COGNITIVE STUDIES | ÉTUDES COGNITIVES, 15: 327–334
Warsaw 2015

DOI: 10.11649/cs.2015.022

RADOVAN GARABÍK[1,A] & **LUDMILA DIMITROVA**[2,B]

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

EXTRACTION AND PRESENTATION OF BILINGUAL CORRESPONDENCES FROM SLOVAK-BULGARIAN PARALLEL CORPUS

Abstract

In this paper the results of the automatic extraction and presentation of bilingual correspondences from Slovak-Bulgarian Parallel corpus are described. The equivalent phrases are extracted from sentence and word level automatically aligned corpus, filtered, indexed and presented in a dictionary-like interface. The bilingual dictionary database contains 80 thousand phrase pairs consisting of approximately 350 thousand words (per each language). Counting unique word forms, the size is 31 thousand in the Slovak part of the dictionary, 26 thousand in the Bulgarian part.

Keywords: translation equivalents; GIZA++; parallel corpora; aligned text; Slovak; Bulgarian

1. Introduction

In this article the authors describe the results of an experimental study on the Slovak-Bulgarian/Bulgarian-Slovak parallel corpus, prepared under the collaborative work in the frame of the Joint research project "Electronic Corpora — Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources" between the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences and Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences. The first joint study under this project consists of the analysis of differences between the Bulgarian and Slovak languages in the MULTEXT-East morphology tagset for corpora annotation (Garabík, Majchráková, & Dimitrova, 2009). The second study is a corpus-based experiment, focusing on the analysis of automatic extraction and visualization of translation equivalents of Slovak-Bulgarian/Bulgarian-Slovak parallel texts with the ultimate goal to obtain useful

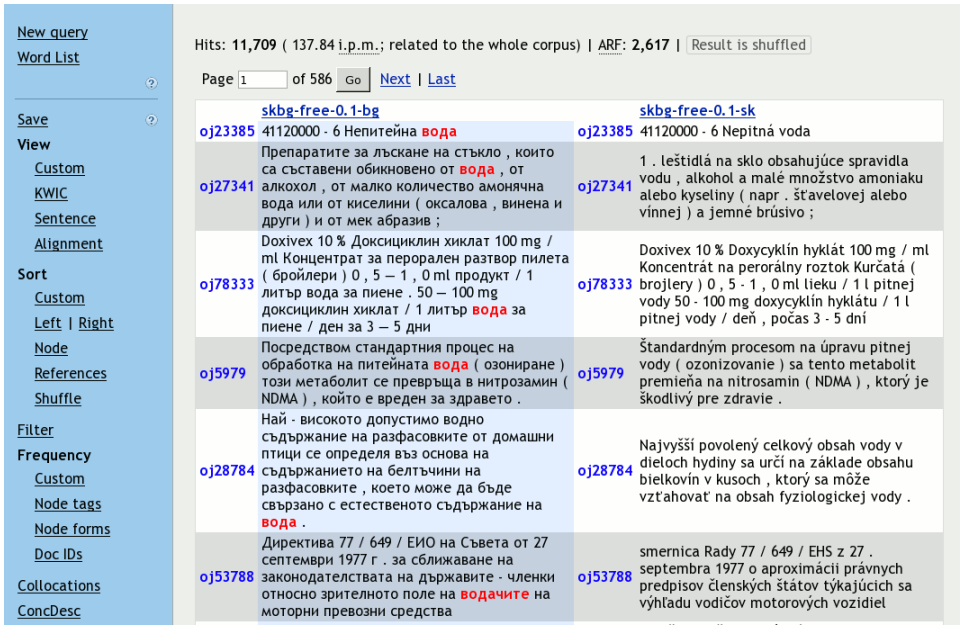

Figure 1: NoSketch interface to the parallel corpus, query word *вода*.

information about Slovak translation equivalents of (definite) articles and demonstrative pronouns in Bulgarian (Dimitrova & Garabík, 2014).

2. Slovak-Bulgarian/Bulgarian-Slovak Parallel Corpus

The parallel sentence-aligned Slovak-Bulgarian/Bulgarian-Slovak corpus is currently under development as a bilingual resource for different kind of language analysis, for research and development of machine and human translation systems, automatic term extraction, etc. A recent version of the corpus is available via a NoSketch Engine web interface at `http://korpus.sk/skbg.html`.

The corpus consists of two parts, fiction texts of about 650 thousand words and over 82 million words (in Bulgarian) and 85 million words (in Slovak) of texts of the EU & EC journals and documents (Dimitrova & Garabík, 2011, 2012).

2.1. Corpus Structure

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 sentence-aligned 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 set of aligned texts includes two Bulgarian novels: Dimitŭr Dimov's *Осъдени души* (Doomed Souls), Pavel Vezhinov's *Бариерата* (The Barrier) and their Slovak translations, the novel of

Slovak writer Klára Jarunková *Brat mlčanlivého vlka* (The silent wolf's brother) and its Bulgarian translation.

The translated Bulgarian-Slovak parallel subcorpus consists of Bulgarian and Slovak translations of works into a third language, namely the Slovak and Bulgarian translations of Jaroslav Hašek's *Osudy dobrého vojáka Švejka za světové války* (The Good Soldier Švejk) and a set of texts of the EU&EC journals and documents.

Recently, the texts of Bulgarian novel — Pavel Vezhinov's *Нощем с белите коне* (In the night riding the white horses), and Ĭordan Ĭovkov's short stories *Песента на колелетата* (The Song of Wheels), *Вечери в Антимовския хан* (Inn at Antimovo), *Ако можеха да говорят* (If they could talk), *Женско сърце* (Women heart) and their Slovak translations were also included in the direct Bulgarian-Slovak subcorpus. The volume of the literature parallel texts is about 650 thousand words per language.

2.2. Morphological Annotation

At the first step of our study we prepare morphologically annotated sentence-aligned parallel texts. The Slovak texts are morphologically annotated automatically by the tagger Morče which has been trained and tuned on tagset, developed by the Slovak National Corpus (Garabík & Šimková, 2012).

For analysis of Bulgarian morphology, we used the Bulgarian parameter file for the Tree-Tagger (Schmid, 1997), using the tagset from the BulTreeBank project (Simov, Osenova, & Slavcheva, 2004).

2.3. Alignment

The bilingual sentence-aligned corpora are valuable resources for many NLP applications: for machine translation research, for searching/extracting of language data, and can be also used as a translation database and language learning materials for training of translators — human and programming tools. The web-presented bilingual aligned corpora are available and oriented both to human and machine users. Such corpora and derived from them special type of lists, as frequency lists and concordances, are useful for language teaching. Concordances have also many applications in contrastive studies: they are used for comparison of different uses of the same word (in a different context), and to locate and analyse phrases and idioms in a given text; to find the translation of the essential elements of text, such as terms (in multilingual texts).

To align the text on the sentence level, we use the hunalign software (Varga et al., 2005). The uses a corresponding bilingual Slovak-Bulgarian dictionary to ensure a higher accuracy of the alignment; we used a small bootstrapped dictionary that has been generated automatically and then manually proofread, removing incorrect word pairs. Alignment on the word level was performed using the GIZA++ software (Och & Ney, 2000), using (for simplicity) only sentence pairs where the alignment was 1:1. Generally, word alignment is M:N (any number of Bulgarian words can map to any number of Slovak ones), although only 1:1 and realistically at most 1:2 (and 2:1) appear in our corpus texts.

3. Phrases Extraction

We use the MOSES (Koehn et al., 2007), a statistical machine translation toolkit to process the corpus. The toolkit uses GIZA++ to obtain an initial word alignment which is subsequently improved by a "grow-diag-final" method.

Throughout this article, we use the term 'phrase' following the MOSES terminology, i.e. a phrase is a short sequence of one or several words that has been selected from the text corpus (and aligned with a corresponding text chunk — phrase — from the second language part of the corpus), and has no connection with a 'phrase' as a term in general linguistics. Although MOSES could be used to build a machine translation system based on our corpus, this was not our goal and we used only the training process which produces aligned and scored bilingual phrase tables.

MOSES training produces four different phrase translation scores:

- inverse phrase translation probability $\varphi(f|e)$
- inverse lexical weighting $lex(f|e)$
- direct phrase translation probability $\varphi(e|f)$
- direct lexical weighting $lex(e|f)$

Ideally, we would like to compute a single score out of these four numbers, reflecting the level of "suitability" of the phrase pair. Since we designed our interface to be language-direction agnostic (i.e. conceptually neither Bulgarian → Slovak nor Slovak → Bulgarian correspondence should be favoured), and since we want to take into account not just phrase correspondence, but also correspondence of individual words, our score must be symmetrical with regard to $\varphi(f|e)$ and $\varphi(e|f)$, as well as to $lex(f|e)$ and $lex(e|f)$, and should reflect the likelihood-like nature of these scores. The simplest function that fulfils these criteria is a simple product, $g = \varphi(f|e) \cdot \varphi(e|f) \cdot lex(f|e) \cdot lex(e|f)$.

In order to be able to quantify the correctness of extracted phrases, we split the phrases into sets according to logarithm of the score g, in intervals two orders of magnitude wide, i.e.

$$g \in (10^{-30}, 10^{-28}] \cup \ldots \cup (10^{-6}, 10^{-4}] \cup (10^{-4}, 0.01] \cup (0.01, 1]$$

In each interval, we randomly selected 10 sentences (population sample) and manually annotated their correspondence, selecting between three options: good, bad and not sure. In the interval $[10^{-14}, 10^{-6}]$ we increased the number of sentences to 30, to get better estimates. Since the sampling of sentences from each interval is without replacement, the probability distribution is hypergeometric; however the number of sentences in each interval (population) is on the order of millions and therefore we can approximate the distribution by a binomial one (this is relevant for confidence interval estimation). For each of the intervals, we calculate the ratio:

$$r = \frac{C(good)}{C(good) + C(bad) - C(notsure)}$$

i.e. we remove the "not sure" sentence pairs from the sample and calculate the ratio of good ones. In order to get a function describing relation of the parameter r to the

score g, we start with several basic assumption. First, phrases with $g = 1$ should be perfect equivalents, $r(g = 1) = 1$. Phrase pairs with very low score should be completely bogus:

$$\lim_{g \to 0^+} r(g) = 0$$

We are therefore looking for a sigmoid function whose values start at zero at zero and saturates when approaches 1. Since we are operating on intervals defined by orders of magnitude, we use the sigmoid function on domain $\ln(x)$. Generic logistic function is defined by:

$$f(x) = \frac{1}{1 + e^{-a \cdot (x - x_0)}} \tag{1}$$

where x_0 is the centre of the function (horizontal shift) and a reflects the slope ('steepness'). Using the logistic function of a variable $x' = \ln(x)$ and simplifying we get

$$f(x) = \frac{1}{1 + \frac{e^{a x'_0}}{x^a}} \tag{2}$$

We then fit our data points with the function (2) to obtain the parameters a and x'_0, which gives us:

$$a = 0.142 \pm 0.021 \tag{3}$$

$$x'_0 = -35.0 \pm 1.2 \tag{4}$$

and subsequently $x_0 = e^{x'_0} = 6.3 \cdot 10^{-16}$, i.e. the ratio of incorrectly aligned phrase pairs will reach 0.5 around $g = 6.3 \cdot 10^{-16}$. The relation between the score g and the ratio r of our population samples is depicted on Fig. 2, together with the function (2).

We can use the function (2) to obtain the value of parameter g where the ratio of correctly aligned phrases drops below certain value — we decided to keep 95% accuracy, so solving the equation

$$r_0 = 0.95 = \frac{1}{1 + \frac{e^{a x'_0}}{x^a}}$$

for x gives us a threshold $g_0 = 6.4 \cdot 10^{-7}$ for the desired 95% accuracy.[1]

After applying the above mentioned threshold, we examined the phrases we obtained, sorted by the score g. At the beginning we have phrase pairs with $g = 1$. This implies that all the factors have to equal one: $\varphi(f|e) = \varphi(e|f) = lex(f|e) = lex(e|f) = 1$. In essence, this means that both Bulgarian and Slovak phrase match

[1] Note that although the uncertainties of a and x'_0 in equation (4) are small in relative values, the parameters themselves are in the exponent and the uncertainty has an enormous effect on absolute value of the function. This analysis should therefore be considered just a very rough estimation.

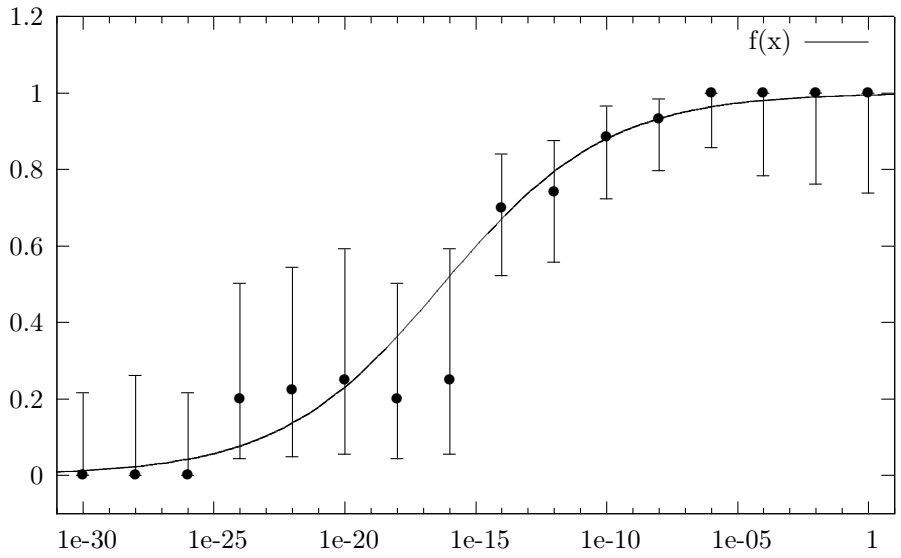

Figure 2: Relation between phrase alignment score g (horizontal) and the ratio of good pairs r (vertical). Vertical error bars display Jeffreys intervals at 95% confidence level (Brown, Cai and DasGupta, 2001).

perfectly — they occur always in the same form, and all the words of the phrases are always translated in the same way. This happens most of the time if there is a (often unique) foreign language (i.e. neither Bulgarian nor Slovak) citation, such as a name of a company or product (most striking are those in a foreign script, e.g. Greek occurs relatively often). Since "normal" sentences do not appear here at all, we included a condition $g \lneqq 1$ in our filter (such "foreign script" phrases would be excluded by following filters anyway).

Additional heuristic filtering consists of excluding phrases that:

- do not start with a letter
- contain punctuation (apart from a comma)
- are not exactly 4 words long
- contain more than 3 words starting with a capital letter
- end with a preposition
- contain characters out of the appropriate alphabet (Slovak or Bulgarian)

4. Search Interface

After filtering, we got 80 thousand phrase pairs, which we indexed by words and lemmas for our dictionary query system — keys (headwords) for each phrase consist of a union of lemmas and word forms from both Bulgarian and Slovak phrase, and their equivalents without diacritics (for Slovak) and transliteration into Latin script (for Bulgarian), to facilitate queries for users without ways of entering Cyrillic or

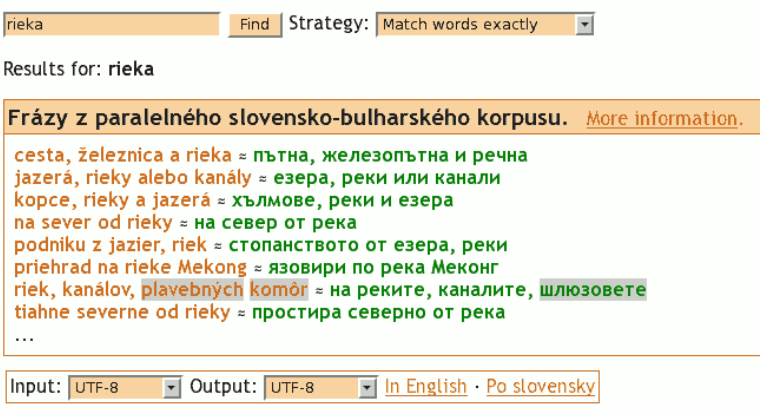

Figure 3: Fragment of the search interface, queried word rieka. Note the highlighted example of word alignment, *plavebných komôr* and *шлюзовете*, featuring 2:1 correspondence.

Slovak diacritics. The amount of words in the database is 350 thousand (per each language). Counting unique word forms, the size is 31 thousand in the Slovak part of the dictionary, 26 thousand in the Bulgarian part.

For the dictionary access, we are using the dict (RFC 2229) server[2] as a backend, with a CGI frontend formatting the results in an intuitive and graphically representative way (see Fig. 3). The interface is accessible at `http://slovniky.korpus.sk/?d=pskbg`.

5. Conclusion

This paper presents results of an experimental study, namely automatic extraction and presentation of bilingual correspondences from Slovak-Bulgarian/Bulgarian-Slovak parallel and aligned corpus. The parallel Slovak-Bulgarian corpus, currently under development, is a valuable bilingual resource for language analysis, automatic term extraction, the research and development of machine and human translation systems, supervised and unsupervised NLP tools training, and machine translation.

References

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 101–117.

Dimitrova, L., & Garabík, R. (2011). Bulgarian-Slovak Parallel Corpus. In *Natural Language Processing, Multilinguality: Proceedings of the 6th International Conference SLOVKO 2011, Modra, Slovakia* (pp. 44–50).

Dimitrova, L., & Garabík, R. (2012). Bilingual corpus — digital repository for preservation of language heritage. In *Proceedings of the International Conference Digital Presentation and Preservation of Cultural and Scientific Heritage DiPP 2012, Veliko Tŭrnovo, Bulgaria* (pp. 132–141).

[2] `http://dict.org`

Dimitrova, L., & Garabík, R. (2014). Translation equivalence of demonstrative pronouns in Bulgarian-Slovak parallel texts. *Cognitive Studies | Études cognitives*, *14*, 65–74. http://doi.org/10.11649/cs.2014.007

Garabík, R. & Šimková, M. (2012). Slovak morphosyntactic tagset. *Journal of Language Modelling*, *0*(1), 41–63. http://doi.org/10.15398/jlm.v0i1.35

Garabík, R., Dimitrova, L., & Koseska-Toszewa, V. (2011). Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies | Études cognitives*, *11*, 227–239.

Garabík, R., Majchráková, D., & Dimitrova, L. (2009). Comparing Bulgarian and Slovak Multext-East morphology tagset. In *Organization and development of digital lexical resources* (pp. 38–46). Kyiv. Dovira Publishing House.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.*

Och, F. J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 440–447). Stroudsburg, PA: Association for Computational Linguistics. http://doi.org/10.3115/1075218.1075274

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. Jones & H. Somers (Eds.), *New methods in language processing* (pp. 154–164). London: UCL Press. (*Studies in Computational Linguistics*).

Simov, K., Osenova, P., & Slavcheva, M. (2004). BTB-TR03: BulTreeBank Morphosyntactic Tagset. BulTreeBank Project Technical Report № 03. Technical report, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005* (pp. 590–596).

Acknowledgment

This work was supported by a bilateral project "Electronic Corpora — Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources (continuation of the project)" as part of the collaboration between Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences and Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences.

The authors declare that they have no competing interests.

The authors' contribution was as follows: Both authors participated in the manuscript; RG focused on technical work and mathematical representation of the ideas, LD on text alignment and accuracy evaluation.

This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 PL License (http://creativecommons.org/licenses/by/3.0/pl/), which permits redistribution, commercial and non-commercial, provided that the article is properly cited.

© The Authors 2015

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation