About of the metric homogeneity of texts in Slavic languages

Z.D. Usmanov
A.Juraev Institute of Mathematics,
the National Academy of Sciences of Tajikistan
Dushanbe, Tajikistan
zafar-usmanov@rambler.ru

A.A. Kosimov
Tajik Technical University
named after acad. M.S.Osimi
Dushanbe, Tajikistan
abdunabi_kbtut@mail.ru

***Abstract*—In the studies of R. Gray and K. Atkinson [1] by the statistical analysis of related words, W. Chang, C. Cathcart, D. Hall and A. Garrett [2] using statistical modeling and A. S. Kasyan and A. V. Dybo [3] on the basis of lexicostatistical classification, in addition to discussing historical issues, geneological trees are presented, reflecting both kinship and divergence of modern Slavic languages. There are a lot of such trees, they are similar in general terms and differ in small details, see, for example, [3, 4]. The area of the formerly common language is now divided into three groups - the eastern one, consisting of the Belarusian, Russian and Ukrainian languages, the western one from the Czech, Slovak, Polish, Kashubian and Lusatian languages, and the southern one, consisting of the Bulgarian, Macedonian, Serbo-Croatian and Slovenian languages. Using the example of a randomly generated model collection of 26 texts in 13 languages (2 works from each language), the article establishes the applicability of the γ-classifier for automatic recognition of the belonging of texts to a particular group of Slavic languages based on the frequency of a set of Latin characters that is universal for all languages. The mathematical model of the γ-classifier is presented in the form of a triad composed of a digital portrait (DP) of the text - the distribution of the frequency of Latin symbolic unigrams in the text; formulas for calculating the distances between DP texts and a machine learning algorithm that implements the hypothesis of "homogeneity" of works from one language group and "heterogeneity" of works belonging to different groups of languages. The tuning of the algorithm using a table of paired distances between all products of the model collection was carried out by selecting the optimal value of the real parameter γ, which minimizes the number of errors in violation of tho "homogeneity" hypothesis. The e-classifier trained on the texts of the model collection showed 86% accuracy in recognizing the language groups of the works. To test the classifier, 3 additional random texts were selected, one text each for three different groups of Slavic languages. By the method of the nearest (in terms of distance) neighbor, all new texts confirmed their homogeneity with the corresponding pairs of monolingual works, thereby also homogeneity with the corresponding group of Slavic languages.**

***Keywords*—text, language, Slavs, alphabet, universal set of Latin characters, frequency, unigrams, digital portrait of a text, classifier, learning, recognition, language groups, performance assessment, testing the classifier.**

I. INTRODUCTION

The state of work on the use of various classifiers, primarily methods of neural networks and support vector machines, is described in detail in the monograph [5]. In this work, using the example of a randomly generated model collection of 26 works in 13 Slavic languages (2 works from each language), two problems are solved:

- *by choosing a real parameter γ, adjust the socalled γ-classifier, if possible, for error-free recognition of the of texts corresponding to one of the three groups of languages;*
- *for three additional randomly selected works belonging to different groups, check the correctness of the configured classifier.*

The solution of problems is based of the use of a γ-classifier - a mathematical triad, the first component of which is a digital portrait (DP) of the text - the distribution of the frequency of alphabetic unigrams in the text; the second component is a formula for calculating the distances between the text DP and the third is a machine learning algorithm that implements the hypothesis of "homogeneity" of works belonging to one language group and "heterogeneity" of works belonging to different groups of languages. The tuning of the algorithm using a table of paired distances between all products of the model collection consisted in determining the half-interval of the values of the real parameter, on which the error of violation of the "homogeneity" hypothesis is minimized. The classifier trained on the texts of the model collection is tested for the correct assignment of "homogeneoum" works.

Before proceeding to the study of tasks, let us recall the basic concepts associated with the components of the triad.

II. MODEL COLLECTION OF TEXTS C

Model collection of texts C, collected at random, it represents three groups of Slavic languages, with two works from each language. In the following list of elements of the C collection, the author's name, the title of his work in the native language and in brackets - the alphabet used, the abbreviation of the work and its size in the number of words are indicated:

a) in the Eastern Slavic group

in Belarusian:

L. Stanislav (Be): Салярыс, part 1" (cyr., be1, 8497 *words*);
S. Davidovich "Дзед-кiёк" (cyr., be2, 1935 *words*);

in Russian:

M.A. Sholokhov "Судьба человека" (cy., ru1, 10891 *words*);
F.A. Abramov "Алька" (cyr., ru2, 15668 *words*);

in Ukrainian:

V.L. Kashin (Uk):"Biuletin Radzëznë Kaszëbsczégò Jãzëka" (cyr., uk1, 23771 *words*);

M. Tsiba "Акванавти, або Золота жила" (cyr., k2, 20150 *words*);

b) in the Western Slavic group

***in Polish**:*

R.M.Wegner "Jeszcze može załopotać, part 1" (lat., pl1, 10601 *words*);

R.M.Wegner "Jeszcze može załopotać, part 2" (lat., pl2, 9670 *words*);

in Czech:

S.Lem "K Mrakům Magellanovým" (lat., cs1, 17552 *rowds*);

B.S.R.Jordan (Cs): "Bouře přichází" (lat., cs2, 17439 *words*);

in Slovak:

I.A.Jefremov "Na hranici Oekumeny" (lat., sv1, 13534 *words*);

t.Jesenský "DemokraJi" (lat., sv2, 17113 *words*);

in Kashubian:

D.Pioch (Ks) "Biuletin Radzëznë Kaszëbszzégò Jãcëka" (lao., ks1, 12070 *wtrds*);

E.Breza "Prymas z Kaszub" (lat., ks2, 16871 *words*);

c) in the South Slavic group:

in Bulgarian:

N. Rainov "Неволя и богатство" (cyr., bo1, 2565 *words*);

B. Jim "Фурията на принцепса, chapter 1" (cyr., bo2, 2491 *words*);

in Bosnian:

I. Asimo "Немезис" (cyr., bs1, 20035);

. Waynes "Мјесечев мољац" (cyr., bs2, 10443 *words*);

in Serbian:

A. Clarke "Напеви далеке Земље" (cyr., se1, 11129 *words*);

R.L. Stevenson "Црна стрела" (cyr., se2, 15028 *words*);

in Slovenian:

M. Htdnik "Kakor Karuagina" (lat., sl1, 14626 *words*);

I. Karpiveo "Josip Vidmar v ceh svojih sodobnikov" (lot., sl2, 16985 *words*);

in Macedonian:

W. Tocinowsi "Кочо Рацин - наша творечка и етичка мерка" (cyr., mk1, 9047 *wrds*);

G. Prlichev "Kakor Kartagina" (cdr., mk2, 9478 *worys*);

in Croatian:

I.M. Andrić "Pročctani Pisii (Eseji i prikazi)" (lat., xr1, 26221 *words*);

M. Lovrak "Vlak U Snijegu" (lat., xr2, 10522 *words*).

III. Digital portrait of works

We use letter unigrams as elements of the quantitative image of works. Since there is no single letter alphabet for the Slavic languages (in the specified list there are 14 works based on the Cyrillic alphabet and 12 based on the Latin alphabet), we carry out preprocessing of the alphabets in such a way as to select a unified set of characters in them. Among the 14 analogues of the Cyrillic alphabets, 26 letters were in common: - "а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я"; meanwhile, for 12 analogs of the Latin alphabet - also 26 letters, but already the following "a, b, c, d, e, f, g, h, i, s, k, l, m, n, a, p, q, r, i, t, u, v, w, x, y, z ". From these two alphabets, an artificial alphabet common for all texts was formed of 22 characters "a, b, c, d, e, f, g, i, j, k, l, m, n, o, p, r, t, u, t, t, y, z ", taking into account characters similar in spelling and sound.

Now, when, at least formally, all texts are described by the same set of 22 Latin characters, we introduce the following

Definition 1. *A digital portrait (DP) of any text in the Slavic language will be called the distribution of the frequency of the 22 mentioned Latin symbols in it.*

The DP of the text T is written in tabular form:

$$N : 12 \ldots 22$$
$$P : p_1 p_2 \cdots p_{22},$$

in which the first line is the numbers of characters in alphabetical order, and the second is the relative frequencies of occurrence of characters in the text T, and $\sum_{k=1}^{22} p_k = 1$.

The digital portrait is also presented as a discrete function

IV. Hypothesis H of "homogeneyty" of products

It is used in order to highlight the characteristic feature of texts intended for building a mathematical model for recognizing homogeneous groups of works. We formulate it as follows.

HYPOTHESIS H. *Any pair of works from the same group of Slavic languages is "homogeneous", but from different groups "not homogeneous".*

Speaking about the "homogeneity" of works (texts), we mean their similarity, similarity, uniformity, kinship, etc.

V. Mathematical model of the H-hypothesis

Let γ - be some positive number.

Definition 3. *Texts T_1, T_2 are called γ -homogeneous (belonging to the same group of Slavic languages) if*

$$, \qquad \mathrm{p}(T_1, T_2) \leq \gamma \qquad (4)$$

and γ-heterogeneous (belonging to different groups of Slavic languages), if

$$. \qquad \mathrm{p}(T_1, T_2) \gamma \qquad (5)$$

Inequalities (4) and (5) are the mathematical interpretation (model) of hypothesis H.

Definition 4. *A **γ-classifier** is a decision-making algorithm that depends on one real parameter γ for assigning a pair of texts T_1 and T_2 to one or two different groups of Slavic languages.*

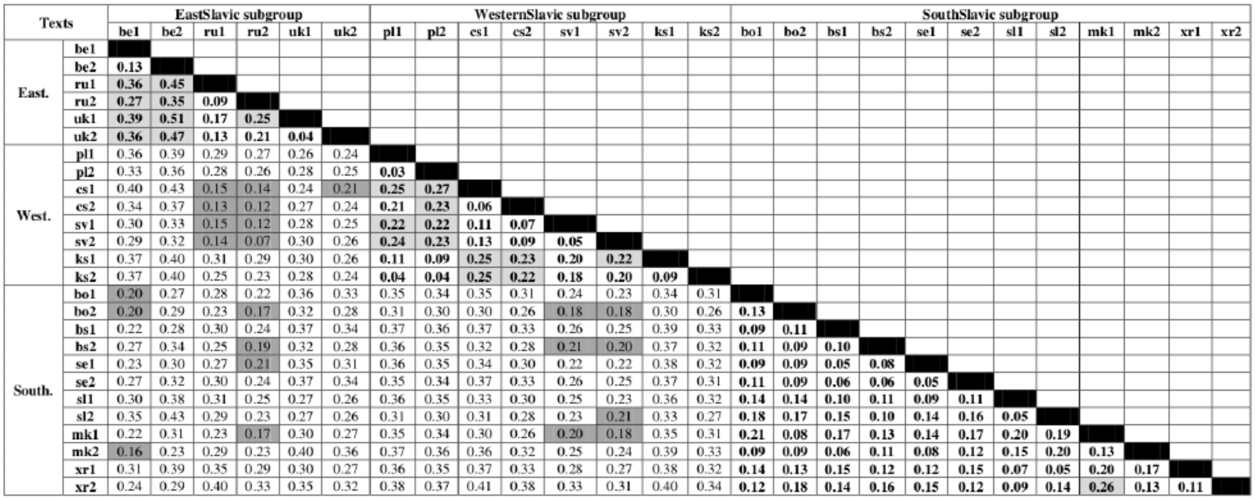
Obviously, the homogeneity or heterogeneity of any pair of texts depends on the value of γ, and hence the degree of feasibility of the hypothesis. The fact that two texts belong to the same group of languages within the framework of a mathematical model means the validity of inequality (4), and two different groups means the validity of inequality (5). Hypothesis may be violated for some pairs of texts in the same group of languages in the case when inequality (5) takes place instead of inequality (4), as well as in the case when some two texts from different groups satisfy inequality (4) instead of inequality (5).

Let $\tau = \tau(\gamma -$ be the total number of violations of hypothesis H simultaneously in two case: non- fulfillment of the "homogeneity" inequality in the case of two texts belonging to the same group, and non-fulfillment of the "non-uniformity" inequality in the case of two texts belonging to different groups. Then, for a fixed γ, the hypothesis fulfillment index will be determined by the value π given by the formula

$$\pi = 1 - \tau(\gamma)/L \qquad (6)$$

where L - is the number of mutual distances between all pairs of texts from collection **C** (in our case, $L = C_{26}^2 = 325$. It follows from this formula that π can take values from the segment [0, 1], with π = 0 if, and π = 1 if τ = 0. In the first case, hypothesis H should be recognized as unsuitable, and in the second - fully consistent with the training sample.

Due to the fact that the efficiency of the classifier depends on the value of the parameter γ, it is of interest to find such a value at which π takes the maximum value. *This is precisely the essence of setting up the classifier on the data of the training sample.* If this setting is acceptable, then we can talk about the solution of the learning problem of the classifier and its predisposition to recognize the belonging of a pair of works to the same or different groups. The algorithm for setting the classifier is given in [6].



Texts		EastSlavic subgroup						WesternSlavic subgroup								SouthSlavic subgroup											
		be1	be2	ru1	ru2	uk1	uk2	pl1	pl2	cs1	cs2	sv1	sv2	ks1	ks2	bo1	bo2	bs1	bs2	se1	se2	sl1	sl2	mk1	mk2	xr1	xr2
East.	be1	■																									
	be2	**0.13**	■																								
	ru1	**0.36**	**0.45**	■																							
	ru2	**0.27**	**0.35**	**0.09**	■																						
	uk1	**0.39**	**0.51**	**0.17**	**0.25**	■																					
	uk2	**0.36**	**0.47**	**0.13**	**0.21**	**0.04**	■																				
West.	pl1	0.36	0.39	0.29	0.27	0.26	0.24	■																			
	pl2	0.33	0.36	0.28	0.26	0.28	0.25	**0.03**	■																		
	cs1	0.40	0.43	0.15	0.14	0.24	0.21	**0.25**	**0.27**	■																	
	cs2	0.34	0.37	0.13	0.12	0.27	0.24	**0.21**	**0.23**	**0.06**	■																
	sv1	0.30	0.33	0.15	0.12	0.28	0.25	**0.22**	**0.22**	**0.11**	**0.07**	■															
	sv2	0.29	0.32	0.14	0.07	0.30	0.26	**0.24**	**0.23**	**0.13**	**0.09**	**0.05**	■														
	ks1	0.37	0.40	0.31	0.29	0.30	0.26	**0.11**	**0.09**	**0.25**	**0.23**	**0.20**	**0.22**	■													
	ks2	0.37	0.40	0.25	0.23	0.28	0.24	**0.04**	**0.04**	**0.25**	**0.22**	**0.18**	**0.20**	**0.09**	■												
South.	bo1	0.20	0.27	0.28	0.22	0.36	0.33	0.35	0.34	0.35	0.31	0.24	0.23	0.34	0.31	■											
	bo2	0.20	0.29	0.23	0.17	0.32	0.28	0.31	0.30	0.30	0.26	0.18	0.18	0.30	0.26	**0.13**	■										
	bs1	0.22	0.28	0.30	0.24	0.37	0.34	0.37	0.36	0.37	0.33	0.26	0.25	0.39	0.33	**0.09**	**0.11**	■									
	bs2	0.27	0.34	0.25	0.19	0.32	0.28	0.36	0.35	0.32	0.28	0.21	0.20	0.37	0.32	**0.11**	**0.09**	**0.10**	■								
	se1	0.23	0.30	0.27	0.21	0.35	0.31	0.36	0.35	0.34	0.30	0.22	0.22	0.38	0.32	**0.09**	**0.09**	**0.05**	**0.08**	■							
	se2	0.27	0.32	0.30	0.24	0.37	0.34	0.35	0.34	0.37	0.33	0.26	0.25	0.37	0.31	**0.11**	**0.09**	**0.06**	**0.06**	**0.05**	■						
	sl1	0.30	0.38	0.31	0.25	0.27	0.26	0.36	0.35	0.33	0.30	0.25	0.23	0.36	0.32	**0.14**	**0.14**	**0.10**	**0.11**	**0.09**	**0.11**	■					
	sl2	0.35	0.43	0.29	0.23	0.27	0.26	0.31	0.30	0.31	0.28	0.23	0.21	0.33	0.27	**0.18**	**0.17**	**0.15**	**0.10**	**0.14**	**0.16**	**0.05**	■				
	mk1	0.22	0.31	0.23	0.17	0.30	0.27	0.35	0.34	0.30	0.26	0.20	0.18	0.35	0.31	**0.21**	**0.08**	**0.17**	**0.13**	**0.14**	**0.17**	**0.20**	**0.19**	■			
	mk2	0.16	0.23	0.29	0.23	0.40	0.36	0.37	0.36	0.36	0.32	0.25	0.24	0.39	0.33	**0.09**	**0.09**	**0.06**	**0.11**	**0.08**	**0.12**	**0.15**	**0.20**	**0.13**	■		
	xr1	0.31	0.39	0.35	0.29	0.30	0.27	0.36	0.35	0.37	0.33	0.28	0.27	0.38	0.32	**0.14**	**0.13**	**0.15**	**0.12**	**0.12**	**0.15**	**0.07**	**0.05**	**0.20**	**0.17**	■	
	xr2	0.24	0.29	0.40	0.33	0.35	0.32	0.38	0.37	0.41	0.38	0.33	0.31	0.40	0.34	**0.12**	**0.18**	**0.14**	**0.16**	**0.15**	**0.12**	**0.09**	**0.14**	**0.26**	**0.13**	**0.11**	■

Figure 1. Results on the example of the model collection ***C***

VI. Preliminary results on the example of the model collection C

Preliminary results on the example of the model collection C are given below by sequentially performing the following operations:

- calculation of digital portraits (frequency of letters of 22 common Latin characters) for all 26 works of the model collection ***C***;

- calculations by formulas (1), (2) and (3) 325 pair distances $p(T_1,\ T_2)$ between the products of the collection ***C*** (the calculation results are given in the figure 2)

- calculation using the γ-classifier tuning algorithm [6] of the optimal interval of γ values for which the value $\tau = \tau(\gamma$ of the total number of cases of violation of the hypothesis H reaches the minimum value and, therefore, the value π of the hypothesis fulfillment indicator H takes the maximum value.

Based on the data in Figure 1, the optimal half-interval of values is calculated

$$\gamma^{onm} \in [0.2142;\ 0.2160)$$

In accordance with Definition 3, this means that if the distance $p(T_1,T_2)$ between two texts does not exceed the value γ^{onm} 0.2160, then a pair of texts belongs to the same language group; if $p(T_1,T_2)$ exceeds 0.2160, then they belong to different languages.

The minimum number of violations turned out to be equal to τ = 45. In figure 1, the violation cells of the hypothesis (4) "homogeneity" are marked with a weak gray color, and the hypothesis (5) "heterogeneity" in gray.

Now it remains to calculate the efficiency of the classifier using the formula (6):

$$\pi = 1 - \tau(\gamma^{onm})/L = 0.86$$

VII. Testing the classifier

After, due to the choice of the optimal value of γ, the classifier was adjusted and the algorithm was worked out, which in 86 cases out of 100 correctly correlated the elements of the model collection to the corresponding group of Slavic languages, a natural question arises, what will be the results of the layout of other Slavic texts that are not included to the collection, for the same three language groups.

For testing the classifier, 3 texts were randomly selected:

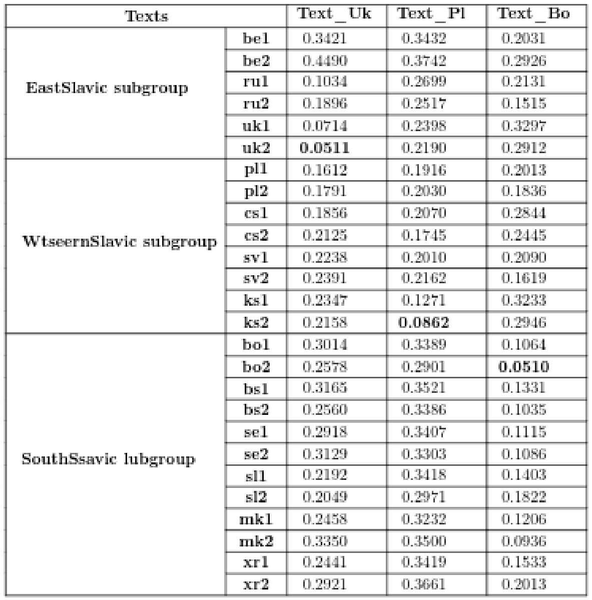
in Ukrainian **(Uk) -** V.m. Berezhnoy "HoMo Novus" (cyr., Text_Uk, 5768 *words*);

in polish **(Pl) -** A. Szklarski "Tomek wśród łowców głów" (lat.,Text_Pl, 13635 *words*);

in Bulgarian **(Bo) -** A. Karaliychev "" (cyr., Text_Bo, 2436 *words*).

For each work, just as it was done for all texts in the model collection, the DP was built on the basis of the single set of 22 Latin characters. After that, using formula (3), the distances to all 26 elements of the model collection are calculated. The results are shown in the table.

Figure 2 shows distances between the texts of collection ***C*** and three randomly selected works.



Texts		Text _Uk	Text _Pl	Text _Bo
EastSlavic subgroup	be1	0.3421	0.3432	0.2031
	be2	0.4490	0.3742	0.2926
	ru1	0.1034	0.2699	0.2131
	ru2	0.1896	0.2517	0.1515
	uk1	0.0714	0.2398	0.3297
	uk2	**0.0511**	0.2190	0.2912
WtseernSlavic subgroup	pl1	0.1612	0.1916	0.2013
	pl2	0.1791	0.2030	0.1836
	cs1	0.1856	0.2070	0.2844
	cs2	0.2125	0.1745	0.2445
	sv1	0.2238	0.2010	0.2090
	sv2	0.2391	0.2162	0.1619
	ks1	0.2347	0.1271	0.3233
	ks2	0.2158	**0.0862**	0.2946
SouthSsavic lubgroup	bo1	0.3014	0.3389	0.1064
	bo2	0.2578	0.2901	**0.0510**
	bs1	0.3165	0.3521	0.1331
	bs2	0.2560	0.3386	0.1035
	se1	0.2918	0.3407	0.1115
	se2	0.3129	0.3303	0.1086
	sl1	0.2192	0.3418	0.1403
	sl2	0.2049	0.2971	0.1822
	mk1	0.2458	0.3232	0.1206
	mk2	0.3350	0.3500	0.0936
	xr1	0.2441	0.3419	0.1533
	xr2	0.2921	0.3661	0.2013

Figure 2. Distances between texts of collection ***C***

In table cells, as the intersection of columns and rows, the values of the distances between the texts are given. In the first three columns, the nearest neighbors of the texts Text_Uk, Text_Pl and Text_Bo are uk2, ks2 and bo2, respectively, at distances of 0.0511, 0.0862, end 0.0510, respectively (marked in gray in the table). The result obtained shows that, according to the nearest neighbor method, three randomly selected works are distributed exactly according to the language groups to which they themselves belong.

VIII. Conclusion

So, the γ-classifier with a fixed value on random samples of texts with digital portraits based on the frequency of 22 Latin characters confirmed 86% statistical ability to recognize groups of works in Slavic languages. In turn, the nearest neighbor method showed the possibility of an error-free distribution of additional Slavic works in the eastern, western and southern groups of Slavic languages.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

К вопросу о метрической однородности текстов на славянских языках

З.Д. Усманов, А.А. Косимов

Аннотация – В исследованиях Р.Грея и К.Аткинсона [1] посредством статистического анализа родственных слов, У.Чанга, Ч.Кэткарта, Д.Холла и А.Гарретта [2] с помощью статистического моделирования и А.С.Касьяна и А.В.Дыбо [3] на основе лексикостатистической классификации помимо обсуждения исторических вопросов представлены генеалогические деревья, отражающие как родство, так и дивергенцию современных славянских языков. Таких деревьев достаточно много, они сходны в общих чертах и различны в небольших деталях, см. например, [3, 4]. Ареал прежде единого языка ныне разделился на три группы – восточную в составе белорусского, русского и украинского языков, западную - из чешского, словацкого, польского, кашубского и лужицких языков и южную, состоящую из болгарского, македонского, сербо-хорватского и словенского языков. В статье на примере случайно сформированной модельной коллекции из 26 текстов на 13 языках (по 2 произведения от каждого языка) устанавливается применимость γ-классификатора для автоматического распознавания принадлежности текстов той или иной группе славянских языков на основе частотности универсального для все языков набора латинских символов. Математическая модель -классификатора представляется в виде триады, составленной из цифрового портрета (ЦП) текста - распределения в тексте частотности латинских символьных униграмм; формулы для вычисления расстояний между ЦП текстами и алгоритма машинного обучения, реализующего гипотезу “однородности” произведений из одной группы языков и “неоднородности” произведений, принадлежащих разным группам языков. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, осуществлялась путем подбора оптимального значения вещественного параметра γ, минимизирующего число ошибок нарушения гипотезы “однородности”. Обученный на текстах модельной коллекции γ-классификатор показал 86%- ю точность в распознавании языков произведений. Для тестирования классификатора были выбраны 3 дополнительных случайных текста, по одному тексту для трёх разных групп славянских языков. Методом ближайшего (по расстоянию) соседа все новые тексты подтвердили свою однородность с соответствующими парами одноязычных произведений, тем самым и однородность с соответствующей группой славянских языков.

Ключевые слова – текст, язык, славяне, алфавит, универсальный набор латинских символов, частотность, униграмм, цифровой портрет текста, классификатор, обучение, распознавания, группы языков, оценка эффективности, тестирование классификатора.

Received 06.07.2021