DOI: 10.15514/ISPRAS-2022-34(5)-10

Data Mining Methods to Compare Englishes

O.V. Donina, ORCID: 0000-0002-1053-540X <olga-donina@mail.ru>
Voronezh State University,
1, Universitetskaya square, Voronezh, 394018, Russia

Abstract. The paper presents the results of the corpus-based research of noun cryptotypes in 20 varieties of English (Englishes). The data for this research collected from Mark Davies' corpora GloWbE and NOW enabled us to focus on variation in the covert classification of nouns in modern Englishes. A noun cryptotype introduced by Whorf is approached as 'a covert type of classification of nouns, marked by lexical selection in a syntactical classifier rather than a morphological tag'. The purpose of the study has been to compare and contrast the covert classification of basic 23 emotions in 20 Englishes (64,702 tokens). 20 Englishes have been clustered with the help of Data Mining methods (such as k-means clustering and a self-organizing Kohonen map). There are six clusters that appeared to be corresponding to geographic areas: American cluster (American and Canadian Englishes); Australian cluster (Australian and New Zealand Englishes); European cluster (British and Irish Englishes); Asian cluster (Indian, Pakistani, Singapore, Hong Kong, Malaysian, Bangladeshi, Sri Lankan, and Philippine Englishes); African cluster (Kenyan, South African, Nigerian, Ghanaian, and Tanzanian Englishes); Caribbean cluster (Jamaican English). The correlation coefficients among Englishes in the Asian and African clusters (the Outer Circle in the World Englishes Paradigm of Braj B. Kachru) range from 0.74 to 0.8 due to little contact among the varieties inside these clusters. The correlation coefficients between Englishes in the American, Australian and European clusters (the Inner Circle, Kachru) range from 0.92 to 0.933, which indicates a high consistency of these varieties owing to the long lasting, enduring linguistic contacts.

Keywords: Data Mining; computer modeling; corpora studies; cryptotype analysis; Englishes

For citation: Donina O.V. Data Mining Methods to Compare Englishes. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 5, 2022. pp. 163-170. DOI: 10.15514/ISPRAS-2022-34(5)-10

Методы интеллектуального анализа данных для сравнения диалектов английского языка

О.В. Донина, ORCID: 0000-0002-1053-540X <olga-donina@mail.ru>
Воронежский государственный университет,
Россия, 394018, Воронеж, Университетская площадь, 1

Аннотация. В статье представлены результаты корпусного исследования криптотипов имен существительных в 20 диалектах английского языка (Englishes). Данные для этого исследования, собранные из корпусов GloWbE и NOW Марка Дэвиса, позволили нам сосредоточиться на вариациях скрытой классификации существительных в современных диалектах английского языка. Криптотип существительного, введенный Уорфом, рассматривается как «скрытый тип классификации существительных, отмеченный лексическим отбором в синтаксическом классификаторе, а не морфологическим тегом». Цель исследования состояла в том, чтобы сравнить и сопоставить скрытую классификацию 23 основных эмоций в двадцати диалектах английского языка (64 702 токена). 20 диалектов английского языков были сгруппированы с помощью методов интеллектуального анализа данных (таких как кластеризация k-средних и самоорганизующаяся карта Кохонена). Шесть кластеров оказались соответствующими географическим областям: американский кластер (американский и канадский английский); австралийский кластер (австралийский и новозеландский диплекты английского языка); европейский кластер (британский и ирландский английский); азиатский кластер (индийский, пакистанский, сингапурский, гонконгский, малазийский, бангладешский, шри-ланкийский

и филиппинский диалекты английского); африканский кластер (кенийский, южноафриканский, нигерийский, ганский и танзанийский диалекты английского); карибский кластер (ямайский английский). Коэффициенты корреляции среди диалектов английского в азиатском и африканском кластерах (внешний круг в парадигме Браджа Б. Качру) колеблются от 0,74 до 0,8 из-за небольшого контакта между диалектами внутри этих кластеров. Коэффициенты корреляции между диалектами в американском, австралийском и европейском кластерах (внутренний круг) колеблются от 0,92 до 0,933, что свидетельствует о высокой согласованности этих диалектов за счет длительных, устойчивых языковых контактов.

Для цитирования: Донина О.В. Методы интеллектуального анализа данных для сравнения диалектов английского языка. Труды ИСП РАН, том 34, вып. 5, 2022 г., стр. 163-170. DOI: 10.15514/ISPRAS-2022-34(5)-10

1. Introduction

Due to the universal digitalization observed in the modern information society, a lot of new scientific fields are emerging, e.g. Digital Humanities (DH) or eHumanities, which is an innovative interdisciplinary field of research that combines methods of the humanities, social and computer sciences intending to explore the possibilities of applying new digital technologies in the humanities. Qualitative data analysis in these sciences can be improved mainly through the digitized texts available for research. It is worth noting the availability and manufacturability of full-text archives (for example, various national corpora), which, instead of a small manual sample (n <100), allow to analyze a statistically representative subset (n> 1,000) or a whole corpus (n> 100,000). The computational linguistics approach allows the researcher to work with large amounts of data and at the same time pay more attention to linguistic details by modeling and visualizing the results obtained.

In sciences, where the object of research is inaccessible to direct observation, such as linguistics, it becomes necessary to model it using various means of visualizing the object under study. Moreover, due to the complexity of studying such a dynamic and multifaceted phenomenon as language, it is advisable, as research in recent years has shown, to use cognitive modeling tools. A review of modern theories of cognitive modeling in linguistics was made by L.S. Abrosimova [1]. The review discusses the success of this approach, including the applications in computational linguistics (for example, in the creation of artificial machine languages and in the improvement of automated translation). In our current work, we use computer cognitive modeling of unobservable objects, using a set of Data Mining methods, the essence of which is the process of discovering new interpretations of knowledge in raw data that are necessary for making decisions in various spheres of human activity with the help of the methods of mathematical statistics. Data Mining is a multidisciplinary field that has arisen and is developing based on such sciences as applied statistics, pattern recognition, artificial intelligence, and database theory.

2. Methodology and related work

Our research methodology of linguistic categories in different linguistic environments is based on three components: cryptotype analysis, methods of corpus linguistics, and Data Mining. We have analyzed noun cryptotypes, i.e. the language categories hidden in the English language. These hidden classes of nouns were described in the works of O.O. Boriskina [2, 3, 4, 5, 6, 7].

A noun cryptotype introduced by Whorf [15] is approached as 'a covert type of classification of nouns, marked by lexical selection in a syntactical classifier rather than a morphological tag' [16]. The purpose of the study has been to compare and contrast the covert classification of basic 23 emotions in 20 Englishes (64,702 tokens).

Noun Cryptotypes represent word classes organized on two main principles: the modeling power of the Cryptotype Core nouns and the inherent potentialities of the Cryptotype Periphery nouns to imitate the Core nouns syntactic behavior, i.e. to borrow the cryptotype Core nouns classifiers and thus to adapt to the Core nouns combinatory characteristics. To illustrate, in context An authentic

feeling was able to penetrate the structure of the debate, the S-position of the verb to penetrate can be substituted by a name of a sharp-pointed object. The noun feeling substitutes the above-mentioned position. This metaphor is meant as the discourse evidence of the noun question belonging to the periphery of cryptotype 'Sharp objects'.

At the moment, 6 cryptotypes of the English language have been identified and described (they correspond with explicit lexical and grammatical categories of some other world languages according to a typological research): the cryptotype Res Liquidae (a class of liquids, the prototype is 'water'), Res Acutae (a class of sharp objects, the prototype is 'thorn'), Res Filiformes (a class of thin objects of unstable form, the prototype is 'thread'), Res Rotundae (a class of round objects, the prototype is 'ball'), Res Parvae (a class of hand-fitting objects, the prototype is 'apple'), Res Longae Penetrantes (a class of solid, long, pointed objects, the prototype is 'stick') [8, 11, 12]. For example, the nominal cryptotype of the English language Res Liquidae includes such nouns as water, blood, milk, and other nominations of objects of reality that exist in a liquid state. These nouns are the prototypes of the cryptotype. At the same time, this cryptotype also contains nouns denoting abstract concepts. Concepts such as life, goodness, or passion do not occur in a liquid state, but a person often categorizes them by analogy with a liquid one. Thus, the names of abstract semantics in metaphorical use can be included in the nominative cryptotypes distinguished in the English language, so that the names of concrete and abstract semantics coexist in the same language class.

Cryptotype analysis was carried out on the names of emotional state and sensory experience (such as anger, fear, love, etc.) in 20 varieties of the English language (Englishes) presented in the Mark Davis' corpus [9]. Along with British and American, it also features rare varieties (for example, Kenyan or Tanzanian Englishes). The volume of the research corpus, formed based on the results of semi-automatic processing of corpus queries, amounted to 65,000 tokens. Previously, to visualize the results of a cryptotype study, Chernov's faces were used [10]. But in our research, given the rather large volume of the resulting research corpus and the need to bring together parameters of different quality (namely, 20 variants of the English language under consideration, 23 names of emotions, and 6 nominal cryptotypes), the use of this method turned out to be impossible. That is why we decided to try to apply the methods of Data Mining and computer-cognitive modeling to achieve our goal.

3. The used approach

First of all, to determine the significance of the factors and the possible reduction of the input parameters before clustering using the Deductor Academic 5.3 program, a factor analysis based on the "varimax" method was carried out. Six currently allocated cryptotypes were used as input data.

The next step was the k-means clustering, as well as the construction of a self-organizing Kohonen map, which is a type of neural network algorithms. Clustering is used to distribute a set of objects into classes that are not initially specified, with the k-means algorithm being the most commonly used. Artificial neural networks, which arose as a model of the biological nervous system, consist of input, hidden and output layers of neurons, and for the input and output layers the parameters are known, while implicit signal transformations take place in the hidden one. In linguistics, neural networks are used in neural network models of a language, in machine translation, automatic clustering of vocabulary (Kohonen maps), etc.

Having employed factor analysis, it was shown that as a result of the performed rotation, there were no significant changes in the structure of the factor space (i.e., the factors set automatically correspond to six cryptotypes by 96.09% - 99.51% (Table 1)). Thus, we can talk about the stability of the data, which indicates the independence of the factors reflected in the correlation matrix, i.e. the high explanatory significance of all factors (cryptotypes) and the possibility of using them for further clustering were proved.

Table 1. Results of factor analysis

	Final factors (Varimax method)					
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Res Acutae			0,9951			
Res Filiformes		0,9841				
Res Liquidae						0,9609
Res Longae Penetrantes					0,9769	
Res Parvae				0,9851		
Res Rotundae	0,9849					

Using the Kohonen neural network, the input maps of the neurons of six cryptotypes were generated, i.e. the internal structure of the input data was visualized by adjusting the weights of the neurons of the maps, where the areas containing approximately the same inputs for the analyzed examples are marked with a certain color. As a result of vector quantization, individual varieties of the English language were grouped into six clusters corresponding to geographic areas (Fig. 1): 1. American area (American English and Canadian English), 2. Australian area (Australian English and New Zealand English), 3. European area (British English and Irish English), 4. Asian area (Indian English, Pakistani English, Singapore English, Hong Kong English, Malaysian English, Bangladeshi English, Sri Lankan English, and Philippine English), 5. African area (Kenyan English, South African English, Nigerian English, Ghanaian English, and Tanzanian English), 6. Caribbean area (Jamaican English).

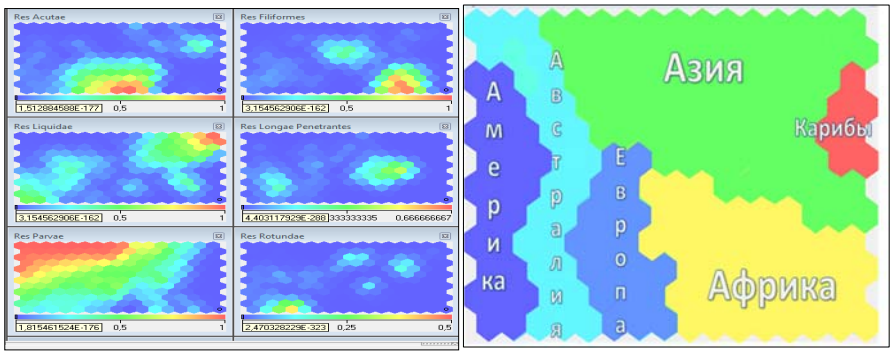

Fig. 1. Kohonen self-organizing maps

The coincidence of the cryptotype categorization with the geographical one emphasizes the importance of the areal influence noted in the work of V.N. Polyakova and E.I. Yaroslavtseva [13]. Within the framework of this work (on the material of the database "Languages of the World"), the phenomenon of a typological shift is studied, the essence of which is that languages in the process of areal contacts acquire new typological features and lose some of the existing ones. At the same time, widespread signs tend to further spread, and low-frequency ones tend to wash out [13: 114-115].

According to the cryptotype analysis for all 23 names of emotions in 20 Englishes, statistics on cryptotypes were calculated, which reflected the general trends characteristic of the research corpus as a whole (Fig. 2). The data obtained showed that in 92.3% of cases the share of the cryptotype Res Rotundae does not exceed 5.6%, which is the lowest value among the cryptotypes, i.e. this class is the least represented in our research corpus. In 93.6% of examples, the representation of the cryptotype Res Longae Penetrantes is in the range from 0% to 11.1%, i.e. it appears next to last in terms of representation. The fourth most common cryptotype is Res Filiformes: in 96.1% of cases, the share of this cryptotype varies from 0% to 22.2%. The cryptotype Res Acutae is the third in frequency: in 79.3% of cases its share of representation does not rise above 22.2%, but at the same time in 3% of examples it is one of the most represented with a value of the cryptotype activity of some emotion names from 88, 9% to 100%. The second place is taken by the cryptotype Res Liquidae, whose share of representation in 94.2% of cases ranges from 0% to 55.6%. The leader in the prevailing majority of word usage is the cryptotype Res Parvae, which share of representation in 55.2% of cases exceeds 55.6%.

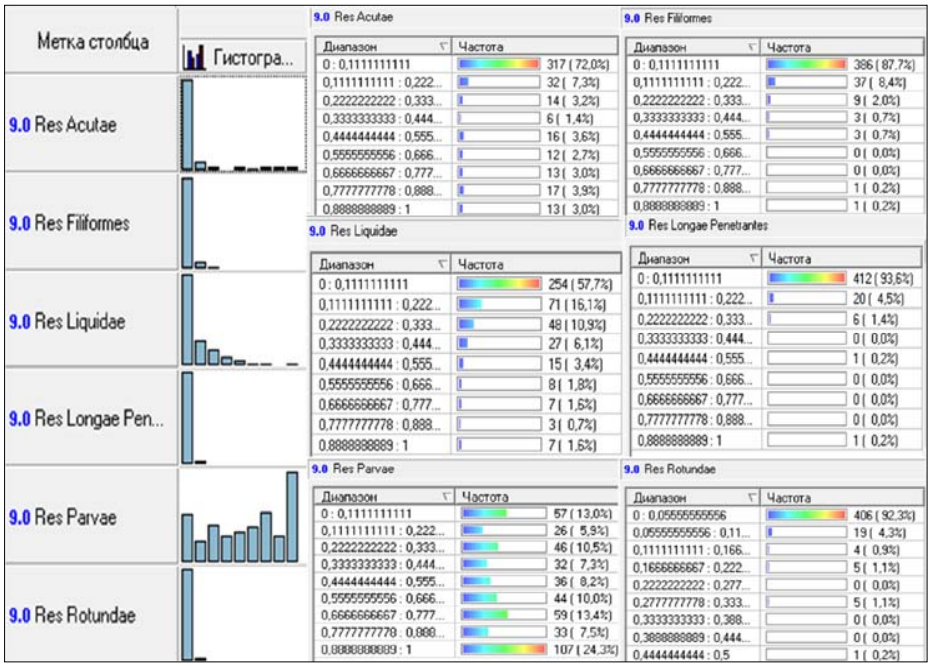

Fig. 2. Frequency of cryptotype representation in the research building

The next step in our research was to create rules based on a decision tree that would allow us to write a computer program capable of establishing the areal affiliation of a language dialect. Decision trees are a method of automatic data analysis that forms a sequential structure of rules, where each object corresponds to a single node that gives a solution. The results of such an analysis can be presented both in the form of a hierarchy (Fig. 3) and in the form of a set of rules describing classes. In the future, thanks to the compiled rules and descriptions of clusters, it becomes possible to trace the dynamics of the influence of language varieties/language areas on each other, conducting a similar study in 10-15 years.

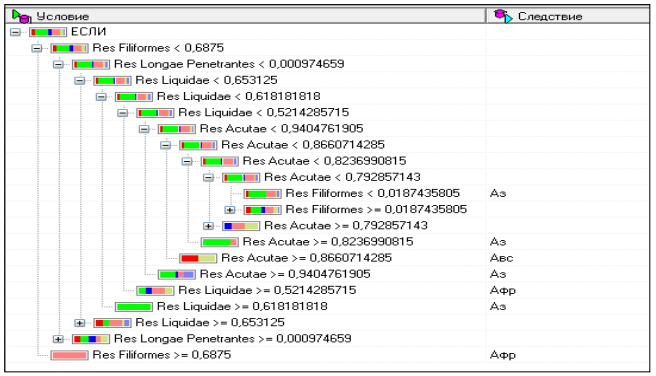

Fig. 3. Fragment of a decision tree

2. Results and conclusion

Data Mining Methods helped to obtain the following main results:

1) All emotions are attributed to one of the six noun cryptotypes in all Englishes:
   - Res Parvae – a class of hand-fitting objects (prototype: stone),
   - Res Liquidae – a class of liquids (water),
   - Res Filiformes – a class of thin objects of unstable form (thread),
   - Res Rotundae – a class of round objects (ball),
   - Res Longae Penetrantes – a class of solid, long, pointed objects (spear),
   - Res Acutae – a class of sharp objects (thorn).
1) 20 Englishes have been clustered with the help of Data Mining methods (such as k-means clustering and a self-organizing Kohonen map). There are six clusters that appeared to be corresponding to geographic areas: American cluster (American and Canadian Englishes); Australian cluster (Australian and New Zealand Englishes); European cluster (British and Irish Englishes); Asian cluster (Indian, Pakistani, Singapore, Hong Kong, Malaysian, Bangladeshi, Sri Lankan, and Philippine Englishes); African cluster (Kenyan, South African, Nigerian, Ghanaian, and Tanzanian Englishes); Caribbean cluster (Jamaican English).
2) The correlation coefficients among Englishes in the Asian and African clusters (the Outer Circle in the World Englishes Paradigm of Braj B. Kachru [14]) range from 0.74 to 0.8 due to little contact among the varieties inside these clusters.
3) The correlation coefficients between Englishes in the American, Australian and European clusters (the Inner Circle, Kachru [14]) range from 0.92 to 0.933, which indicates a high consistency of these varieties owing to the long lasting, enduring linguistic contacts.
4) The most relevant to the metaphorical categorization of emotions in all Englishes is Res Parvae. Notably, there is a grammatical category of hand-fitting objects in the grammar systems of some indigenous languages of African, American and Australian clusters, and there is a classifier for counting hand-fitting objects in some Asian languages.
5) Res Liquidae is the second cryptotype frequently associated with emotions for English-speakers of Australian, American, African and Caribbean clusters, whereas for Asian and European clusters it is the class of sharp objects. Presumably, it could be due to close contacts of the English language with the languages of the indigenous population.

References / Список литературы

[1] Abrosimova L.S. Word formation in the linguistic categorization of the world. Rostov-on-Don, SFU Publishing House, 2015, 328 p. (in Russian) / Абросимова Л.С. Словообразование в языковой категоризации мира. Ростов на Дону, Изд-во ЮФУ, 2015 г., 328 стр.

[2] Boriskina O.O Linguistic categorization of the elements. In Proc. of the 2nd International Conference on Philology and Culture, 2019, pp. 149-157 (in Russian) / Борискина О.О. Языковая категоризация стихий. Материалы 2-й международной конференции «Филология и культура», 1999 г., стр. 149-156.

[3] Boriskina O.O. Cryptoclasses of primary elements as an element of the ontognostic description of language. In Problems of linguistic prognostics, issue 1, Voronezh, Central Black Earth Book Publishing House, 2000, pp. 121-126 (in Russian) / Борискина О.О. Криптоклассы первостихий как элемент онтогностического описания языка. В сборнике статей «Проблемы Лингвистической Прогностики», вып. 1, Воронеж, Центрально-Черноземное книжное издательство, 2000 г., стр. 121-126.

[4] Boriskina O.O. National-specific linguistic consciousness and borrowed word. In Intercultural communication and problems of national identity, Voronezh, Voronezh State University Press, 2002, pp. 406-410 (in Russian) / Борискина О.О. Национально-специфическое языковое сознание и заимствованное слово. В сборнике статей «Межкультурная коммуникация и проблемы национальной идентичности», Воронеж, издательство Воронежского государственного университета, 2002 г., стр. 406-410.

[5] Boriskina O.O. The dynamic noun combinatory profile. Issues of cognitive linguistics, 2008, issue 3, pp. 64-69 (in Russian) / Борискина О.О. Моделирование синтагматической динамики слова. Вопросы когнитивной лингвистики, вып. 3, 2008 г., стр. 64-69. .

[6] Boriskina O.O. Cryptotype projection of abstract entities: applications of cryptotype approach to noun combinations study. Proceedings of Voronezh State University. Series: Linguistics and intercultural communication, issue 1, 2009, pp. 32-37 (in Russian) / Boriskina O.O. Криптоклассные проекции мира непредметных сущностей: опыт криптоклассного анализа словосочетаемости. Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация, вып. 1, 2009 г., стр. 32-37.

[7] Boriskina O.O. Explanation of the unexplained or about the motivation of the unmotivated. Vestnik of Saint Petersburg University. Series 9. Philology. Oriental studies. Journalism, issue 1, 2010, pp. 95-100 (in Russian) / Борискина О.О. Объяснение необъяснимого или мотивация немотивированного. Вестник Санкт-Петербургского университета. Серия 9. Филология. Востоковедение. Журналистика, вып. 1, 2010 г., стр. 95-100.

[8] Boriskina O.O., Marchenko T. An Algorithm for Analysis of Distribution of Abstract Nouns in Cryptotypes. In Proc. of the 2010 International Conference on Artificial Intelligence, 2010, pp. 907-913.

[9] NOW Corpus (News on the Web). Available at: https://www.english-corpora.org/now/.

[10] Donina O.V. The study of cryptopytes: vizualization of reserach results. Proceedings of Voronezh State University. Series: Linguistics and intercultural communication, issue 3, 2015, pp. 105-112 (in Russian) / Донина О.В. Способы визуализации результатов криптоклассного исследования. Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация, вып. 3, 2015 г., стр. 105-112.

[11] Donina O.V., Boriskina O.O. Emotive lexemes from the perspective of areal variativity. Proceedings of Voronezh State University. Series: Linguistics and intercultural communication, issue 4, 2016, pp. 41-45 (in Russian) / Донина О.В., Борискина О.О. Эмотивная лексика в аспекте ареальной вариативности. Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация, вып. 4, 2016 г., стр. 41-45.

[12] Kretov A.A., Boriskina O.O, Vasilyeva N. 2004. «Flight of thought» and methods of cryptoclass investigation. Proceedings of Voronezh State University. Series: Linguistics and intercultural communication, issue. 1, 2004, pp. 61-65 (in Russian) / Кретов А.А., Борискина О.О., Васильева Н.Е. «Полёт мысли» и методика исследования криптоклассов. Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация, вып. 1, 2004 г., стр. 61-65.

[13] Polyakov V.N., Yaroslavtseva E.I. The Quantitative Parameters of Typological Shift. Scientific notes of Kazan State University. Series: Humanitarian sciences, vol. 150, issue 2, 2008, pp. 97-118 (in Russian) / В.Н. Поляков, Е.И. Ярославцева. Квантитативные закономерности типологического сдвига в языках Евразии (на материале БД «Языки мира» ИЯ РАН). Ученые записки Казанского университета. Серия Гуманитарные науки, том 150, вып. 2, 2008 г., стр. 97-118.

[14] Kachru B. Models for Non–native Englishes. The Other Tongue: English across cultures. Urbana: University of Illinois Press, 1992, 416 p.

[15] Whorf B.L. Language, Thought and Reality. Selected Writings of Benjamin Lee. The MIT Press, 1964, 290 p.

[16] Boriskina O.O. The Main Criteria for the Exploration of Noun Cryptotypes. In Proc. of the VI International Scientific Conference on Language, Culture, Society, 2011. Available at: http://www.mosinyaz.com/conferences/mnk6_s3_12/.

Information about the author / Информация об авторе

Olga Valer'evna DONINA – Candidate of Philological Sciences, Associate Professor of the Department of Theoretical and Applied Linguistics. Research interests: computational linguistics, corpus linguistics, cognitive linguistics, metaphor study, the study of Englishes.

Ольга Валерьевна ДОНИНА – кандидат филологических наук, доцент кафедры теоретического и прикладного языкознания. Научные интересы: компьютерная лингвистика, корпусная лингвистика, когнитивная лингвистика, метафороведение, изучение диалектов английского языка.