IT TOOLS IN DISCOURSE ANALYSIS: STATE OF AFFAIRS AND PROBLEMS

[a]IGOR O. GURYANOV, [b]RADIF R. ZAMALETDINOV, [c]ELZARA V.GAFIYATOVA, [d]GERMAN I. LASHEVICH, [e]OLGA S. SAFONKINA

[a]Kazan Federal University, Kazan, 18 Kremlevskaya Street, Kazan, 420008, Russia
[b] Kazan Federal University, Kazan, 18 Kremlevskaya Street, Kazan, 420008, Russia
[c]Kazan Federal University, Kazan, 18 Kremlevskaya Street, Kazan, 420008, Russia
[d]Kazan Federal University, Kazan, 18 Kremlevskaya Street, Kazan, 420008, Russia
[e] Ogarev Mordovia State University, ul.bogdana khmelnitskogo, 39, Saransk,capahack , 430005, Russia

email: [a]igor.goor@gmail.com, [b]r.r.zamaletdinov2017@gmail.com, [c]e.v.gafiyatova2016@gmail.com,[d]german.i.lashevich2015@gmail.com, [e]o.s.safonkina2016@gmail.com

Abstract: The article deals with the analysis of statistic instruments used for the search and distinction of modified idioms in the discourse. At the present time the analysis with the help of statistic instruments is fundamental to all experiments. The article presents the study of main research papers as well as the achievements made by internationally well-known scholars in the field of linguistics and famous mathematicians. The research analyses the impact made by influence of the creation of new area of research, linguistic statistics. The article also gives ideas of the application of statistical soft wares, such as Google Ngram viewer, Ruscorpora and Mystem, for the analysis of the use of modified and non-modified bookish idioms in the discourse of different languages, such as Russian, English and German. The research material was selected from monolingual and bilingual phraseological dictionaries. The results of the study are supported by the examples taken from applied software. The materials of the study may be useful worldwide by educators and researchers involved in professional linguistic research and training.

Keywords: modified idioms, corpus linguistics, discourse, statistics, Ngram, grammatical modifications, quantitative linguistics.

1 Introduction

In linguistics there are different applications that contribute to the implementation of different research that requires analysis of texts. Large companies such as “Google” and “Yandex” provide applications that contribute to linguistic analyses of different textual unities. Some mechanisms of search engines that these companies use to provide their services are in open-access. XXI century is officially recognized as the age of computers. Computing resources contribute to processing of vast amount of data. Therefore, to start with we consider to review the origins of linguistic statistics and linguistic mathematics in Russia to analyze the development of this area of research for the period of one and a half century.

“The language is … a socio-historic phenomenon reflecting social events and the structure of the society” (Solnyshkina et al, 2014). The first ideas of applying mathematical methods for studying languages were introduced in the XIXth century by I.A. Badouin de Courtenay, a prominent scholar of Kazan Linguistic School (Jakobson ,1972). However practical implementation of statistic processing of the language did not start until the beginning of the XXth century. It was V.Ya. Bunyakovskiy and A.A. Markov who contributed to the development of creation statistical methods applied to the language functioning. In 1847, the journal “Sovremennik” published an article titled "On the Possibility of Introducing Determinative Confidence Building Measures to the Results of Some Sciences of Observational and Primarily Statistics" by V. Ya. Bunyakovskii (Sheynin,1991). In this work the author shared his thoughts on the application of mathematical methods for studying Grammatical and Etymology of the language within the paradigm of Comparative Linguistics.

Later, in 1906 A.A. Markov introduced “Markov’s Chain” or “Markov’s Process” by (Rozanov,2012). This is a variation of probability theory where the future and the past of the system are independent of each, so we do not take into account the history of the system and we make our prediction on present state of the system. As that reminds the functioning of such advanced systems as languages, “Markov’s Chain” firstly was applied for linguistics, mainly in textology.

Only in the XX century with the advent of structural linguistics, scholars began to apply mathematical methods for the analysis of linguistic processes much more broadly. Scholars considered language elements as a suitable illustrative material for constructing quantitative methods of research or for the promotion of statistical theorems, since according to certain regularities the language functions can be identified and confirmed by statistical calculations. However scholars did not focus on considering these results from linguistic point of view and only later, in the middle of the 50ies of the XXth century, an American scholar George Zipf established the relationship between the frequency of the word in the text and its rank in the dictionary. This dependence was called “Zipf’s Law” (Powers,1998), which is applicable to linguistic statistics basing on the assumption that any linguistic unit has a certain degree of probability of being used in the text. The identification of the degree makes it possible to define the use of a given unit. Linguistic statistics treats the text as a sequence of units on a fixed level.

2 Methods

The goal of the paper is to evaluate the possibility of applying quantitative linguistics methods to the search of modified idioms. The research was conducted on the bookish idioms selected from monolingual and bilingual phraseological dictionaries. During the research, the following resources as “Google Ngram viewer”(2017), “Ruscorpora” (2017) and “Mystem” (2017) by Yandex were used.

N-gram is a sequence of n-elements. The scope of this term is broad coverage of the areas from theoretical mathematics to genetics. From the semantic point of view, it can be a sequence of sounds, syllables, words or letters. A sequence of two consecutive elements is often called a bi-gram, a sequence of three elements is called a trigram. This sequence allows to represent any phrase in the form of a mathematical sequence, of two elements and find the complete correspondence of these units in the text. N-grams suit well to classify and make a structure of any text. As we know “Automatic event extraction is an important task in knowledge acquisition step” (Solovyeve et al, 2016).

The application of the methods of quantitative linguistics with idioms is a complex task. This is due to the fact that an idiom, the main subject of the study can undergo certain modifications in the context. To illustrate that E.F. Arsenteva (Arsenteva et al, 2016) distinguished 12 types of modifications that can be roughly divided into grammatical and semantic modifications though in the context we may observe fusion of these modifications. This situation makes it difficult to apply statistical programs in fully automated mode to analyze how idioms are used in the text.

During our research, we selected 30 bookish idioms from the English-Russian dictionary edited by A.V. Kunin (1984), 30 units from the Russian-English dictionary edited by S.I. Lubenskaya (2004) and 30 units from German -Russian dictionary edited by L.E. Binovich (1995).

It is noted that a significant number of bookish idioms include composition components related to obsolete vocabulary, which resulted in a significant raise of number of units in XIX or the beginning of XX century. To prove this hypothesis, we turned to two resources that are used in the processing of statistical language data.

Thus, the main source of the collection of statistical information for the study of the functioning of Russian idioms is the national

corpus of the Russian language (ruscorpora.ru) . The advantage of Russian corpus is that the frequency of use of each unit can be considered both synchronically and diachronically. The most important feature is the ability to adjust the distance of the components from each other, which allows you to search for units transformed by wedging, breaking, and lexical replacement.

An important aspect that affects the convenience of processing and using the program is the tab "statistics" and "distribution by years". The system constructs a graph of the frequency of use of a given unit on the time axis. In the "statistics" tab, we observe statistics on meta attributes, such as the stylistics of the text, and the author's gender, which makes the research attractive from gender linguistics' point of view of, as well as other indicators. A meta attribute is a sign or signs by which the body is marked. The main meta attributes are laid by the developers of the corps at the development stage.

It should be noted that, despite a significant amount of statistical data, the obtained results require careful verification, since among the units with a rethought meaning there may be free word combinations. This feature can be manifested by setting a large interval between the components of a unit. In our study, we limited ourselves to the maximum distance between components of 7 words.

The third program "Mystem" provided by Yandex has the least user friendly interface used in the console mode. Although it has the most flexible and advanced functions, we can input a large text that would be analyzed by the program. In this case the output will be in the .txt extension. Each word of the text can receive a certain characteristics (gender, number, case, etc.). However, it still requires extra review of the context and distinction of idioms from word-combinations.

3 Results

The first program Google Ngram Viewer allows to track the frequency of the use of words and phrases, as well as idioms. The program searches through the vast library of Google Books and notes the number of units frequency. A distinctive feature of the resource is that a large number of languages are available, namely English, Russian, Spanish, German, and Chinese. With the help of this program, we state that a significant number of phraseological units of the book style refer to obsolete expressions. In English, for all selected units, the raise of particular use referred to the XIX century and to the beginning of XX. For example: The wings of Azrael, was used mostly in 1923; Appeal to Caesar, in – 1824. In English, we can see such a specific feature that, when the article is added to the construction of phraseology, the raise of use shifts to the side.

In German discourse we observed the same trend. Bookish idioms tend to raise in the XIX century. Thus, Eine Dornenkrone tragen ("*to wear the crown of thorns*") was mostly used in 1832; Einklang der Herzen ("*harmony of hearts*") in - 1826; Goldenen Zeiten entgegengehen ("to meet golden times") in - 1810.

In Russian discourse we observed idioms that tend to reach the maximum use in the XIX century. Обетованная земля (obetovannaya zemlya) (*"promised land")* was mostly used in 1847; Почивать на лаврах (pochivat' na lavrah) ("*to rest on laurels*") in - 1824. According to E.V. Gafiyatova "cultural literacy allows independent use of communication tools and knowledge" (Gafiyatova,2014). Although we noticed that nearly half of the Russian bookish idioms have peak of use in XXI century: (potemkinskie derevni) ("Potemkin's villages) – idioms refers back to the history of Russia. Duke Potemkin built fake villages in the region he was responsible for to convince Empress Catherine II that region has a sustainable economic growth" peak of use 2007 and still grows; Заплечных дел мастер (Zaplechnyh del master) ("torturer") was used in 1997 year, mostly.

4 Discussion

While investigating the frequency of the use of Russian idioms, we faced with the of correct spelling of phraseological units took place since a spelling reform in 1917-1918. The Google Ngram Viewer database consists of original written sources translated into electronic form, which complicates the detection of the frequency of use of idioms. The statement that a significant number of bookish idioms belongs to the outdated layer of vocabulary was confirmed only in English and German, with the highest use of 30 selected units took place in the XIXth and first half of the XXth century. In Russian, on the contrary, there was a tendency towards the growing popularity of some units.

According to E.F. Arsenteva's (10) classification, there are the following types of transformation of idioms (Arsenteva and other scholars use the term Pharseological units (PU)): 1. Substitution or replacement of a PU Component (s); 2. Permutation; 3. Addition; 4. Insertion; 5. Cleft use, which is interrelated with insertion; 6. Deletion of a PU component (s) or ellipsis; 7. Phraseological allusion, which is closely connected with ellipsis; 8. Phraseological reiteration; 9. Phraseological pun; 10. Contamination of two PUs; 11. Extended phraseological metaphor; 12. Phraseological saturation of discourse.

Within these transformations we can trace in semi-automated mode with only grammatical modifications such as: substitution or replacement of component, permutation, addition, insertion, cleft use, and deletion of a PU component (s) or ellipsis. Other modifications require analysis from a scholar as modifications deal with semantics and cognition of idioms. As D.N. Davletbaeva states "that the mechanisms of cognitive processing of figurative base of a phraseological unit work simultaneously, as they are responsible for different aspects of nonce phraseological unit meaning. Figurative base presents not only base for conceptualization and categorization of objective reality but also the emotion stimulus, motivating stimulus, a "hint" for cultural interpretation of the meaning, causing native speaker's emotive attitude" (Davletbaeva et al, 2015).. These features of cognition of idioms and their simulation by the computer are the goals of computer linguists and AI specialist.

As for the use of "Mystem" we faced the same problems when the programme provides data by the each unit but cannot distinguishes the idioms from word combinations although grammatical information can be useful in analysis of the idioms and distinction of the core component of each unit.

5 Conclusion

It is worth noting the feature of using Google Ngram Viewer. In the search window, only conventional units can be entered, and the search for sematically modified idioms is impossible at that stage of application development.

We stated the following problems of using Google Ngram Viewer programme for the search of modified idioms. However, the following problems occur, namely that the program cannot distinguish an idiom from free word-combination, so it is worth limiting research to structural modifications.

1. When we search for insertion as a type of modified idiom, its components can be separated within the sentence, sometimes the gap is beyond the sentence. As a result, the program can skip this modification.
2. Modern software makes it easier to find phraseological units and how to use them. The main problem in the search for phraseological units and their transformations is the impossibility for the program to distinguish idiom from word-combination.
3. As a result, the programs for deducing statistics, for example, determining the frequency of the use of the idiom greatly facilitate the work of the scientist in the study of phraseological units in the diachronic aspect. Nevertheless, there is a number of developments and tools that can

facilitate and accelerate the processing of large amounts of data, and in particular, the use of information retrieval systems.

Acknowledgement

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Solnyshkina M, Gafiyatova E. ,*Modern forestry English: macro- and microstructure of low register dictionary*. Journal of Language and Literature. 2014- Vol. 5. No. 4. pp. 220-224.
2. Jakobson, R. *The Kazan school of Polish linguistics and its place in the international development of phonology*. In: Jakobson, R. (ed), Selected Writings. Vol. II: Word and Language. Hague: Mouton.1972.
3. Sheynin O. B. ,*On V. Ya. Buniakovsky's work in the theory of probability*. Article. Archive for History of Exact Sciences.1991.
4. Rozanov Y.A. ,*Markov Random Fields. Springer Science & Business Media*. 2012, p. 58. ISBN 978-1-4613-8190-7.
5. Powers, David M. W. (1998). *Applications and explanations of Zipf's law*. Association for Computational Linguistics.
6. Ngram Viewer. URL: https://books.google.com/ngrams (accessed 19.05.2017)
7. Nacional'nyj korpus russkogo yazyka. URL:http://www.ruscorpora.ru/ (accessed 16.05.2017)
8. Mystem. URL:http:// tech.yandex.ru/mystem (accessed 16.06.2017)
9. Solovyev V., Ivanov V ., *Knowledge-Driven Event Extraction in Russian: Corpus-Based Linguistic* /Resources Hindawi Publishing Corporation Computational Intelligence and Neuroscience. 2016-Volume p. 1-10
10. Arsenteva E.F, Arsentyeva Y.S. ,*Discoursal analysis of phraseological of euphemisms: Experimental data in teaching English*// Journal of the Social Sciences. 2016-Vol.11, Is.6.
11. Kunin A.V. ,*Dictionary of English phraseological units*. – Moscow.: Russkiy Yazik.1984.
12. Lubenskaya S.I. *Russian-English phraseological dictionary*. – Moscow: AST-Press kniga,2014.
13. Binovich L.E. ,*German-Russian phraseological dictionary*– Moscow: Akvarium.1995.
14. Gafiyatova E., Pomortseva N . T*he Role of Background Knowledge in Building the Translating/Interpreting Competence of the Linguist* // Indian Journal of Science and Technology,2016, Vol 9(16), p 2-11
15. Davletbaeva D.N, Ivanova A.M, Kozlova Y.A. *Psycholinguistic criteria for understanding phraseological units*. Mediterranean Journal of Social Sciences. 2015-Vol.6, Is.4S2.