], accessible through a web interface. It was later followed by the parallel Slovak-French corpus [VG07], using very similar structure and interface. Both of these corpora are lemmatised and POS tagged.

The Slovak-Czech parallel corpus contains mostly translations of fiction between Czech and Slovak (in both directions), with a small part of independent translations from other languages (75 % English, the rest German, Polish, Italian, French and Ancient Greek) into both Czech and Slovak. Both Czech and Slovak parts of the corpus are morphologically anaysed and lemmatised with the *morče* software, the Czech part uses the tagset used in the Czech National Corpus.

The Slovak-English corpus consists of original English language fiction and their Slovak translations. The texts were obtained from various sources; the bulk of the Slovak translations were already collected in the Slovak National Corpus, some of them were scanned, OCRed and then proofread. A small amount of OCR-induced errors remained in the texts, but presumably, it will not have significant effects on the overall corpus quality. The English texts are lemmatized and POS-tagged, using the TreeTagger software [Sch97].

All the parallel corpora use the Hunalign software [VNH+05].

4 Structure of the Bulgarian-Slovak/Slovak-Bulgarian corpus

The corpus currently contains translations of fiction in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak. The main part of the parallel corpus contains texts in other languages translated into both Bulgarian and Slovak.

The parallel 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 translations in Slovak, and original texts in Slovak, such as literary works by Slovak writers and their translations in Bulgarian.
- The translated Bulgarian–Slovak parallel subcorpus consists of Bulgarian and Slovak translations of literary works in the third language.

The Bulgarian–Slovak corpus contains parallel texts, aligned at the sentence level. It is composed of translations of literary works (mainly fiction) in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak (or translations into both of these languages from a third language).

To align the text on the sentence level, we use the Hunalign software. 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 first version of 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 *Осъдени Души* (Doomed Souls) and Pavel Vezhinov's *Бариерата* (The Barrier) and their Slovak translations (*Odsúdené duše* and *Bariéra*), the novel of

Slovak writer Klára Jarunková *Brat mlčanlivého vlka* (The silent wolf's brother) and its Bulgarian translation (*Братът на мълчаливия вълк*), and the Slovak and Bulgarian translations of Jaroslav Hašek's *The Good Soldier Švejk*.

5 Format of the texts and search

```
<doc lang="sk" origlang="sk"
     biblio="Odsúdené duše,
             Slovenský spisovateľ,
             Bratislava 1976,
             preklad Milan Topoľský">
<s id="1" link="1">
Odsúdené      odsúdený      Gtfp1x 12
duše          duša          Ssfp1  04
</s>
<s id="2" link="2">
Dimitār       dimitār       %      01
Dimov         dimov         %      24
</s>
<s id="3" link="3">
Prvá          prvý          Nafs1  02
časť          časť          Ssfs1  03
Koniec        koniec        Ssis1  04
jedného       jeden         Nfns2  04
dobrodružstva dobrodružstvo Ssns2  01
</s>
```

Fig. 1. Example of Slovak vertical file, wordforms with lemmas, morphosyntactic tags and number of disambiguation possibilities, sentences with links into Bulgarian text.

The structural annotation allows the texts in the two languages (Bulgarian/Slovak and vice versa) to be aligned at the corresponding level in order to produce aligned bilingual corpora. Alignment means the process of relating pairs of words, phrases, sentences or paragraphs in the texts in different languages which are translation equivalent. Currently, the corpus is automatically aligned at the sentence level without the help of a bilingual dictionary. In order to achieve reasonable quality of the corpus, the alignment should be as precise as possible. The project aims to create a small (several thousand words) dictionary suitable for automatic alignment (i.e. the dictionary will not be meant to be a general dictionary for human consumption).

The access to the recent version of the corpus is provided via a simple web interface (Fig. 2).

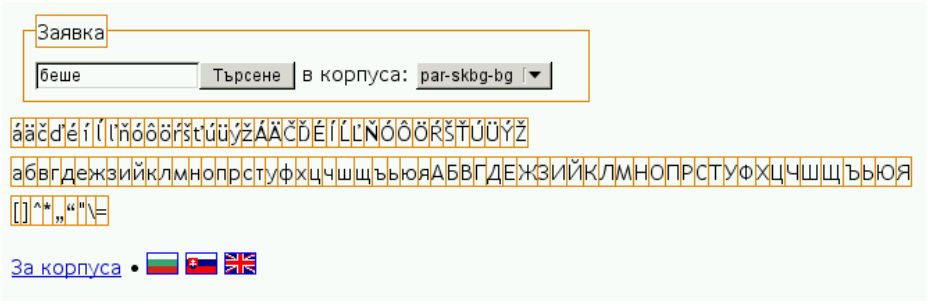

Fig. 2. Web search interface – a dialogue box in Bulgarian. A virtual Slovak/Bulgarian keyboard is visible under the query input field.

6 Applications of the bilingual corpus and future development

One of the uses of parallel corpora is the machine translation technology, either for training statistical models or translation evaluation (the current version of the corpus is well usable for the latter, but its size should be significantly increased before it will be useful for the machine translation training). The parallel and comparable corpora can be also used as a translation memory and language learning materials for training of (human) translators.

Besides, the bilingual corpora are envisaged to be used in education, in schools as well as universities for the foreign language instruction.

Fig. 3. Concordances of the Bulgarian adjectives *добър* in the corpus

In addition, the corpus is useful as a language material for bilingual lexical and terminological databases and on-line dictionaries development [DK09, ŠGD09]. The Bulgarian–Slovak corpus could be also useful to linguists-researchers for research purposes, e.g. in comparative and contrastive studies of Bulgarian and Slovak (Fig. 3, 4).

Fig. 4. Concordances of Bulgarian verb *беше* – (imperfect form of verbs of imperfective aspect in Bulgarian) of the verb *съм* (Dimitar Dimov's Doomed Souls and the Slovak translation thereof)

Further work will involve enriching the annotation of the corpora that will increase their representativeness, with regard to its composition and size. We will evaluate existing possibilities for automatic morphosyntactic annotation and lemmatisation of Bulgaria texts and for automatic syntactical annotation of both Slovak and Bulgarian texts of the corpus.

7 Conclusion

The paper describes Bulgarian-Slovak and Slovak-Bulgarian parallel and aligned corpora. The corpus is aimed as a widely-distributable dataset for language engineering, for the development of machine translation and other multilingual technologies, such as tools for the development of lexical databases and digital dictionaries, for human translation training and contrastive and terminology studies. Although intellectual property right prohibit open distribution of the corpus, we try to make it as accessible to the wide scientific community as possible, either by providing statistics or language models on request.

References

[DEI+98] L. Dimitrova, T. Erjavec, N. Ide, H. Kaalep, V. Petkevič, and D. Tufiş. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the COLING-ACL'98*, pages 315–319, Montréal, Québec, Canada, 1998.

[DK09] L. Dimitrova and V. Koseska–Toszewa. Bulgarian-Polish Corpus. *Cognitive Studies/Études Cognitives*, 9:133–141, 2009.

[DKRR10] L. Dimitrova, V. Koseska–Toszewa, R. Roszko, and D. Roszko. Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus. *Cognitive Studies/Études Cognitives*, 10:217–240, 2010.

[IBR00] N. Ide, P. Bonhomme, and L. Romary. XCES: An XMLbased Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference.*, pages 825–830, Paris, 2000. ELRA.

[R. 06] R. Garabík and В. П. Захаров. Параллельный русско-словацкий корпус. In *Труды международной конференции Корпусная лингвистика*, pages 81–87, Санкт-Петербург, 2006. Издательство С.-Петербургского университета.

[Sch97] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pages 154–164. UCL Press, London, GB, 1997.

[ŠGD09] M. Šimková, R. Garabík, and L. Dimitrova. Design of a multilingual terminology database prototype. In V. Koseska, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open Workshop*, pages 123–127, Warsaw, Poland, 2009.

[VG07] D. Vasilišinová and R. Garabík. Parallel French-Slovak Corpus. In *Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007*. Tribun, Brno, 2007.

[VNH+05] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 590–596, 2005.