TOWARD MEASURING LINGUISTIC COMPLEXITY: GRAMMATICAL HOMONYMY IN THE RUSSIAN LANGUAGE

Olga Nevzorova, Alfiya Galieva, Vladimir Nevzorov

Abstract*: Currently linguistic complexity is one of the most debatable concepts in linguistics, and there are different ways of understanding this complexity depending on linguistic domains, research aims and theoretical background. We proceed from the assumption that linguistic complexity becomes apparent in those parameters that can be measured. Grammatical homonymy is an important manifestation of structural complexity of a language, and many aspects of it are computable.*

The study of grammatical homonymy from the point of view of linguistic complexity requires development of appropriate methodology. We examined this phenomenon on linguistic data of the extended version of A. Zaliznyak dictionary using the software of Ontointegrator system. We distinguished four structural parameters that enable to disclose statistical aspects of grammatical homonymy relevant for language processing. The distribution of grammatical homonyms manifests basic typological features of a language.

Keywords*: grammatical homonymy, linguistic complexity, the Russian language, word forms, parts of speech.*

ITHEA Keywords*: H.3.1 Content Analysis and Indexing.*

Introduction

The phenomenon of linguistic complexity (language complexity) became one of the topics of great importance in linguistics in the last decades. Various researches represent dissimilar ways of theoretical understanding of the phenomenon of complexity and propose different parameters of measuring and different ways of practical evaluation of this complexity ([Kusters, 2003; Dahl, 2004; Bane, 2008; Gil, 2008; Juola, 2008; Miestamo, 2008; Newmeyer, 2014; Becerra-Bonache, 2015] and other works).

Modern science worked out different approaches to determine the complexity of an object, and these approaches may be reduced to two basic types:

1) Complexity as a characteristic of objective (structural, dynamic and other) properties of the system;
2) Complexity as a characteristic of the process of cognition and studying an object, rather than of the system.

In [Rastrigin, 1981] the following basic properties of complex objects are specified:

1. Lack of necessary mathematical description.
2. "Noisiness" of complex systems which is evoked not by special generators of random hindrances, but rather by individual complexity of an object and by resulting inevitable abundance of secondary processes, so the object behavior seems in many cases unexpected to the researcher.
3. Intolerance to external control.
4. Non-stationarity of the complex system that shows up in drifting characteristics of the system, in changing its parameters, and in evaluation of the system in time.
5. Impossibility in many cases to reproduce the experiments, due to "noisiness" and non-stationarity of the complex system [Rastrigin, 1981].

The properties of complex objects named above, are generally applicable for characterizing natural languages, nevertheless, with certain provisos. In particular, experiments on natural language are irreproducible in the sense that results of analyzing diverse texts and diverse text collections may significantly differ, which is caused by complicated interaction of systemic, functional, individually authored and other factors. With respect to external influences, different subsystems of the language behave differently, so we can distinguish two types of these subsystems:

– open ones – vocabulary (languages easily accept new words), lexical semantics (words of a language get new senses);
– closed ones – from a synchronic viewpoint grammar is a closed system, because new grammatical categories and grammatical meanings hardly emerge.

Word formation may be regarded as a borderline domain: new words appear with ease, but as a rule, only in derivation models that are admissible for the language system itself.

Putting the question of linguistic complexity requires development of definitions and objective criteria of this complexity. Apparently, the degree of complexity may significantly differ depending on who would

assess the language, within what linguistic theory, on what layer data and what form (written or oral) of the language.

Studying the data of grammatical dictionaries and grammatically annotated corpora may be regarded as a tool for measuring quantitatively expressed parameters of linguistic complexity on the level of grammar. This paper is a first step to understanding the phenomenon of linguistic complexity basing on data on distribution of part of speech homonymy in Russian; the data is retrieved from the extended version of Grammatical Dictionary of A. Zaliznyak [Zaliznyak, 1987].

We aim to uncover certain formalized and measurable parameters of linguistic complexity; the main focus is on grammatical homonymy in the Russian language. The rest of the paper is organized as follows: Section 2 presents a brief description of related works. Section 3 defines main factors influencing linguistic complexity. Section 4 gives an analysis of part of speech distribution of homonymous word forms from the viewpoint of linguistic complexity. Section 5 concludes by summarizing main results and indicating future research.

Related works

Although the concept of linguistic complexity seems intuitively clear, it has scarcely undergone formalization and analysis. Researchers regard different criteria and parameters of linguistic complexity and get different, even opposite results for the same language.

A. Berdichevsky [Berdichevsky, 2012] gives an overview of approaches to theoretical understanding of language complexity and concludes that there are three main illations confirmed by most researches. First, commonly accepted ideas about equal complexity of all languages is not true. Not only can researchers rank languages by complexity, but they also aim at measuring the complexity of a language, or, at least, of a fragment of a language, using quantitative methods. At last, such measuring, as well as certain qualitative studies, illustrate that linguistic complexity is influenced by social factors [Berdichevsky, 2012].

The dissertation of W. Kusters *Linguistic Complexity. The Influence of Social Change on Verbal Inflection* [Kusters, 2003] investigates the influence of extralinguistic factors on internal language structure. The author studies verbal inflection in certain languages (Arabic, Scandinavian, Quechua and Swahili) and argues that a large number of non-native speakers of a language, social cohesion within a speech community, and enlargement of external contacts can lead to decreasing the complexity of verbal inflection.

In [Dahl, 2004] is represented methodologically significant delimitation of a number of essential concepts: *complexity, cost, difficult*y and *demandingness.* According to this researcher, *complexity* is a theoretical construct aimed at determining "objective' parameter of a language, important for language

processing that must not be related to a user or an agent. The notions of *cost* and *difficulty* are relevant for adult language learners. Cost implies essentially "the amount of resources – in terms of energy, money or anything else – that an agent spends in order to achieve some goal" [Dahl, 2004]. High cost does not necessarily imply high degree of complexity – the relationship between these phenomena is not direct. "Difficulty is a notion that primarily applies to tasks, and is always relative to an agent: it is easy or difficult for someone" [Dahl, 2004]. Demandingness is a link between complexity and difficulty: for instance, acquiring a human language natively is certainly demanding (only human children seem to fulfil the requirements), but it does not necessarily follow that children find it difficult [Dahl, 2004].

The paper of P. Juola [Juola, 2008] discusses some definitions proposed in literature, and shows how complexity can be assessed in various frameworks. The author focuses on mathematical and psychological aspects of complexity, and attempts to validate available complexity measurements.

We may say that the topic of complexity of languages has different dimensions and nowadays attracts a great deal of interest. Researchers maintain that language complexity may be regarded and evaluated on different levels: of the language as a whole, and of its separate layers; thus parameterisation of linguistic complexity needs further research, and work results must be considered in the general theory of language.

Parameters of complexity: toward a definition

The notion of complexity is conceptualised and defined differently in different domains. To specify this notion we are to take into consideration peculiarities of the internal organisation of the system, its evolution, interaction with the external world, etc. We are to realise that the actual diversity of internal relations of a complex object is not easy to merely describe and parameterise, but also to discover in many cases. That is essential for such a multidimensional phenomenon as language.

Assessment of linguistic complexity supposes search for objectively evaluating and finding comparable criteria. To determine the absolute value of complexity many researchers ([Dahl, 2004], [Juola, 2008] and other) use a categorical apparatus of information theory, and Kolmogorov complexity may serve as an example of that. Kolmogorov complexity may be defined as a way of measuring the amount of information in a given string – as the length of the shortest possible algorithm required to describe/generate that string [Juola, 2008]. Because of practical uncomputability and nonapplicability of Kolmogorov complexity for linguistic phenomena, P. Juola applies a purely technical expedient and considers file compression method as an attempt to approximate this kind of complexity within a tractable formal framework [Juola, 2008].

J. McWhorter, assessing linguistic complexity, relies upon the assumption that an area of grammar is more complex than the same area in another grammar to the extent that it encompasses more overt

distinctions and/or rules than another grammar [McWhorter, 2001]. This assumption is deployed in the following way:

1. A phonemic inventory is more complex to the extent that it has more marked members.
2. A syntax is more complex than another to the extent that it requires processing more rules, such as asymmetries between matrix and subordinate clauses.
3. A grammar is more complex than another to the extent that it gives overt and grammaticalized expressions to more fine-grained semantic and/or pragmatic distinctions than another.
4. Inflectional morphology renders a grammar more complex than another one in most cases [McWhorter 2001].

Reduction to a common denominator of great variety of grammatical phenomena of different languages remains an insoluble problem; nevertheless the first steps in this direction may be made by means of automatic text processing.

Grammatical homonymy from the linguistic complexity view point

Linguistic literature does not present similar views on homonymy. Disputable items are the content of the concept, the principles of classification and classification schemes. The most general classification distinguishes lexical homonyms which represent the same category of parts of speech, and grammatical homonyms which are related to different parts of speech. Grammatical homonymy is an important manifestation of structural complexity, and formal and quantitative characterization of homonymous structures within and across languages can provide a complexity ranking for them in many respects.

In this paper we consider grammatical (part of speech) homonymy in Russian on the data of grammatical dictionary of A. Zaliznyak. The work is aimed at examining statistical characteristics of grammatical homonymy and at eliciting complexity parameters of this phenomenon.

Investigation of statistical properties is carried out by means of *Ontointegrator* software system developed by O. Nevzorova and V. Nevzorov [Nevzorova, 2009]. As the linguistic data source we used the extended version of A. Zaliznyak dictionary that was deployed in the system as a paradigmatic list of words. Total volume of the dictionary is 133,040 lexemes (3,162,600 word forms). Each word form is coded by two numerical characteristics that define constant and variable grammatical characteristics of the word form, the latter depending on the part of speech. Each homonym is marked by two or more sets of grammatical characteristics.

Figure 1 shows the basic screen form of *Ontointegrator* system for work with a grammatical dictionary. The *Ontointegrator* system has the Russian interface. Figure 2 presents distribution of word forms by

parts of speech in a grammatical dictionary (by the time the article was written). Figure 3 displays distribution of words by parts of speech.

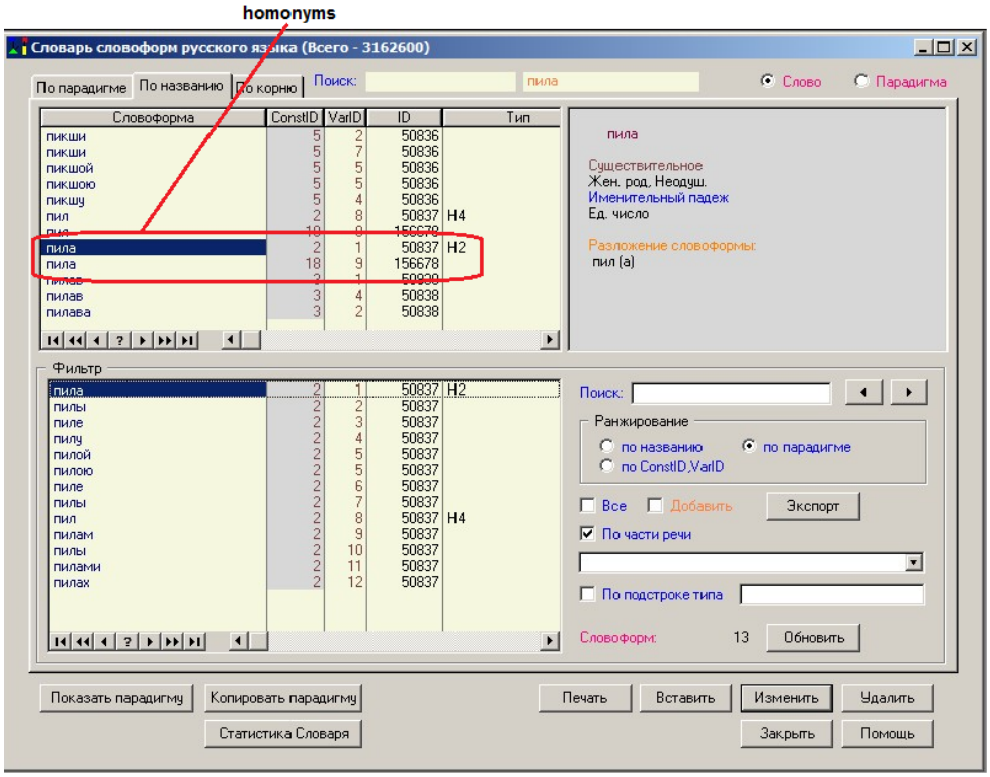

Figure 1. Basic screen form of *Ontointegrator* system for work with a grammatical dictionary

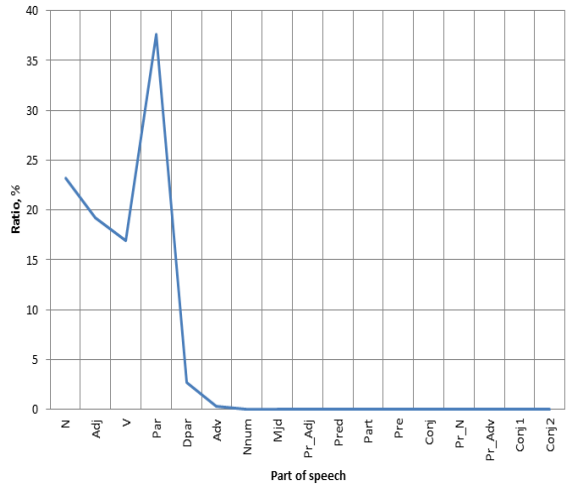

Figure 2. Distribution of word forms by parts of speech in a dictionary

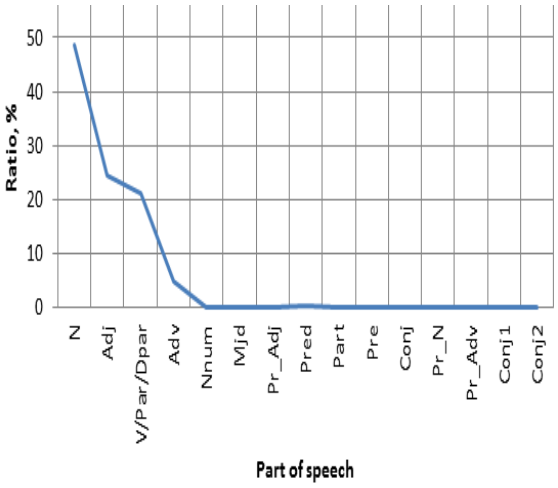

Figure 3. Distribution of words by parts of speech in a dictionary

For designating parts of speech the following abbreviations are used:

N – noun; ADJ – adjective; V – verb; PAR – participle; DPAR – the gerund; ADV – adverb; CONJ – conjunction; PRE – preposition; PART – particle; MJD – interjection; PRED – predicate word; NNUM – numeral; PR_ADJ – pronominal adjective; PR_N – pronominal noun; PR_ADV – pronominal adverb; CONJ1 – syndetic word of type 1; CONJ2 – syndetic word of type 2.

Based on grammatical characteristics of word forms we built statistical distributions by different characteristics to get a full picture of performance of grammatical homonymy in Russian.

Figure 4 displays distribution of grammatical homonyms/non-homonyms within each part of speech. Figure 4 illustrates the contribution of homonyms into each part of speech (class), i.e. into total number of elements of a given part of speech. For instance, all elements of class Conj1 (in Russian: *chego, kogo, chem, chto, kom, komy, kem, chemy*) – 8 items in all) are grammatical homonyms; class ADV contains 75,1% of grammatical homonyms, and in classes N, ADJ, V and PAR grammatical homonymy is imperceptible in relation to total number of members of these classes.

Proceeding from the analysis of distribution of homonyms within classes (parts of speech) we may identify *homonymity parameter* which may help us distinguish between strong (containing large percentage of homonyms) classes and weak (containing small percentage of homonyms) classes. 10%. is taken as a conditional threshold of division.

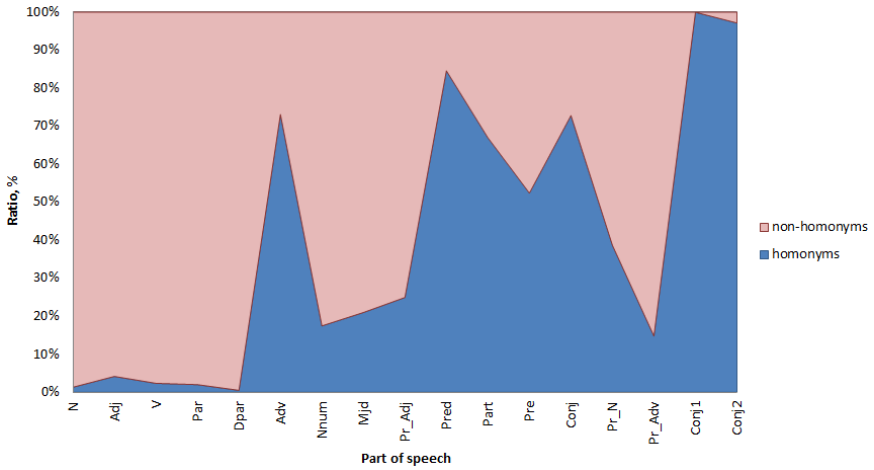

Figure 4. Distribution of grammatical homonyms/non-homonyms within each part of speech

Strong classes are ADV, NNUM, MJD, PR_ADJ, PRED, PART, PRE, CONJ, PR_N, PR_ADV, CONJ1, Conj2. Weak classes are N, ADJ, V, PAR, DPAR. Homonymity parameter reflects basic typological features of Russian morphology and syntax, where for example, homonymy of adverbs and predicate words, and of pronouns and syndetic words is a stumbling-block for disambiguation.

Figure 5 shows distribution of grammatical homonyms/non-homonyms of each part of speech (contribution to total volume of the dictionary); for visual assessment of correlation between them the amount of homonyms is displayed with the scale factor of 10. Figure 6 displays the same distribution for strong classes on an enlarged scale.

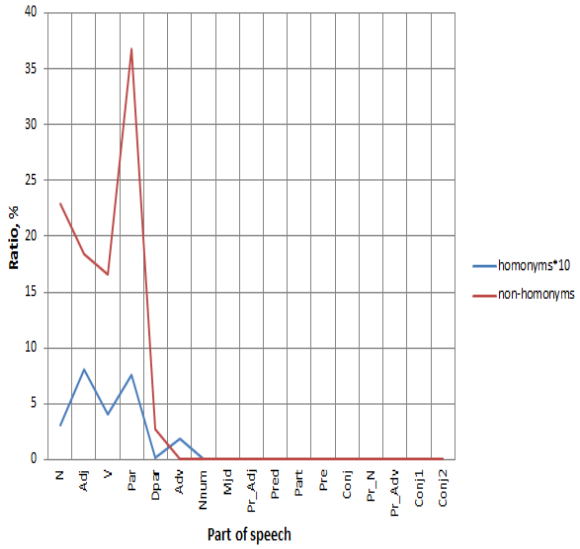

Figure 5. Distribution of grammatical homonyms/non-homonyms of each part of speech

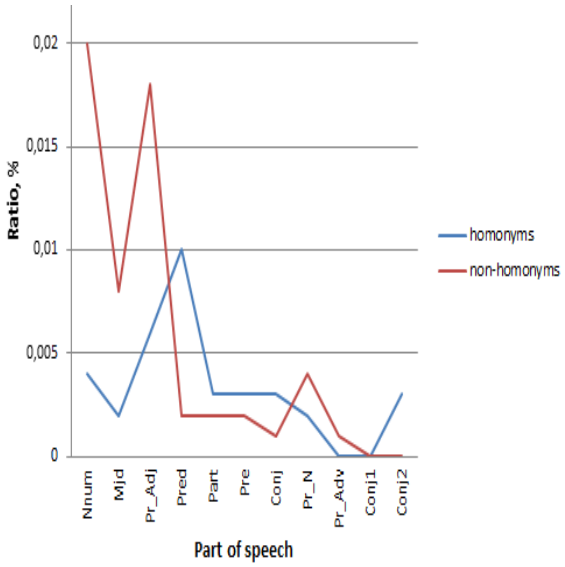

Figure 6. Distribution of grammatical homonyms/non-homonyms of each part of speech (on an enlarged scale) for strong classes

Another characteristics of the homonymy system of a language is the *power of types of grammatical homonymy*. The value of this parameter determines the number of different combinations of part of speech categories (classes) in the set of all word forms of the language. The Russian language by power of types of grammatical homonymy has value 126. The parameter of power of types of grammatical homonymy gives us the absolute numeric value of types of homonymy fixed in the dictionary.

Table 1 represents data on all types of grammatical homonymy and power of classes (estimated on belonging to the selected range of values).

The next step of the study was to examine the system of grammatical homonymy of the binary type, i.e. the object of study was grammatical homonyms with two characteristics. The first characteristic was fixed (given), and the second could vary depending on the class. In this way we built systems like Y/X, where Y and X were change within the spectrum of classes i.e. for Y=N this is N/ADJ, N/V, N/PAR, N/DPAR etc.

Table 1. Distribution of types of grammatical homonymy

Types of grammatical homonymy	Number of types	Range of values of power of the type
ADJ/PAR, PAR/V	2	5000-10000
ADJ/ADV, ADJ/N, N/V	3	1000-4999
N/PAR	1	500-999
ADJ/N/PAR, DPAR/N, ADJ/V	3	200-499
ADJ/ADV/PRED, ADV/N	2	100-199
ADJ/ADV/V, ADJ/PAR/V, ADJ/ADV/PAR, CONJ2/PR_ADJ, N/NNUM	5	50-99
ADJ/DPAR, ADJ/N/V, ADJ/PRED, ADV/PRE, CONJ/PART, MJD/N, PR_ADJ/PR_N	7	20-49
ADJ/PR_ADJ, ADV/CONJ, ADV/NNUM, ADV/PRED, ADV/V, N/PRED, N/PR_ADJ, N/PAR/V,	8	10-19
ADJ/ADV/N, ADJ/ADV/PRE, ADV/DPAR, ADV/CONJ/PART, ADV/PART, CONJ2/V, DPAR/PRE, N/PART, N/PRE, PART/V, PR_ADJ/V	11	5-9
ADJ/ADV/CONJ/PART, ADJ/ADV/N/PRED, ADJ/ADV/PART, ADJ/ADV/PART/PRED, ADJ/DPAR/N, ADJ/N/PART, ADJ/N/PR_ADJ, ADJ/N/PAR/V, ADJ/PR_N, ADV/N/PART, ADV/N/PRE, ADV/N/V, ADV/PAR, ADV/PRED/V, CONJ/PART/PR_ADV, CONJ/V, CONJ/CONJ1/PR_N, CONJ1/N/PR_ADJ, CONJ1/PR_N, CONJ1/N/PR_N, CONJ2/N/PR_ADJ, DPAR/N/V, DPAR/PAR, DPAR/V, DPAR/PR_ADJ, MJD/PART, N/NNUM/PAR, NNUM/PAR/V, NNUM/PR_N, NNUM/PR_ADJ/PR_N, NNUM/V, PAR/PR_ADJ/V, PART/PR_ADJ/PR_N, PRED/V	34	2-4
Другие типы	50	1
Total number of types	**126**	

Figure 7 shows distribution of the system of binary homonyms N/X with respect to all homonyms from N. Figure 8 represents this distribution on an enlarged scale.

To characterize the homonymy systems of binary type we entered third parameter - *binary expression* with ranking values (strong, average, weak). The binary expression parameter for the system N/X has average value (9 pairs from 17 are valuable). For comparison we provide Figure 9 and 10 displaying the system of binary homonyms of the PAR/X type. The binary expression parameter for this system is weak (6 pairs from 17 are valuable).

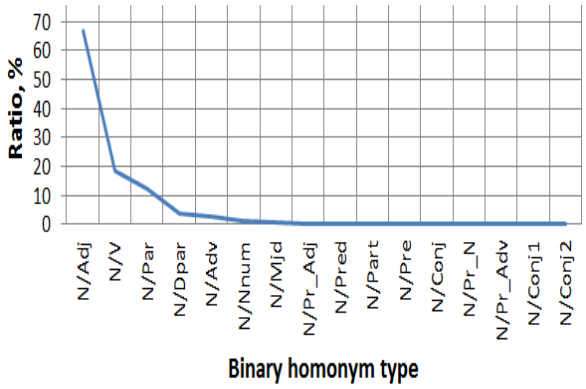

Figure 7. Distribution of binary homonyms of N/X type with respect to all homonyms from N

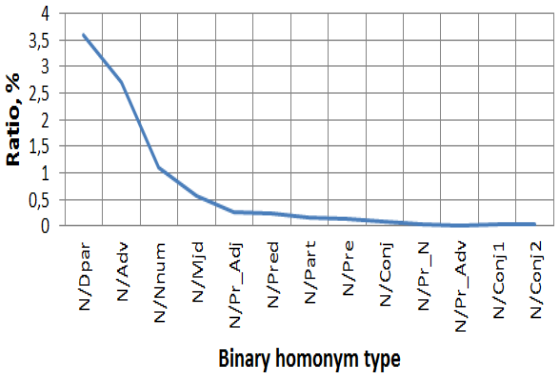

Figure 8. Distribution of binary homonyms of N/X type with respect to all homonyms from N (on an enlarged scale)

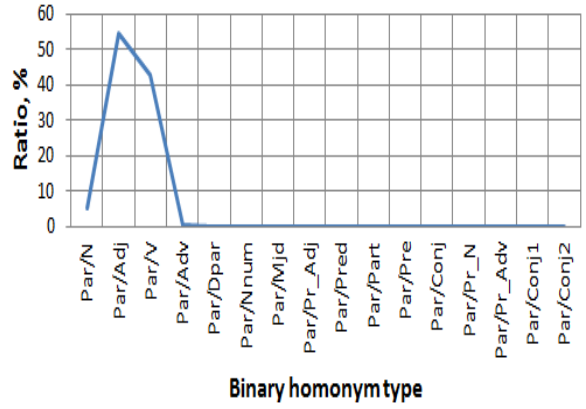

Figure 9. Distribution of binary homonyms of PAR/X type with respect to all homonyms from PAR

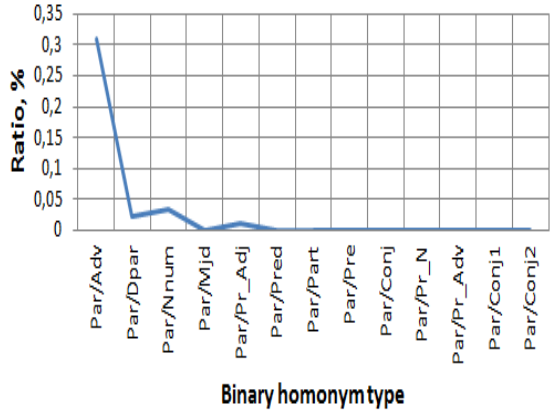

Figure 10. Distribution of binary homonyms of PAR/X type with respect to all homonyms from PAR (on an enlarged scale)

The fourth characteristics – *complication of binary parameter* for the system of binary homonyms is connected to the analysis of complication types in the structure of binary characteristics. For each binary parameter we obtained data on types of its complication, i.e. what additional classes (parts of speech) may extend the state of characteristics of the homonym. So we detected homonyms with 3 and 4 members, for example N/ADJ/V (in Russian: *zeleney, krylo*) or N/ADJ/ADV/PAR (in Russian: *gorychim*). Figure 11 presents the picture of complication for binary homonym N/ADJ. For this binary homonym complication parameter has high value (9 from 17) and in the structure of complication we find groups of homonyms of 3 members (6 groups) and 4 members (2 groups). The same plots are built for all binary parameters of all parts of speech.

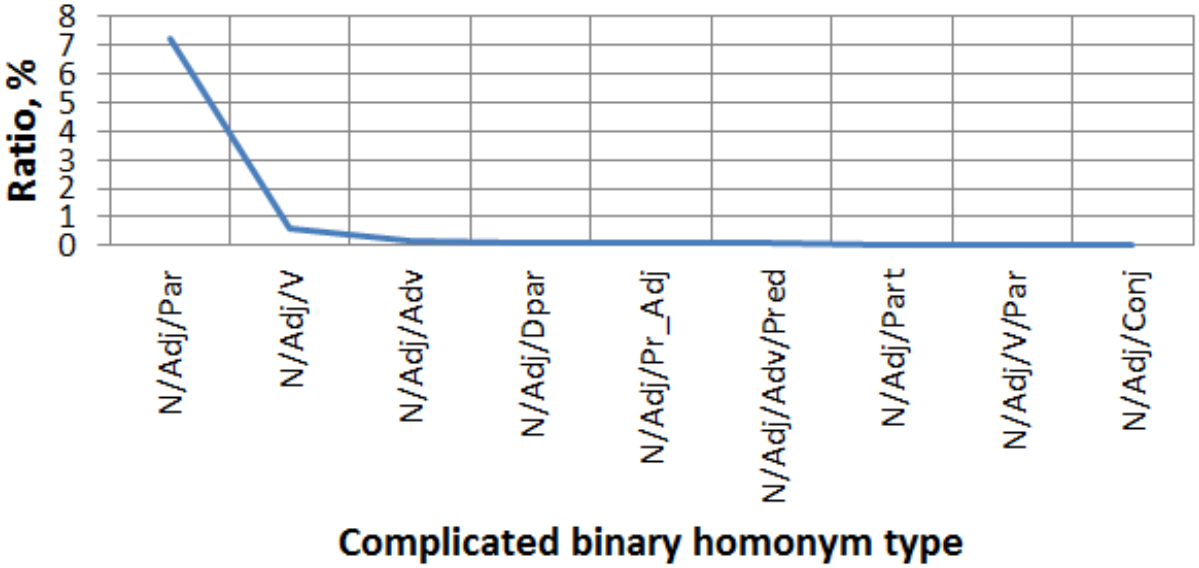

Figure 11. Distribution of complication for binary homonym of N/ADJ/X type

So for describing part of speech aspects of grammatical homonymy in Russian we distinguished four structural parameters:

- *homonymity* parameter (12 strong classes, 5 week classes);
- power of types of grammatical homonymy parameter (126 types);
- binary expression parameter (strong, average, weak values based on different binary characteristics);
- complication of binary parameter (strong, average, weak values based on different binary characteristic).

The detection of distribution of grammatical homonyms implemented by means of different parameters, manifests basic typological features of the Russian language and clarifies intricate transformations of parts of speech in the language use.

Conclusion

Formal and quantitative characterization of comparable structures within and across languages can lead to their complexity ranking. Grammatical homonymy is an important manifestation of structural complexity of a language, and many aspects of it are computable.

This paper deals with problems related to measuring complexity of natural languages on example of determining parameter of describing grammatical homonymy. Grammatical homonymy was examined on linguistic data of the extended version of A. Zaliznyak dictionary, with the help of the software of *Ontointegrator* system.

We distinguished four relevant structural parameters that enable to disclose statistical aspects of part of speech homonymy. Investigation into quantitative aspects of grammatical homonymy implemented by means of different parameters, sheds light on basic typological features of the Russian language and clarifies complicated interconnection of parts of speech in language use.

Grammatical homonymy in Russian can not be reduced merely to part of speech homonymy, so we are planning to expand research area engaging other grammatical categories. The methodology we propose may be used for comparing grammatical homonymy and related phenomena in different languages. In future we are planning to investigate grammatical homonymy in Tatar and to compare it with that of Russian.

Study of complex aspects of grammatical homonymy has important theoretical as well as practical significance, primarily for computer systems for natural language processing, machine translation and machine learning. To refine the methods of machine learning it is necessary to prepare a training collection. In the case of grammatical homonymy, the training collection can and should be built taking into account the complexity and statistical aspects of this phenomenon, relying on the structural model of grammatical homonymy.

Acknowledgement

The work is supported by the Russian Foundation for Basic Research (project # 15-07-09214).

Bibliography

[Bane, 2008] Bane, M. Quantifying and Measuring Morphological Complexity. In Proceedings of the 26th West Coast Conference on Formal Linguistics, 2008. pp. 69-76.

[Becerra-Bonache, 2015] Becerra-Bonache, L., Jimenes-Lopez, M.D. A Grammatical Inference Model for Measuring Language Complexity In Advances in Computational Intelligence. 13th International Work Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part I, 2015. pp. 4-17.

[Berdichevsky, 2012]. Berdichevsky A. Language Complexity. In Voprosy Yazykoznaniya, Vol. 5, 2012. pp. 101–124. (in Russian)

[Dahl, 2002] Dahl, Ö. The Growth and Maintenance of Linguistic Complexity. John Benjamins Publishing, Amsterdam, 2004.

[Gil, 2008] Gil, D. How Complex are Isolating Languages? In Language Complexity: Typology, Contact, Change. Ed. Miestamo, K. Sinnemäki & F. Karlsson. John Benjamins Publishing, Amsterdam. pp. 109-131.

[Juola, 2008] Juola, P. Assessing Linguistic Complexity. In Language Complexity: Typology, Contact, Change. Ed. Miestamo, K. Sinnemäki and F. Karlsson. John Benjamins Publishing, Amsterdam, 2008. pp. 89 – 108.

[Kusters, 2003] Kusters, W. Linguistic Complexity: The Influence of Social Change on Verbal Inflection. LOT, Netherlands Graduate School of Linguistics, Utrecht, 2003.

[McWhorter, 2001] McWhorter, J. The World's Simplest Grammars are Creole Grammars. In Linguistic Typology. Vol. 5, Issue, 2001: pp. 125-66.

[Miestamo, 2008] Miestamo, M., Sinnemäki, K. and Karlsson, F. (eds). Language Complexity: Typology, Contact, Change. Vol. 94. John Benjamins Publishing, Amsterdam, 2008.

[Nevzorova, 2009] Nevzorova O., Nevzorov V. The Development Support System "OntoIntegrator" for Linguistic Applications. In International Book Series "INFORMATION SCIENCE AND COMPUTING". Number 13. Intelligent Information and Engineering Systems. Supplement to the International Journal "Information Technologies & Knowledge". Vol. 3. ITHEA, Rzeszow-Sofia, 2009. pp. 78-84.

[Newmeyer, 2014] Newmeyer Frederick J. and Preston LaureL B. (ed.) Measuring Grammatical Complexity. Oxford University press, 2014.

[Rastrigin, 1981] Rastrigin L.A. Adaptation of Complex Systems. Methods and Applications. Zinatne, Riga, 1981. (in Russian).

[Zaliznyak, 1987] Zaliznyak A. A. Grammatical dictionary of the Russian Language. Inflection. Russky Yazyk, Moscow, 1987.

Authors' Information

Olga Nevzorova *– Research Institute of Applied Semiotics of Tatarstan Academy of Sciences; Deputy Director. Kazan Federal University. P.O. Box: 420111, Levobulachnaya str., 36a, Kazan, Russia; e-mail: onevzoro@gmail.com*

Major Fields of Scientific Research: Natural language processing, Artificial intelligence

Alfiya Galieva *– Research Institute of Applied Semiotics of Tatarstan Academy of Sciences; Senior researcher. P.O. Box: 420111, Levobulachnaya str., 36a, Kazan, Russia; e-mail: amgalieva@gmail.com*

Major Fields of Scientific Research: Semantics, Grammar of Turkic Languages, Philosophy of Language

Vladimir Nevzorov *– Kazan National Research Technical University named after A.N. Tupolev; Associated Professor the Department of Computer-Aided Design. P.O. Box: 420111, K. Marks str., 10, Kazan, Russia; e-mail: nevzorovvn@gmail.com*

Major Fields of Scientific Research: Natural language processing, Artificial intelligence