Query Expansion for Slovak to Bulgarian Language Machine Translation Using Parallel Search

VELISLAVA STOYKOVA[1] DANIELA MAJCHRAKOVA[2]
[1]Institute for Bulgarian Language, BAS
52, Shipchensky prohod blvd., bl. 17, 1113 Sofia, BULGARIA
[2]Ludovit Stur Institute of Linguistics, SAS, Panská 26,
811 01 Bratislava, SLOVAKIA

Abstract: - The paper presents results of the application of a statistical approach for Slovak to Bulgarian language machine translation. It uses Information Retrieval inspired search techniques and employs several algorithmic steps of parallel statistical search with query expansion in Slovak-Bulgarian EUROPARL 7 Corpus using the Sketch Engine software and its scoring. The search includes the generation of concordances, collocations, word sketch differences, word sketches, and thesauri of the studied keyword (query) by using a statistical scoring, which is regarded as intermediate (inter-lingual) semantic standard presentation by means of which the studied keyword (from the source language) is mapped together with its possible translation equivalents (onto the target language. The results present the study of adjectival collocabillity in both Slovak and Bulgarian language from the corpus of political speech texts outlining the standard semantic relations based on the evaluation of statistical scoring. Finally, the advantages and shortcomings of the approach are discussed.

Key-Words: - Artificial Intelligence, Information Retrieval, Semantic Search, Machine Translation, Collocations

Received: December 18, 2020. Revised: May 12, 2021. Accepted: June 10, 2021. Published: June 27, 2021.

1 Introduction

Recent theories in the area of Machine Translation (MT) use the specific algorithm sets to support a more semantically relevant and logically unambiguous translation of related words contexts from a source to a target language. In its early beginning, the MT approaches use a knowledge-based semantic model [10] and employ linguistic rules (formal grammars) and a controlled vocabulary, so to encode and parse the text from the source language by creating a formal intermediate presentation which later decodes the result into the target language. Additionally, the large lexical database for both source and target languages is employed as well which requires time and efforts. As the main area of application is the translation of technical documentation, the approach interprets the task as a domain-specific text translation, and a development of the domain-specific lexical and terminological electronic resources like controlled vocabularies (as for legal texts, weather forecasts, etc.) is of great importance.

Similarly, the statistical approaches to MT use statistically-based models which take into account statistical parameters of bilingual text corpora using word-based, phrase-based, syntax-based, etc. statistical techniques. Recently, some statistical approaches from the area of Information Retrieval (IR) based on the estimation of association, distribution, and combinatorial properties of words shown that the words statistical similarity is connected to their semantic similarity [2, 3] and some other semantic relations (hierarchy) employing the vector-space scoring techniques as a measurement of words semantic properties. Moreover, those techniques appear to give promising results for the translation of both related and non-related languages. In addition to that, the Neural Networks based MT techniques are used [16]. The research connected to studying a semantic clustering and languages inter-relatedness are given in [9].

Further, we are going to present the results of the application of IR approach based on query expansion of parallel search and retrieval of statistically similar words for Slovak to Bulgarian language translation. We use the statistical scoring of the Sketch Engine software as the standard semantic intermediate (inter-lingual) presentation and the texts of the Slovak-Bulgarian parallel electronic corpora which are part of the EUROPARL 7 Corpus [6], and we perform several types of a word search (algorithmic steps) showing that the statistical words search from the source corpus (in the Slovak language) give more meaningful and elaborate translation equivalents in

the target corpus (in the Bulgarian language) when the query is retrieved by expanding the scope of the search.

2 Europarl Parallel Corpus

Our research is indebted and based on some preliminary results related to parallel corpora keywords search in the Slovak-Bulgarian parallel EUROPARL 7 Corpus [13]. Thus, the main Europarl parallel corpus collects the proceedings of the European Parliament sessions. It contains texts from 21 European languages including language families of related and non-related languages. From its early beginning, the corpus was created to process pairs of sentence-aligned bilingual sub-corpora for statistical machine translation [6].

The corpus preprocessing includes the identification of sentence boundaries and the alignment of bilingual pairs of related parallel corpora. In its last version, the Europarl 7 corpus uses a small number of mark-up annotations like <CHAPTER id>, <SPEAKER id name and language> and paragraph <p> without annotating any formal grammars. The existing parallel corpora relate most European languages to the English language.

2.1 Slovak-Bulgarian parallel EUROPARL 7 Corpus

The Slovak-Bulgarian parallel EUROPARL 7 Corpus consists of texts from both Slovak and Bulgarian parts of Europarl 7 corpus incorporated into and using the Sketch Engine statistical software improved with options for processing parallel corpora. The corpus is not annotated for part-of-speech, not lemmatized, and does not use any formal grammars. A similar approach to language data is presented in [15] and the Big Data approach is used for corpus processing. The Europarl 7 corpus includes about 13 000 000 words of Slovak language and 9 591 100 words of Bulgarian language allowing parallel statistical and CQL regular expressions search.

3 Sketch Engine

The Sketch Engine (SE) software [5] allows the use of various approaches to extract lexical-semantic properties of words and most of them are with multilingual application [4]. Extracting *keywords* is the most common and widely used technique to define the basic terms of a particular domain. The SE's software standard options for keyword extraction are based on the use of word frequency lists. However, semantic relations can be extracted by the generation of related word contexts through word *concordances*. Concordances define the context in quantitative terms and further work is needed to be done to define *semantic relations* by searching for co-occurrences and *collocations* of a related keyword.

Co-occurrences and collocations are words that are most probably to be found with a related keyword. They assign the semantic relations between the keyword and its collocated word which might be of similarity or a distance.

The statistical approaches used by the SE to search co-occurrence and collocated words are based on defining the probability of their co-occurrences and collocations. We have used techniques of *T−score*, *MI−score,* and MI^3 *− score* incorporated in SE for corpus processing and searching. Basically for all, the following terms are used: N - corpus size, f_A - number of occurrences of the keyword in the whole corpus (the size of the concordance), f_B - number of occurrences of the collocated keyword in the whole corpus, f_{AB} - number of occurrences of the collocate in the concordance (number of co-occurrences). The related formulas for defining *T−score*, *MI−score,* and MI^3−score are as follows:

$$\text{MI-Score } \log_2 \frac{f_{AB} N}{f_A f_B}$$

$$\text{T-Score } \frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

$$\text{MI}^3\text{-Score } \log_2 \frac{f_{AB}^3 N}{f_A f_B}$$

The *T−score*, *MI−score,* and MI^3 – score is applicable for processing parallel corpora as well.

4 Search Results from Slovak-Bulgarian parallel EUROPARL 7 Corpus

Some previous research done with the Slovak-Bulgarian EUROPARL 7 Corpus presents the study of time expressions translations equivalency [12] by employing the parallel search which uses statistical scoring (*T−score*, *MI−score,* and MI^3−score). Further, we shall study the translation of adjectives relating to time (and time expressions) by using the same scoring and expanding the query of the search

at every step using the generation of concordances, collocations, word sketches, and thesauri.

First, we search the Slovak part of the EUROPARL 7 Corpus, so to detect the semantic relations of the adjective *dnešný* (EN – *today's*) studying its combinatorial properties by generating the keyword's concordance (Fig. 1) which present all its occurrences in the corpus (3 484) with their related semantic contexts.

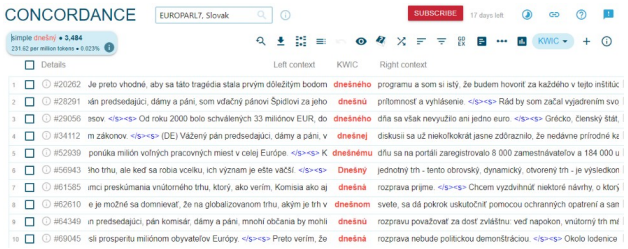

Fig. 1. The concordance of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus.

However, the result does not present structured information about the semantic relations of keyword *dnešný* (*today's*) so that, we expand the query by searching for its collocational words (using statistical scoring of *T−score*, *MI−score,* and MI^3−score) and we generate related collocation candidates.

The theoretical background of the concept of collocations has a long tradition in linguistic research regarding collocations as language-specific semantically relevant word combinations [8] and differentiating among them collocations, typical collocations, and idioms. At the same time, the statistical corpus-based research regard collocations as statistically relevant word combinations which can be studied, presented, and extracted using their association, distribution, and combinatorial properties. We accept collocations as statistically relevant words which hold certain semantic relations and which can be extracted by using statistical scoring.

Thus, we generate the collocation candidates of Slovak keyword *dnešný* (*today's*) which first seven ranked results are given in Fig. 2 and include words like *rozprava* (EN - *debate*), *deň* (EN - *day*), *hlasovanie* (EN - *vote*), *doba* (EN - *time*), etc. The related meaningful combinations are *dnešná rozprava* (EN - *today's debate*), *dnešný deň* (EN – *today's (day)*), *dnešnó hlasovanie* (EN – *today's vote*), *dnešná doba* (EN - *today's times*), etc.

	Word	Cooccurrences	Candidates	T-score	MI	MI3
1	rozprava	160	3,199	12.61	8.51	23.16
2	rozpravy	132	2,418	11.46	8.64	22.73
3	rozprave	99	1,526	9.93	8.89	22.15
4	dňa	78	1,130	8.81	8.98	21.55
5	hlasovanie	119	3,091	10.87	8.14	21.93
6	dobe	56	830	7.47	8.95	20.56
7	hlasovaním	48	713	6.91	8.94	20.11

Fig. 2. The collocation candidates of keyword *dnešný* (*today's*) in Slovak EUROPARL7 Corpus (7 coll. Span).

However, for further filtering we need to generate not only statistically similar collocations but also statistically distant words (antonyms), so to obtain all semantic relations of the studied keyword. Thus, we extent the statistical search and generate word sketch difference of keyword *dnešný* (*today's*) (Fig. 3).

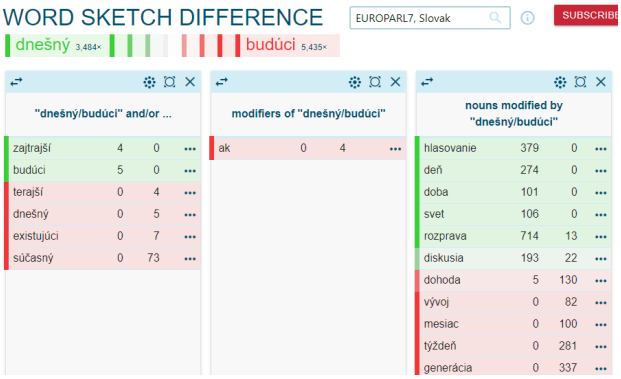

Fig. 3. The word sketch difference of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus.

The results include semantically different (statistically distant) words (antonyms) which are subsumed into two general types of semantic relations – 'nouns modified by' and 'and/or'. The relation 'nouns modified by' presents both the collocations of *dnešný* (*today's*) which are given in Fig. 2) together with the collocations of its antonym *budúci* (EN - *future*) (like *budúca generácia* (EN – *future (next) generation*), etc.). The visual presentation of that relation is given in Fig. 4.

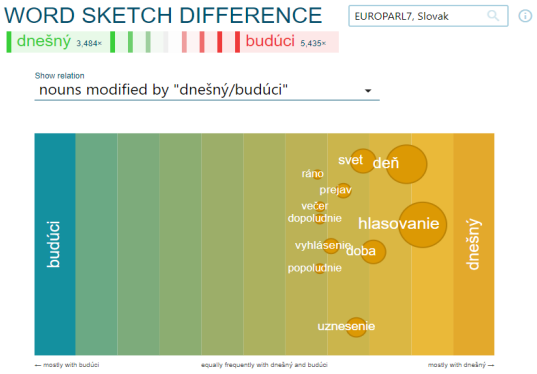

Fig. 4. The word sketch difference of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus for the relation 'nouns modified by'.

The relation 'and/or' presents the antonyms of *dnešný* (*today's*) (like *zajrtajši* (EN – *tomorrow's*) and *budúci* (EN - *future's*)) as well as the antonyms of *budúci* (*future's*) (like *terajši* (EN – *now*), *existujúci* (EN – *existing*), *sučasný* (EN – *current*), etc.) which are synonyms of *dnešný* (*today's*). The related visualization of that relation is given in Fig. 5.

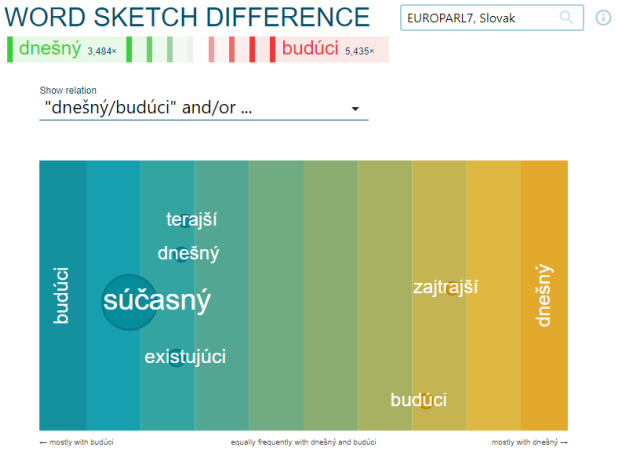

Fig. 5. The word sketch difference of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus for the relation 'and/or'.

Additionally, the visualization of all semantic relations with the keyword's collocational profile (word sketch) including the antonyms (outlined in different colors) is given in Fig. 6.

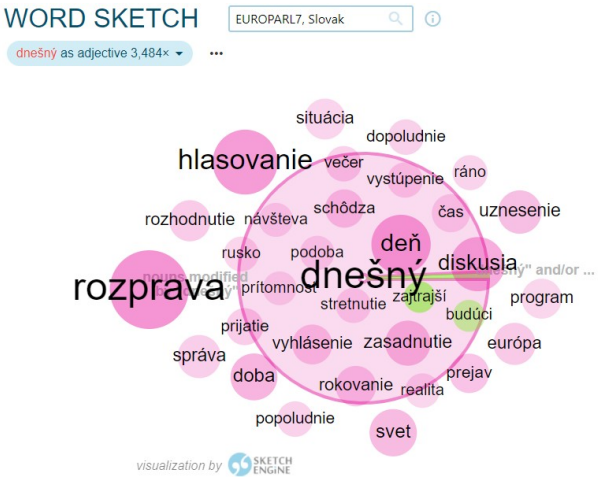

Fig. 6. The word sketch visualization of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus (24 collocations span).

The use of statistical search and retrieval allows further query expansion by searching also for words with common collocations. The search relay on the fact that if two words have common collocations, they share similar meaning. That type of search reveals hidden semantic relations and can give as a result the generation of a thesaurus.

The visualization of search results for words that have common with the keyword *dnešný* (*today's*) collocations is given in Fig. 7 and it presents differently than previous search words which outline more complex semantic relations (hierarchical).

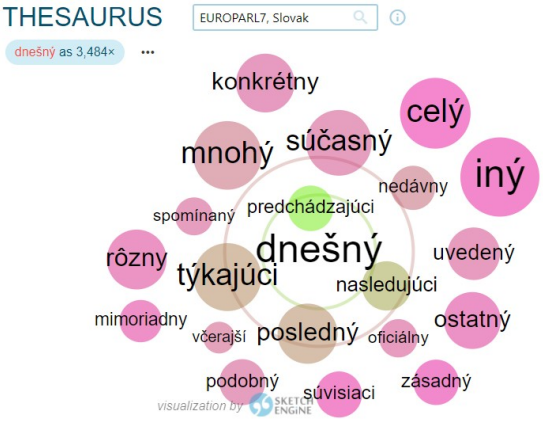

Fig. 7. The thesaurus of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus (20 collocations span).

4.1 Parallel Search Results from Slovak-Bulgarian EUROPARL 7 Corpus

The previous results show that by expanding the collocational search, we obtain more complex

keyword semantic relations. So that, we generate the collocation candidates of the Bulgarian adjective *днешен* (SK – *dnešný,* EN - *today's*) from Bulgarian EUROPARL 7 Corpus using the same scoring. Some previous research shows that the extraction of parallel collocations is a reliable approach to MT [11]. The results are given in Fig. 8 and present similar to Slovak adjective *dnešný* collocations like *разискване* (SK – *rozprava,* EN - *debate*), *гласуване* (SK -*hlasovanie,* EN - *vote*), *ден* (SK – *deň,* EN - *day*), etc.

	Word	Cooccurrences	Candidates	T-score	MI	MI3
1	разискване	337	4,017	18.32	9.07	25.86
2	гласуване	197	2,303	14.01	9.09	24.34
3	дата	99	348	9.94	10.83	24.09
4	ден	88	1,600	9.35	8.46	21.38
5	вот	54	1,304	7.32	8.05	19.56
6	заседание	57	1,573	7.52	7.85	19.52
7	свят	54	2,132	7.30	7.34	18.85

Fig. 8. The collocation candidates of the Bulgarian adjective *днешен (today's*) from Bulgarian EUROPARL 7 Corpus (7 coll. span).

Hence, we suppose that word sketch difference for the Bulgarian keyword *днешен* (*today's*) get similar semantic results as for its Slovak counterpart *dnešný* (*today's*) since both are generated using the same scoring and can be regarded as possible translations.

Consequently, we expand the search for word sketch of keyword *dnešný* (*today's*) with the Bulgarian EUROPARL 7 Corpus performing a parallel bilingual search within both corpora and expecting to obtain the results which include words with the same semantic relations (the same word sketches) but in the Bulgarian language. The results are given in Fig. 9 and include both Slovak and Bulgarian words sharing the same semantic relations ('nouns modified by' and 'and/or').

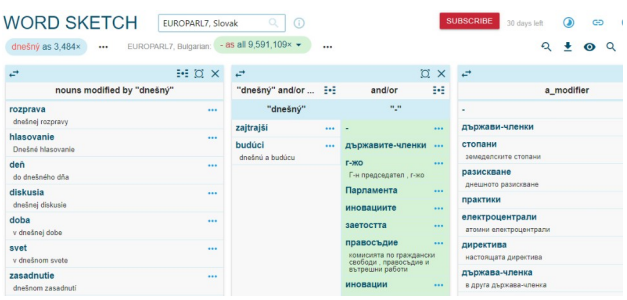

Fig. 9. The parallel bilingual word sketch of keyword *dnešný* (*today's*) from Slovak-Bulgarian EUROPARL 7 Corpus.

Alternatively, the visualization of parallel bilingual word sketch search is presented in Fig. 10.

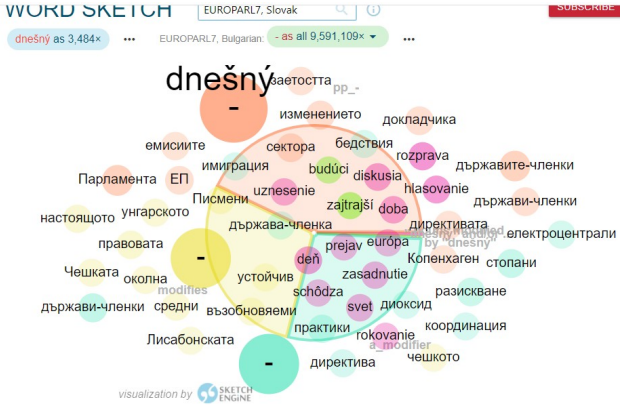

Fig. 10. The visualization of parallel bilingual word sketch of keyword *dnešný* (*today's*) from Slovak-Bulgarian EUROPARL 7 Corpus (12 collocations span).

Following the above assumptions, we can expect that if we generate a parallel bilingual Slovak-Bulgarian concordance of keyword *dnešný* (*today's*), we can get its possible Bulgarian language translations. The results are given at Fig. 11 and present several possible Bulgarian translations of the Slovak adjective *dnešný* (*today's*) like *днешен* (SK – *dnešný*, EN - *today's*), *сегашен* (SK – *terajší,* EN – *now)*, *съществуващ* (SK – *existujúci,* EN – *existing*), *днес* (SK – *dnes*, EN – *now*), etc., and the majority of which were previously generated as word sketch difference synonyms of *dnešný* (*today's*). The results also imply that the generation of various semantic relations (like word sketch difference, word sketch, and thesaurus) suggests more alternative translation equivalents.

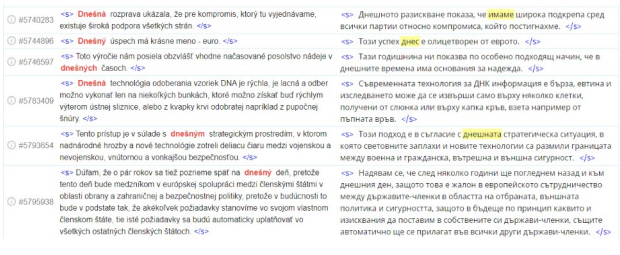

Fig. 11. The parallel bilingual concordance of the word *dnešný* (*today's*) from Slovak -Bulgarian EUROPARL 7 Corpus.

Some previous applications of query expansion for MT [14] use well structured specialized texts and report results only for term extraction however, in our research we present results for translation of political speech trying to enlarge the scope of the application and to elaborate the techniques to obtain more complex types of statistically relevant word associations.

Thus, the collocation search can be expanded also with the n-grams search (3-4-grams) to obtain more typical for the register of political speech collocations. Hence, we generate 3-4-grams of the studied keyword from both Slovak and Bulgarian language parts of EUROPARL 7 Corpus. The related results are presented in Fig. 12 and Fig. 13 and show typically for both languages collocations (SK – *dnešný deň*, BG – *ден днешен*) which can be studied further from a linguistic point of view. A comparison of collocation *dnešný deň* with the Slovak adjectives collocation dictionary [7] shows that it is outlined there as a typical collocation.

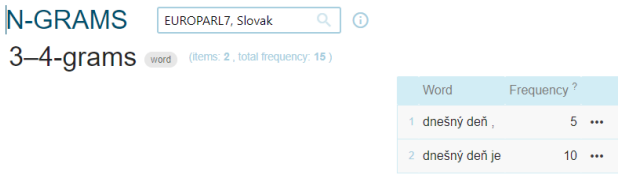

Fig. 12. N-grams search results of keyword *dnešný* (*today's*) from Slovak EUROPARL 7 Corpus.

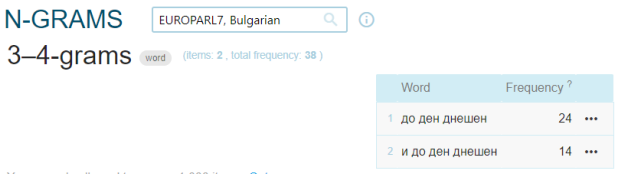

Fig. 13. N-grams search results of keyword *днешен* (*today's*) from Bulgarian EUROPARL 7 Corpus.

5 Discussion

The presented approach uses a statistical scoring in order to evaluate words statistical similarity and distance (relating both to words lexical semantic relations) by accepting them as universal inter-lingual presentation. The technique is based mostly on query expansion and search iterations within a bilingual electronic text corpora, and presupposes a related structure of bilingual linguistic data (corpora sentence alignment), however without structuring, the result would be different.

The advantage of the approach is that it is fast, and does not require formal grammar and lexical encoding of large-scale texts but mostly a bilingual sentence alignment preprocessing. Since the technique used is language-independent, it can be applied for language translation but also for language learning (regarded in AI paradigm as the algorithmic steps) [1].

6 Conclusion

The proposed research presents a statistical approach to MT for Slovak to Bulgarian language translation. The approach uses IR search and retrieval techniques, mainly a parallel statistical search and a query expansion at every algorithmic step to generate concordances, collocations, word sketch differences, word sketches, and thesauri of the studied keyword (query) by using a related SE scoring, which is regarded as intermediate (inter-lingual) standard semantic presentation by means of which the studied keyword (from the source language) is mapped together with its possible translation equivalents (onto the target language).

The result, also, is based on studying the adjectival collocability in both Slovak and Bulgarian language from the corpus of political speech texts, and they may have more complex linguistic interpretation concerning the further study of the use of typical collocations and idioms and their role in political speech.

References:

[1] Almutairi, A., Gegov, A., Adda, M., Arabikhan, F. (2020). Conceptual Artificial Intelligence Framework to Improving English as Second Language. WSEAS Transactions on Advances in Engineering Education, vol. 17, 87-91.

[2] Baroni, M., Evert, S. (2008). Statistical Methods for Corpus Exploitation. *Corpus Linguistics: An International Handbook*, vol. 2, 777–803.

[3] Baroni, M., Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4), 673-721.

[4] Kilgarriff, A., Reddy, S., Pomikalek, J., Avinesh, P. (2010). A Corpus Factory for Many Languages. *Proceedings of LREC 2010,* 904-910.

[5] Kilgarriff, A. et al. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1, 7-36.

[6] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings from MT Summit 2005,* 79–86.

[7] Majchraková, D., Chlpiková, K., Bobeková, K. (2017). *Slovnik kolokácii pridavných mien v slovenčine*. VEDA, SAV, Bratislava

[8] Melcuk, I. (2012). Phraseology in the Language, in the Dictionary, and in the Computer. *Yearbook of Phraseology*, 3, 31-56.

[9] Mutabazi, B., Revesz, P. Z. (2019). A Quantitative Lexicostatistics Study of the Evolution of the Bantu Language Family, WSEAS Transactions on Computers, vol. 18, 97-100.

[10] Nirenburg, S. (1989). Knowledge-based machine translation. *Machine Translation,* 4, 5–24.

[11] Novitskiy, V. (2011). Automatic Retrieval of Parallel Collocations for Translation Purposes. Pattern Recognition and Machine Intelligence,. *Lecture Notes in Computer Science*, 6744, 261–267, Springer

[12] Stoykova V., Šimková M., Majchráková D., Gajdošová K. (2015). Detecting Time Expressions for Bulgarian and Slovak Language from Electronic Text Corpora. *Procedia Social and Behavioral Sciences*, 186, 257-260, Elsevier

[13] Stoykova, V. (2016). Using Statistical Search to Discover Semantic Relations of Political Lexica – Evidences from Bulgarian-Slovak EUROPARL 7 Corpus. Mathematical Aspects of Computer and Information Sciences, *Lecture Notes in Computer Science*, 9582, 335–339, Springer

[14] Stoykova, V., Stankovic. R. (2019). Using Query Expansion for Cross-Lingual Mathematical Terminology Extraction. Artificial Intelligence and Algorithms in Intelligent Systems, *Advances in Intelligent Systems and Computing*, 764, 154-164, Springer

[15] Tarasov, D. (2020). Language Attribution of an Unmarked Text Corpus. WSEAS Transactions on Systems and Control, vol. 15, 754-759.

[16] Zhang, P. (2019). Evaluation of Two Different Models in Neural Machine Translation. WSEAS Transactions on Information Science and Applications, vol. 16, 87-93.

Sources of funding for research presented in a scientific article or scientific article itself

The reported study was done within the framework of the project "*IT-based Analyses of Bulgarian – Slovak / Slovak – Bulgarian Lexical Data*" (2018-2021) within the Bulgarian Academy of Sciences grant scheme and the bilateral agreement with the Slovak Academy of Sciences.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)