QUANTITATIVE AND QUALITATIVE METHODS OF CORPUS TEXT ANALYSIS

[a]DIANA I. KHRIPKOVA, [b]GALINA V. KOLPAKOVA, [c]DIANA F. KAJUMOVA, [d] ELENA A.OGANOVA

[a]*Kazan Federal University, Leo Tolstoy Institute of Philology and Intercultural Communication , 18 Kremlevskaya Street, Kazan, 420008, Russia.*
[b]*Kazan Federal University, Leo Tolstoy Institute of Philology and Intercultural Communication, 18 Kremlevskaya Street, Kazan, 420008, Russia.*
[c] *Kazan Federal University, Leo Tolstoy Institute of Philology and Intercultural Communication, 18 Kremlevskaya Street, Kazan, 420008, Russia.*
[d] *Lomonosov Moscow State University, Ulitsa Leninskiye Gory, 119991, Moskva, Russia.*

email: [a]missdiana7@mail.ru, [b]Rusia@Prescopus.Com, [c]d.g.kajumova2015@gmail.com ,[d] a.elena2015@gmail.com

Abstract:The article is devoted to the identification and classification of semantic distribution, models of (speech environment models) German synonyms naming person, and comparative analysis of the frequency of the speech realization of German synonyms ("Textkorpora" on the computer program material) and their fixation in the German synonymous dictionaries. The stages of analysis of computer program "Textkorpora" with the use of quantitative and qualitative methods of corpus analysis are presented in the article. The results of the body of research speech functioning of German synonyms, names of persons, in terms of their comparison with the data of lexicographical sources are developed in the article. A classification of semantic distribution models of German synonyms, identified as a result of the semantic interpretation of the program "Textkorpora" are established, the general patterns of speech implementation of German synonyms, names of persons, and various trends in their functioning. The semantic interpretation of "Textkorpora" computer program is an experience with the methods for quantitative and qualitative analysis of the text corpus. Analysis of the program "Textkorpora" leads to the following conclusion: the presence of similar structurally models the speech of German synonyms environment does not guarantee them the same type of qualitative implementation of the German synonyms, members of the same synonymous row. Synonymous dictionaries reflect the cash fund language synonymous units in the system of language, in the lexicon (static aspect), but do not reflect the linguistic reality using of speech units in speech (dynamic aspect). Results of the analysis of verbal functioning of German synonyms, names of the person (on the computer program material "Textkorpora") not only indicate significant fluctuation frequency of speech realization of German synonyms, but they are members of one synonymous row, and identify the voice of preference in the use of synonyms members of the same linguistic community, but and evidence of "mobility", the changes over time of the formation synonymous vocabulary. Article submissions, the results of the study are of interest for scientists studying the corpus relying on statistical methods, bypassing the stage of pre-formulated hypotheses, applying qualitative analysis method for the semantic interpretation corpus only at a subsequent and final step.

Keywords: quantitative method, qualitative method, corpus, IdS "Textkorpora" program, semantic distributive model, the realization of speech.

1 Introduction

Considerable interest of linguists to the lexicographical research, in-depth study of the paradigmatic relations in the lexical system of language, the analysis of the semantic structure of the word due to the wide spread of cognitive and corpus linguistics, application of computer technologies in the processing of texts (Sprachkorpora, 2007; Lexikalische Semantik und Korpuslinguistok, 2006). *The topicality* of the study is the fact that the alleged problems in the article, were studied with a large array of text corpus in the computer program "Textkorpora" des IdS als empirische Basis für die linguistische Forschung" with the use of modern methods of corpus linguistics (quantitative and qualitative analysis). *The purpose of the study* is to detect the *lexicographical reflection* of synonymy (on materials of synonyms dictionary by H.Gerner and G.Kempke the editions of 1974, 1984, 1999 years (Synonymwoerterbuch, 1974; Synonymwoerterbuch, 1984; Woerterbuch Synonyme, 1999) and the frequency of "the natural speech" verbal use of German synonyms, names of person in the language of reality. The computer program of the Institute of Mannheim German "Textkorpora der IdS als empirische Basis für die linguistische Forschung" contains 1.736 billion words of text with examples totaling 4340000 book pages. Niladri Sekhar Dash throws corpus linguistics as a multidimensional area. Corpus Linguistics is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication and comprehension (Niladri Sekhar, 2010).

A wide range of technical options for the creation of electronic enclosures, and large amounts of experimental facts, electronically prompted scientists to find effective methods of analysis of texts using computer technologies. According to L.M.Ayhinger "case studies are an attempt to get closer to the linguistic reality through statistical analysis and mathematical modeling in abstraction from the linguistic competence of the subject and his method of introspection" (Eichinger,2007). "Quantitative and qualitative methods of analysis applicable to the corpus of texts and annotations to the levels in the body" (Safina et al, 2015). Abstract K.Sherer in treatment is "an additional structural information that goes beyond the text in the body, encoded in the text by means of special markings" (Scherer,2006). With abstract implicit information contained in the text, translated into explicit form, that speeds up the process of collecting information in the annotated case. The basis of any quantitative analysis, believes A.Lyudeling is qualitative analysis and categorization of data. Quantitative analysis depends on the preceding stage of categorization, the latter is not always indisputable, as often in studies based on quantitative analysis of the buildings, there is no information about the categorization conducted employed categories, categories of selection criteria (Luedeling,2007). Mick O'Donnell presents language corpora used to enhance the classroom experience in several ways. For the teacher, corpora can be used to improve their understanding of classroom interactions, and of the language learning process itself. Alternatively, a corpus of student writings can be explored to identify what students do wrong, and thus target teaching to their problem areas (Mick O'Donnell,2010-2012).

2 Methods

In the first stage of the study, using the method of quantitative analysis, we can learn to what extent the lexicographical practice reflects the linguistic reality. Due to the significant array of text examples of "Textkorpora" program and given the high frequency of the speech use of German synonyms (eg, Hase "coward" - 1874, Fuchs is "cunning" - 10023), we chose to study a fragment of this program, that includes the 9 most common synonymous series (SS), whose members represent the quality of human characteristics (SS Ange-ber - proud, SS Schmeichler - smoothie, SS Kriecher - groveller, SS Heuchler - hypocrite, SS Lügner - liar, SS Geizhals - miser, SS Feigling - coward, SS Schlappschwanz - gruel, SS Duckmäuser - demure). The total number of text examples demonstrating the uses of the 9 members of the Set-governmental synonymous series, was about 14,200 texts. As a result the characteristic changes in the lexical composition and structure of synonymous series in three editions of the German dictionary of synonyms and H.Gernera G.Kempke (1974, 1984, 1999.) were given. The comparative analysis of these words-ray showed relative stability synonyms for the duration length of time. Only in the last edition of the thesaurus (1999) some changes in the structure and lexical filling of some SS were recorded.

In the second phase of the study the method of qualitative analysis was used to identify and describe the distribution of semantic models of the German synonyms members 9 analyzed synonymous series. As a result of the semantic interpretation of the computer program "Textkorpora" 7 semantic distributional patterns were identified: 1) adjective, pronouns, participles in an agreed definition of the function, 2) prepositional phrases in the function of inconsistent definitions (mostly in postposition), 2) prepositional phrases in an inconsistent definition of the function, 3) subordinate attributive sentence specifying the meaning synonym often represents its semantic motivation, 4) semantic correlates of having identical with synonymous or different from the reference classifying, 5) the transfer of semantic correlates with identical or different reference attributing, belonging to different CP 6) the main focus is on the

semantic functions (identification, generalizing, quality-term characteristics of a person) of the German article, combined with a synonym, in the naming of a person-having reference assignment to a specific person, group of persons, an indefinite multitude of individuals, 7) the game of words. The most common model of the speech environment are semantically close and distant correlates synonym. Enumerating various semantically correlates with the assignment of a single reference in the text to actualize the potential seme. The semantics of the individual as a synonym has a regulating effect on the environment synonym, prejudging its possible range of contextual semantic correlates. A typical model of the speech environment of lexemes Schönredner – "talker, Smoothie", Kriecher – "under-Halim, sycophant, groveler", Speichellecker – "sycophant, flatterer, meanly due", Feigling – "coward" is a comparison with different semantic correlates having identical or different from their reference concerns, for example: Er ist ein Miesmacher und Angsthase, ein Weichling, Defätist und Nichtsnutz. "He's a whiner"' (skeptic knocker) and a coward, sissy, defeatist and a slacker (parasite, it-dyay, idler). A token Charmeur "charming man" is found in the enumeration semantically different nouns, comprehensively characterized by a human and having identical with identical Charmeur referenting relatedness: Als witziger Charmeur, souveräner Frackträger und eleganter Gentleman-Abenteurer knüpfte er fast nahtlos an die Karriere seines Vaters an. "How witty charming, sovereign, wearing a suit and elegant man-tion gentleman-adventurer, he slowly continued his father's career." There was no single case of the use of words among Charmeur having different reference relatedness. Comparison with semantic correlates not typical tokens der Geizige – "stingy", Geizkragen – "hunks", Pfennigfuchser – "cheapskate, cheapskate," relating to the CP der Geizige – "miser" and tokens Pharisäer – "Pharisee, the hypocrite," a member of the CP Heuchler – "a hypocrite."

3 Results

According to the results of the comparative analysis on the materials of program "Textkorpora" the discrepancies between the frequency of the speech realization of synonyms and their markedness in a dictionary of synonyms of H.Gerner and G.Kempke (1999) were revealed. For example, a synonym Drückeberger, expelled from the SS Feigling - coward of the dictionary, is used in 146 texts, and included in the SS Feigling - coward of the same dictionary synonyms Wagenichts, Trauminet not recorded in any of the text. 6 synonyms members of SS Schmeichler - Smoothie (Schmuspeter, Schmuskopf, Fuchsschwänzer, Flaumstreicher, Schmeichelzunge, Schmeichelkatze), were not used in the texts of the program even once, although they are on-labeled in a dictionary of synonyms of H.Gerner and G.Kempke (1974 and 1984's edition). Significant differences in the frequency of using the synonyms were revealed. They were the members of the same synonymous series. Thus, the frequency of using the members of the SS Heuchler - hypocrite ranges from 1177 (Biedermann) to two occurrences (falscher Fuffziger), frequency of use of the dominant Heuchler is to 1736, members of the CP Pharisäer - 70, Wolf im Schafspelz - 77 cases of abuse.

As a result of the semantic interpretation of the computer program "Textkorpora" 7 semantic distributional patterns were identified. The most common model of the speech environment are semantically close and distant correlates synonym. Enumerating various semantically correlates with the assignment of a single reference in the text to actualize the potential seme. The semantics of the individual as a synonym has a regulating effect on the environment synonym, prejudging its possible range of contextual semantic correlates.

4 Discussions

Q.Cai, J. Zhang consider that corpus linguistics is a newly developed subject with its specific characteristics and can be widely used in many aspects of language research and application (Qiang Cai, 2013) Many linguists throw corpora as very important for computational linguistics and offer a survey of how corpora for computational linguistics is "currently used in different fields of the discipline, with particular emphasis on anaphora and coreference resolution, automatic summarisation and term extraction" (Orasan et al, 2007). M.Tymoczko provides a discussion of the centrality of corpus-based studies within the entire discipline of translation studies (Maria Tymoczko,1998).

The debate is the problem of "experimental data" in linguistics. Linguists trust introspection as a method of study and consider the real facts of how inaccurate reflections of abstract principles, see the analysis of the buildings further opportunity to expand the theoretical knowledge of the language (Eichinger, 2006). Y.Asmussen treats the body as a large digital collection of texts, "serving as a representative sample for a specific and in general context is the target sample of language in general" (Asmussen, 2007). But modern housing, believes Kr.Lemann may also include language material, specially created for the body. The requirement for it texts and their existence as private collections preserved today, the requirements of the exhaustion of the body and the natural existence of the texts prior to their analysis, void (Lehmann et al, 2007).

5 Conclusion

The results of the semantic interpretation of a computer program "Textkorpora" using the methods of quantitative and qualitative analysis show that the lexicographical practice does not reflect the linguistic reality: if dictionaries of synonyms fix a non-volatile, stable character fairy-nomen language system synonyms, while the studing the speech synonyms of a computer program "Textkorpora" the dynamic character of verbal synonymy was found. It is in constant change and development. If you have a common structural semantic models of the German Synonyms, names of the person, the members of the 9 various synonymous series, that we analyzed and characterized by a qualitatively distinct features of speech realization of these models. The research is performed according to the Russian Governement Program of Competitive Growth of Kazan Federal University.

Acknowledgement

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

References

1. Asmussen J. *Korpusliguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen* // Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Institut fuer Deutsche Sprache. Jahrbuch 2006. Berlin, New Jork: Walter de Gruyter, 2007. S.123-151.
2. Eichinger L.M. *Linguisten brauchen Korpora und Korpora Linguisten // Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Institut fuer Deutsche Sprache. Jahrbuch 2006. Berlin, New Jork: Walter de Gruyter, 2007. S. 1-8.
3. Lexikalische Semantik und Korpuslinguistok. *Tuebinger Beitraege zur Linguistik*. Bd. 490. Tuebingen: Narr, 2006. 498 S.
4. Luedeling A. *Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik* // Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Institut fuer Deutsche Sprache. Jahrbuch 2006. Berlin, New Jork: Walter de Gruyter, 2007. S. 28-48.
5. Lehmann Chr. Daten. Korpora. Dokumentation // *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Institut fuer Deutsche Sprache. Jahrbuch 2006. Berlin, New Jork: Walter de Gruyter, 2007. S. 9-27.
6. Mick O'Donnell "*Corpus linguistics and the application of new technologies in the foreigh language classroom: Part 1: Corpus linguistics and the foreigh classroom*", 2013-14, p.9.// Masters Programme in Applied

Linguistics. Madrid: Departamento de Filología Inglesa, 2011- 2012, P. 1-9
7. Maria Tymoczko *"Computerized Corpora and the Future of Translation Studies"*// Meta : journal des traducteurs. Meta: Translators' Journal, vol. 43, n°4, 1998, P. 652-660
8. Niladri Sekhar Dash *"Corpus Linguistics: A General Introduction"// Proceedings of the Workshop on Corpus Normalization.* LDCIL, CIIL, Mysore, India, 2010, P.1-25
9. Orasan C., Ha L.A., Evans R.J., Hasler L., Mitkov R. *"Corpora for computational linguistics"*// Ilha Do Desterro: A Journal of English Language, Literatures in English and Cultural Studies. Special Issue in Corpus Linguistics, 52 (1), 2007, P. 65-102
10. Qiang Cai, Jianping Zhang "*Corpus Resources and Their Use in English Teaching*"// Paris, France: Atlantis Press, 2013, P.137.
11. Sprachkorpora – *Datenmengen und Erkenntnisfortschritt Institut fuer Deutsche Sprache.* Jahrbuch 2006. Berlin, New Jork: Walter de Gruyter, 2007. 265 S.
12. Synonymwoerterbuch. *Sinnverwandte Ausdruecke der deutschen Sprache (Hrsg. von H.Goerner u. G.Kempcke).* Leipzig: VEB Bibliographisches Institut, 1974. 643 S.
13. Synonymwoerterbuch. *Sinnverwandte Ausdruecke der deutschen Sprache (Hrsg. von H.Goerner u. G.Kempcke).* Wiesbaden: Drei Lilien Verlag, 1984. 643 S.
14. Woerterbuch Synonyme *(Neu bearb. u.hrsg. von H.Goerner u. G.Kempcke).* Muenchen: Deutscher Taschenbuchverlag, 1999. 818 S.
15. Safina R.A., Varlamova E.V., Tulusina E.A. "*The stylistic potencial of the contextual usage of phraseological units as hybrid formation*// Asian social science, Volume 11, Jesue 19, 2015, P.65.
16. Scherer C. *Korpuslinguistik – Kurze Einfuehrungen in die germanistische Linguistik.* Bd. 2. Heidelberg: Universitaetsverlag Winter, 2006. 98 S.