COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 65–74
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.007

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

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

TRANSLATION EQUIVALENCE OF DEMONSTRATIVE PRONOUNS IN BULGARIAN-SLOVAK PARALLEL TEXTS

Abstract

In this paper we describe our automatic analysis of several parallel Bulgarian-Slovak texts with the goal to obtain useful information about Slovak translation equivalents of (definite) articles and demonstrative pronouns in Bulgarian. Rather than focusing on individual translation equivalents, we present a method for automatic extraction and visualization of the translations. This can serve as a guide for pinpointing interesting features in specific translated documents and could be extended for other parts of speech or otherwise identifiable textual units.
Keywords: translation equivalents, demonstrative pronouns, parallel corpora, aligned text, Slovak, Bulgarian.

1 Introduction

In this article we briefly describe results of the second experimental study on the Bulgarian and Slovak digital resources. The first one consists of the analysis of differences between the Bulgarian and Slovak languages MULTEXT-East morphology tagset for the corpora annotation (Dimitrova, Garabík & Majchráková, 2009). The second experimental corpus-based study is prepared under the collaborative work in the frame of Joint research project "Electronic Corpora — Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources" between Institute of Mathematics and Informatics of BAS and Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences.

The parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus[1] contains 376 200 words of fiction 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). The set of aligned literature texts we used in our article includes two Bulgarian novels: Dimitar Dimov's *Осъдени души* (Doomed

[1]The recent version of the corpus is available via a simple web interface at http://korpus.sk/skbg.html.

Souls) and 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 Slovak and Bulgarian translations of Jaroslav Hašek's *Osudy dobrého vojáka Švejka za světové války* (The Good Soldier Švejk). The contrastive study we describe here is a corpus-based attempt to compare the translation correspondences of the articles and demonstrative pronouns in Bulgarian and Slovak. Since a difference between the classifications of the demonstrative pronouns exists, we need to harmonize in some way the specifications that describe the Bulgarian and Slovak demonstrative pronouns.

2 Bulgarian Demonstrative Pronouns

Pronouns are words that under certain conditions and circumstances can replace nouns (respectively collocations with these nouns as a main word) adjectives or numerals. Some types of pronouns used in the speech to indicate the different objects and their signs or to ask a question (query) for them. Lexical meaning of pronouns is very general and undefined. These pronouns, that replace nouns, adjectives and numerals, express lexical meaning of the word instead which they are used. Therefore, they have not their own lexical meaning in the strict sense, and realize an anaphoric service in the speech, i.e. to direct the mind of the message's perceiver to the names already mentioned or to the known objects. The pronouns are words that belong to various grammatical categories. Some types of pronouns have retained case forms that are used in the modern Bulgarian language, while the nouns have lost their flexion. Some pronouns possess also the grammatical categories of gender, number and person. Bulgarian pronouns are classified as: personal pronouns, possessive pronouns, reflexive pronouns, demonstrative pronouns, interrogative pronouns, relative pronouns, indefinite pronouns, negative pronouns, and summative pronouns — 9 types in total.

The demonstrative pronouns indicate different objects (person or objects) and their qualitative and quantitative characters, for example:

— този човек /this man/; тази река /this river/; това цвете /this flower/; тая вечер /tonight/; тези планини /these mountains/; тия камъни /these stones/;

— онова дърво /that tree/; онези дървета /those trees/;

— полезни храни /useful foods/ ↔ тези храни /these foods/,

— добра жена /good woman/ ↔ такава жена /a woman/,

— 10 тетрадки /10 notebooks/ ↔ толкова тетрадки /so notebooks /,

— хубав град /nice city/ ↔ такъв град /this town/.

Many authors have published research on the origin and development of demonstrative pronouns in Slavic languages, among them Lane (1961), Kortlandt (1983), Catasso (2011). The demonstrative pronouns in the modern Bulgarian language are inherited from the Old Bulgarian language (Stoianov, 1993), having undergone some changes mainly by adding intensifying modal particles to the original forms. Such particles are: *-в*, *-ва*, *-зи*, *-я*.

Table 1 Development of personal pronouns from Old to Modern Bulgarian

Old Bulgarian	Modern Bulgarian
тъ, та, то	*то-зи*, *то-я*, *та-зи*, *та-я*, *то-ва*, *те-зи*
онъ, она, оно	*он-зи*, *он-я*, *она-зи*, *она-я*, *оне-зи*
такъ, така, тако	*такъ-в*, *така-ва*, *тако-ва*, *таки-ва*

There are three types of demonstrative pronouns: for persons and objects, for quality and for quantity. The demonstrative pronouns agree in number and gender with the noun they refer to (except for this for quantity).

Each demonstrative can not only modify a noun, but also be used on its own.

Personal demonstrative pronouns have two forms (for persons and objects): for nouns that are close to the speaker or writer and for far nouns.

Table 2 Classification of Bulgarian personal pronouns according to gender and proximity

	Sg. Masc.	Sg. Fem.	Sg. Neuter	Pl. (all genders)
For nouns close to the speaker or writer (this)	този/тоз тоя	тази/таз тая	това туй	тези/тез тия
For nouns far from the speaker or writer (that)	онзи/оня оная	онази/оназ онуй	онова ония	онези/онез

Quality demonstrative pronouns have also two forms:

1) Positive, that specifies that the noun has a particular quality (such; this kind of/this sort of; that sort of/of that type; (такова /so/))

Sing. Masc.	Sing. Fem.	Sing. Neuter	Plural (all gender)
такъв	такава	такова	такива

2) Negative, that specifies that the noun doesn't have a particular quality or has a different one (not this kind of/not this sort of; not of that type).

Sing. Masc.	Sing. Fem.	Sing. Neuter	Plural (all gender)
онакъв	онакава	онаково	онакива[a]
инакъв	инаква	инакво	инакви
толкав	толкава	толкаво	толкави[b]
толчав	толчава	толчаво	толчави[c]

[a] for persons and objects

[b] a conversational tone styling / with nuance of a colloquialism /

[c] "толчав" and its word-forms hardly used in the literary language of the XXI century

Demonstrative pronouns for quantity: толкова, толкоз

The demonstrative pronoun for quantity "толкова" is used with nouns and adjectives. It both specifies the exact quantity of something — this many/this much, and indicates the large extent or degree of something — so (many/much). Both *толкова* and *толкоз* translate in English as: so; this, that; that much/many; so much/many; this/that much.

Some authors add the pronominal adverbs to different groups of pronouns, including the demonstrative pronouns (Rusinov, 1993). Bulgarian pronominal adverbs form a group of adverbs, usually formed by means of some pronominal roots, or by merging of two prepositions, of a preposition and a pronoun, an adverb, or a noun.

The demonstrative pronominal adverbs are classified as follows: for time: close (сега) and far (тогава); for manner: close (така), far (иначе); for place: close (тук(а)), far (там); for direction: close (насам), far (натам); for reason (затова), and for quantity (толкова) — the same word as demonstrative pronoun of quantity. This list includes the main representatives of Bulgarian demonstrative pronominal adverbs, without listing all formal and archaic ones.

This is the reason we decided to include them in the analysis, as well as the fact that they are unambiguously classified as pronouns in Slovak grammars.

3 Slovak Demonstrative Pronouns

Table 3 Classification of Slovak personal pronouns according to their proximity

For nouns close to the speaker or writer (this)	tento	takýto	toľkýto	toľkoto	takto	tadeto tadiaľto	potiaľto
For nouns far from the speaker or writer (that)	ten	taký	toľký	toľko	tak	tade tadiaľ	potiaľ

For nouns close to the speaker or writer (this)	odtiaľto stadiaľto	pre toto	za toto	na toto	tu	sem
For nouns far from the speaker or writer (that)	odtiaľ stadiaľ	preto	zato	nato	tam	ta

The canonical classification of Slovak pronouns is described in (Ružička, 1966); the demonstrative pronouns indicate specific object, person, or quality. As opposed to traditional Bulgarian classification, there is a class of demonstrative pronouns with adverbial behaviour — while in Bulgarian these would be considered (demonstrative) adverbs, they are classified as pronouns in Slovak grammars.

The distinction between proximal and distal demonstratives is present in Slovak too, however the description in (Ružička, 1966), as pointed out in an analysis

by Rybák (1969) is neither complete nor consistent. We reproduce here the table describing proximity classification from the analysis mentioned (Table 3); as the author himself notes, the table should not be considered definitive since the proximal/distal dichotomy is broken for some pronoun pairs and the research was just preliminary. Also note that the author introduced demonstrative pronouns with a preposition as a (potential) proximal counterpart — a controversial position in our opinion, since these combinations are not really perceived as a single semantic unit and are used rather rarely in Slovak. The author also indiscriminately mixes demonstrative pronouns and demonstrative adverbs. In the table we refrained from displaying all the three genders — the table would get too unwieldy since apart from the expected masculine, feminine and neuter forms in the singular there are also separate forms for masculine and non-masculine in the plural; we display only masculine singular.

Unlike Bulgarian, the distal demonstrative pronouns are the 'neutral' ones — as noted in (Ružička, 1966) and elaborated in (Rybák, 1969), there is a semi-regular process deriving proximal pronouns (including adverbial ones) from the distant ones.

For completeness, we mention another class of demonstrative pronouns — the ones derived from the root *on-*, e.g. *onen. oná, onam, onaký.* These are the true distal demonstratives, indicating persons, objects or qualities that are either temporally or spatially removed from the speaker (and listener). These pronouns are however used rather infrequently and their usage is rather stylized.

4 Morphology Tagging

In the following text, we use short identification strings for different analysed texts — 'barrier' for Vezhinov's *The Barrier*, 'brother' for Jarunková's *The Silent Wolf's Brother*, 'souls' for Dimov's *Doomed Souls*, 'švejk' for Hašek's *The Good Soldier Švejk* and 'eu' for the collection of European Union texts.

It should be noted that only *barrier* and *souls* were translated from Bulgarian to Slovak; *švejk* is a translation from a third language (Czech) while *eu* contains texts that were most probably translated from English or French. We include these for completeness, the analysis was oriented predominantly to Bulgarian→Slovak translations.

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

We tried to estimate the precision and recall of correct tagging of Bulgarian demonstratives. The recall (of demonstrative pronouns) has been calculated in a straightforward manner by comparing the annotation with a fixed list of Bulgarian pronouns — the recall for various texts in Table 4, the precision was virtually 100 % (the only mistakes were 8 words from the *eu* texts where some Cyrillic letters were replaced by identically looking Latin ones).

Table 4 Size of input texts and recall of TreeTagger for demonstrative pronouns

source	size [tokens]	recall
barrier	36740	0.9940
brother	72981	0.9880
souls	106389	0.9904
švejk	76686	0.9932
eu	87974907	0.9987
all	88267703	0.9986

To estimate the precision of article taging, we took a sample of random 1000 tokens (i.e. nouns, numerals, adjectives, pronouns and participles) tagged with article. Out of these 1000 tokens, there were 35 false positives (out of these 8 proper nouns ending with suffixes similar to Bulgarian articles). This gives the precision of 96.5 % (with an error bound on the estimate being 3 %), and the results concerning articles should be taken with this accuracy in mind. The most frequent mistake was the tagging of тези as `Ncfpi` (that is, a substantive).

Table 5 Percentage of Bulgarian determiners and various types of pronouns for separate input texts.

source	articles [% of tokens]	pronouns [% of tokens]	close [%]	far [%]	other [%]
barrier	6.04	2.25	53.3	1.33	45.4
brother	7.73	1.49	44.4	3.04	52.3
souls	9.63	1.88	65.8	2.55	31.6
švejk	8.48	1.92	56.3	3.06	40.7
eu	7.74	0.48	82.8	0.63	16.5

For our analysis, we separated the Bulgarian articles and demonstrative pronouns[2] into four distinct classes (according to traditional semantic division): articles as a separate class, close demonstrative pronouns, far demonstrative pronouns, and "other" demonstrative pronouns (for quality and quantity) and also parts of speech that are usually classified as pronominal adverbs (see Table 5).

For each of the Bulgarian sentence we calculate the number of determiners in each of the four classes. We exclude sentences where there are determiners of two or more classes present in one sentence (this means we keep the sentences with zero determiners), to get a 'pure' class, and then look for Slovak demonstrative pronouns in the translated sentence.

We should note (and expect) that Slovak demonstrative pronouns do not correspond directly to Bulgarian determiners — for example the most frequent demonstrative pronoun *to* is used predominantly as an impersonal 3rd person pronoun.

[2]We use the term 'determiners' for both articles and demonstrative pronouns.

To estimate the influence of unrelated Slovak demonstrative pronouns in the translation, we define the parameter $p_0(s) = d_{sk}/|s|$, where d_{sk} is the number of demonstrative pronouns in the Slovak sentence and $|s|$ is the length of the sentence. We then calculate the mean value:

$$\overline{p_0} = \frac{1}{|\mathfrak{S}|} \sum_{s \in \mathfrak{S}} p_0(s)$$

where $\mathfrak{S}$ is the set of all considered sentences and $|\mathfrak{S}|$ its size (i.e. number of sentences). To get an appropriate value of this 'background noise', we calculate this parameter for each Slovak demonstrative pronoun, taking into account sentences where the original Bulgarian text did not contain any determiner (the 'pure' class). This parameter therefore shows the frequency of given pronoun scaled by the importance of the pronoun in the sentence (as opposed to a more straightforward notion of frequency as the ratio of pronouns to the total number of words in the whole text).

Table 6 Most frequent Slovak pronouns according to the parameter $p \cdot 1000$ for Bulgarian sentences lacking determiners

barrier		brother		souls		švejk	
to	12.85	to	12.13	to	9.53	to	16.45
tam	0.77	tak	2.00	taký	1.05	ten	1.97
taký	0.35	ten	1.18	tam	0.59	tak	1.50
vtedy	0.28	tá	1.17	sem	0.57	tam	1.37
tak	0.16	taký	1.08	tak	0.44	taký	0.92

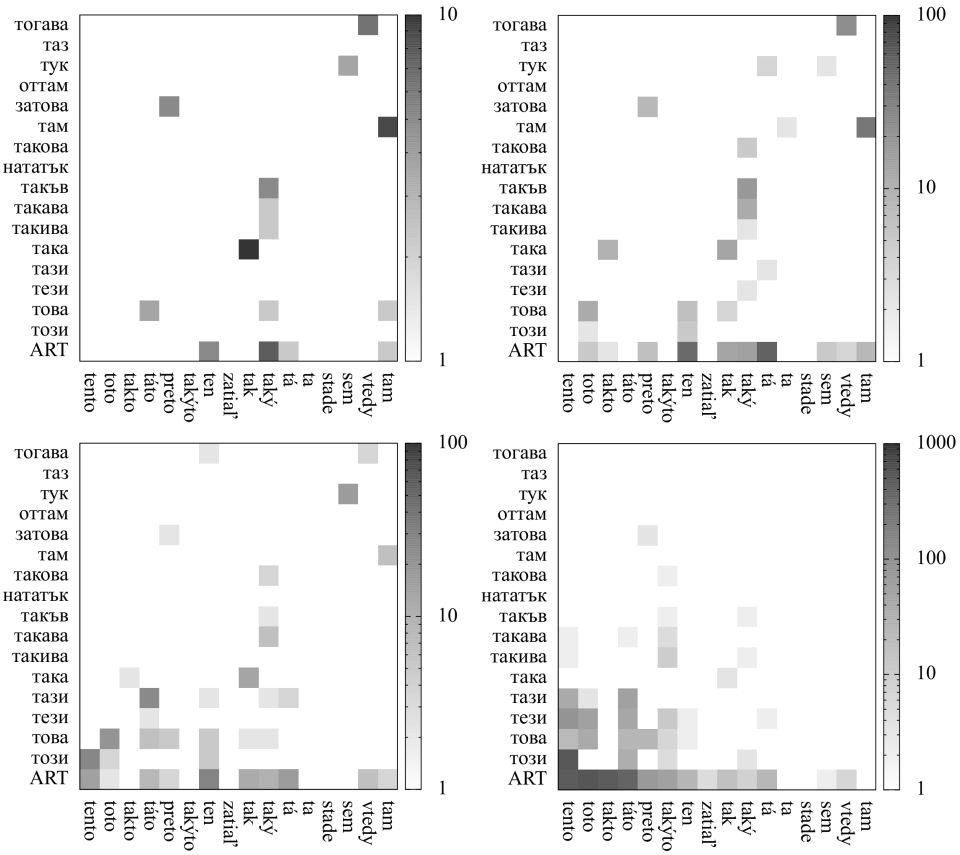
As we see from the Table 6, the Slovak pronoun *to* is indeed used very frequently (compared to other demonstrative pronouns) even in sentences where the original Bulgarian text did not contain any determiners; therefore we do not analyze occurrences of this pronoun further, because we would not be able to maintain reasonable confidence about the accuracy of the results.

5 Visualization and Analysis

To easily see the translational equivalent between Bulgarian and Slovak pronouns, we organize them in a matrix where each row m corresponds to a specific Bulgarian determiner, each column n corresponds to a specific Slovak pronoun and the value at the (m, n) cell shows the number of translations — we take into account only those sentences where there is at most one Bulgarian determiner and at most one Slovak demonstrative pronoun. The matrices are then displayed using a 'heatmap', where the intensity of each cell corresponds to the number of translations (Figure 1). For space reasons, we depict only four texts, with only the most frequent Bulgarian and Slovak determiners visible. To facilitate comparisons, we kept the determiners in the same order for each of the texts; to improve visibility and contrast, the images are modified — the values are mapped to intensity levels on a logarithmic scale and the matrices are 'cleaned up' — the cells with the number equal to one are

removed. We also omitted the cells corresponding to 'empty' translation (if either the original or the translation lacks determiners), this information is important for the analysis, but drowns the rest of the picture with its rather high values.

Figure 1 Heatmap of translation equivalents in several texts. From left to right, top to bottom: *barrier, brother, souls, eu.* `ART` stands for (any) article



From the visualisation we can immediately derive some conclusions — surprisingly, *barrier* does not use the words *toto* and *tento* in translations. This is in constrast with *souls*, where *tento* is a preferred translation of този, and is also used to translate some articles; *toto* is the preferred translation of това and is also used to translate този. Texts in *brat* are translated from Slovak to Bulgarian — we see that (as expected) both *ten* and *tá* were translated mostly by deploying articles. The translation of тогава as *vtedy* is straightforward in all the texts with the exception of *eu*, where тогава does occur very rarely; this can be explained by different nature of the texts. Similarly missing word from the *eu* is там/*tam*, on the other hand, *toto* and *tamto* correspond to Bulgarian articles — this is missing from fiction texts, but this could be explained by translation from e.g. English (and

the predominance of English demonstrative pronouns and articles, and a different custom in translating non-fiction texts).

6 Concluding Remarks

In this paper we presented the results of the corpus-based experiment — to compare the translation correspondences of the demonstrative pronouns in Bulgarian and Slovak.

Maybe here we must mention the main role of human translators in a selection of Slovak words, translating Bulgarian demonstrative pronouns. According to the statistics, the preferable translated equivalences of some Bulgarian demonstrative pronouns and demonstrative pronominal adverbs are listed in Table 7.

Table 7 Preference of translation equivalents for some Bulgarian demonstrative pronouns (DP) and demonstrative pronominal adverbs (DPA)

Ratio	Bulgarian DP	Slovak translation
1:3	толкова	tak / taký / to
	такъв	tá / taký / takýto
	такова	tá / taký / takýto
1:4	такава	tá / taký / takýto / táto
	Bulgarian DPA	Slovak translation
1:1	тогава	vtedy
	тъй	taký
1:2	така	tak / to
	там	tam / tamto
	тук	tu / sem

The analysis we presented does not try to be exhaustive; rather it just presents some possibilities that are offered by sentence-aligned, linguistically (morphologically) annotated texts. We extracted not only information about translation equivalents of Bulgarian articles and demonstrative pronouns, but we also presented in a visually informative way the differences between translations of several Bulgarian texts. Similar techniques could be used for further comparisons of various parameters of translated texts.

References

Catasso, N.(2011). The Grammaticalization of Demonstratives: A Comparative Analysis. *Journal of Universal Language*, *12*(1), 7–46.

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

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* (pp. 132–141), Veliko Tarnovo, Bulgaria.

Dimitrova, L., Garabík, R., & Majchráková, D. (2009). Comparing Bulgarian and Slovak Multext-East morphology tagset. In *Development of Digital Lexical Resources. Proceedings of the 2 nd MONDILEX Second Open Workshop conference* (pp. 38–46), Kiev.

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.

Kortlandt, F. (1983). Demonstrative Pronouns in Balto-Slavic, Armenian, and Tocharian. *Studies in Slavic and General Linguistics*, *3*, 311–322.

Lane, G. S. (1961). On the Formation of the Indo-European Demonstrative. *Language*, *37*(4), 469–475. doi: 10.2307/411348.

Rusinov, R. (1993). Adverb. In S. Stoianov (Ed.), *Grammar of Modern Bulgarian Literary Language* (Vol. 2: Morphology, pp. 387–408). Sofia: Publishing House of BAS.

Ružička, J .(1966). Morfológia slovenského jazyka (1st Ed.). Bratislava: Vydavateľstvo SAV.

Rybák, J. (1969). K systému ukazovacích zámen v slovenčine. *Slovenská reč*, *34*(6), 358–362.

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

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.

Stoianov, S. (1993). Demonstrative Pronouns. In S. Stoianov (Ed.), *Grammar of Modern Bulgarian Literary Language* (Vol. 2: Morphology, pp. 198–200). Sofia: Publishing House of BAS.

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