Yarema O.B.

Applying statistical methods to the analysis of allusion in literary text

Yarema Oksana Bohdanivna, advanced student
Easteuropean National University named after Lesia Ukrayinka, Ternopil, Ukraine

Abstract. Applying statistical methods to the analysis of allusion in literary text. The aim of this article is the study of various allusive types via the methods of statistics. The statistical methods used are chi-square criterion and the coefficient of the degree of similarity. The peculiarities of the functioning of allusions are revealed through the comparison of different allusive types in three genres of literature. The process of comparison of three genres, the identification of the nucleus of each genre and other feature are presented in the article.

Keywords: *allusion, genre, chi-square, coefficient of similarity*

Text is a population of units from the point of view of statistics that can undergo observation and counting. The use of formulas with an aim to detect some characteristics of language units, seems to be difficult. But with the detecting of laws and specifics of functioning of these variables the inner content of the text messages can be decoded, as well as the reasons of text fullness with these units, their correlation and interconnection. These signs are essentially diagnostic, by means of which, on the one hand, the identification of texts is carried out, on the other, efforts are applied to penetrate into the depth of the text, not accessible to direct observation [6]. However, the text in its classical meaning [2, 7, 9, 11] is not the only object of our analysis in this article. Text can be a set or corpus of the works of one author, provided they belong to one genre. For example, a sample of poems by T.S. Eliot can be considered a set. Such text is also considered a collective one, although the level of integrity is slightly different. Integrity is ensured by the unity and stability of individual style, author affiliation to a particular school or literary movement and so on. The formation of one homogenous population of the texts is based on concrete statistical and linguist parameters. Thus, the aim of this article is to detect the variance of allusive types in three genres of literature through statistic methods. Though the phenomenon of allusion was analyzed by a great number of foreign and Ukrainian linguists, such as P. Allan, R. Leppihalme, M.V. Vorobiova, N.Y. Novokhachiova, V.P. Moskvin and others, the methods of statistic were yet not applied to the analysis of allusion. To cope with this task, the works of V.I. Perebyinis and B.N. Golovin served as the basis of this investigation.

In the course of forming the sample of text excerpts, linguists primarily define its specificity and selection criteria – linguistic and statistical, which should reflect the homogeneity of the text and provide a quality selection of material to be tested. On the one hand, this selection should be based on the general principles of selection of texts, on the other – certainly the features of units to be detected in an array of text should be considered.

An array of text that consists of identical by multitude fragments of texts by various authors is likened by the criteria of selection and requirements to the selection of material for corpus text [4]. Therefore, selection of material is carried out on the following criteria:

1) diachronic – texts should belong to the same time period, which lasted from the late 19th century to the mid-20th century, and is united by the literary movement of modernism;
2) representative – texts reflect the real state of a language in the aforementioned period preserving author's or regional language features;
3) stylistic – texts constitute fiction by prosaic writers, poets and playwrights. Critical essays, essays and other kinds of genres were not taken into account;
4) authentic – all texts are not adapted or modified and so reflect the real state of the language and organization of a work;
5) quantitative – the number of words in the passages is clearly determined and varies between 100 word forms.

Statistical calculations are based on the comparison of three main genres of literature, that of narrative, lyrical and dramatic [10, p. 10], namely the comparison of features of allusive load within these genres. Due to differences in definition and classification of genres and kinds of literature, to the group of genre we include a number of literary works similar in the type of their linguistic and cognitive orientation to an object or subject, or the act of artistic expression: a word either depicts the objective world (in case of prose work) or expresses the state of the narrator (in case of verse) or reproduces the process of verbal communication (in case of drama) [5, p.322].

To the corpus of three metasamples excerpts of works or entire texts of authors whose work is related to the literary movement of modernism were included. In forming the samples of drama and narrative genres we selected randomly text excerpts with length from 900 to 1100 word forms, which is in average of 1000 word forms each. The variation in the length of passages is caused by logical completeness of the piece of the text – a paragraph or a replica. Formation of the sub-samples in lyrical genre metasample is characterized by grouping smaller poems into one sub-sample with the number of units of 1000 word forms.

As V.I. Perebyinis states, the basis for the recognition of random sampling as one that gives reliable results of the study is the hypothesis that a large number randomly selected units from the general population should adequately represent it. Random sample corresponds to statistics law on accidental events [8, p.20].

In general 210 sub-samples, that is 70 sub-samples for each genre (or metasample), respectively, served as material for the research. According to the table of "Sufficiently big numbers" such quantity of samples will ensure reliability of the results (p) with 99% reliability and relative error (ε) at 0,09%, which is quite satisfactory for linguistic studies [3, p.57]. After all, as B.N. Golovin indicates, the experience to use statistics for the study of the basic phenomena of morphology and syntax in different styles of literary language convinces us that for sufficiently plausible data of mean values and quota sufficient will be a number of 10-20 samples with amount of 500 words. However, phenomena with small frequency require a larger number of observations of investigated frequency and quotas [3, p. 58]. Therefore, the study of allusive units, to one general population, which consists of 70 samples, were included 10 sub-samples based on the excerpts from the texts of seven authors that have been selected to represent the phenomenon in a particular genre of literature randomly.

General population of narrative genre is comprised form the literary texts of such authors as J. Joyce, V. Woolf, J. Conrad, H. D. Lawrence, W. S. Maugham, G. Orwell, A. Huxley; lyrical genre – T.S. Eliot, W.B. Yeats, H.D. Lawrence, W.H. Auden, W. Owen, E. Sitwell, D. Thomas; dramatic genre – S. Beckett, A. Wesker, W.B. Yeats, W.S. Maugham, J. Osborne, S. O'Casey, H. Pinter.

As this number of excerpts is selected from the relatively small amount of text, it is insufficient to draw conclusions about the characteristics of the functioning of allusions in three genres of literature in general, but enough to talk about the peculiarities of the allusive units in analyzed texts of the British author-modernists.

The first stage of the analysis is to determine whether the sample or some of them such as drama and prose, prose and poetry, drama and poetry are included to one general population, or there are stylistic differences among them due to linguistic or stylistic factors. Obviously, the comparison of the frequency characteristics of allusions between three metasamples is possible with nonparametric tests, one of which is a chi-square test (χ^2). This is the most commonly used criterion for the study of linguistic phenomena, because, unlike the Student's t test, it enables comparison of more than two samples. Student test, however, is useful to verify the materiality of differences and distances between pairs of arrays of comparable texts (prose-drama, prose-poetry, poetry-drama).

With "chi-square" test the empirical observation data are compared to the theoretical tabulated values and hypothesis about the distribution of the studied units is defined.

In assessing the difference between empirical and theoretical distributions we need to know the value of the probability and degrees of freedom that meet a specified level of significance. For this purpose K.Pearson developed a standard table [1, p.162], in which the values of the parameters at the intersection of probability and number of degrees of freedom give the probabilities that assess the value of χ^2.

The distribution of absolute data of each sample is represented in the table (see Table 1) together with the sum of the absolute frequencies in each metasample – 83, 293 and 579 units respectively for drama, prose and poetry, as well as the amount of each type of allusions in three metasamples.

That is, we analyze the eight types of allusions, where 107 units is the sum of mythological allusions, 267 – theological, 239 – literary, 100 – historical, 40 – folklore, 45 – common, 14 – art-allusions and 143 – personal. The number of 955 is a total amount of allusions in all metasamples.

Table 1. The distribution of the absolute frequencies of different types of allusion in three metasamples

	M	T	L	H	F	C	A	P	Σ
Drama	9	33	15	9	2	4	0	11	**83**
Prose	9	66	82	49	11	21	5	50	**293**
Poetry	89	168	142	42	27	20	9	82	**579**
Σ	**107**	**267**	**239**	**100**	**40**	**45**	**14**	**143**	**955**

Thus, $\chi^2 = 955 \cdot (1{,}07\text{-}1) = 66{,}85$.

In order to check whether the data states the significance of these differences, it is necessary to determine the number of degrees of freedom equal to the product of the number of columns minus one to the number of rows minus one [8, p.36]: f = (8-1) · (3- 1) = 14. The null hypothesis is accepted at $\chi^2 \le \chi^2_{5\%}$ and discarded when $\chi^2 \ge \chi^2_{1\%}$. For f = 14 we obtain the value $\chi^2_{5\%}$, which corresponds to a value of 23,7, and $\chi^2_{1\%} = 29{,}1$. In our case, index $\chi^2 > \chi^2_{1\%}$, or 66,85 > 29,1, that is much bigger than the tabulated values, so the null hypothesis of homogeneity of the studied metasamples is rejected. This shows the significance of the difference between the observed texts and is caused by the specifics of allusion in every text and its structural features of speech. It should be added that if χ^2 test showed significance, according to which all three samples could be combined into one general population, the distribution of allusive load would not differ between metasamples and it could be argued that the saturation of the texts by allusions is independent of generic nature of the texts or is dictated by other laws.

Obviously, if every genre of literature has a structured set of different types of allusive elements, among them are those that occur most often, so they are the nucleus of the system, and those that are extremely rare or do not occur at all, that is they are at the periphery of the matrix. It should not be forgotten that the frequency of each type of allusions not only depends on the genre of the work, but also on the ideological orientations of writer's literary and cultural aspirations, methods of construction of the text and its saturation by stylistic elements.

To determine the nucleus of the system and the periphery, we need to calculate the relative frequency of different types of allusive elements, or the percentage of the absolute frequency of the total number of items in the corpus (see Table 2). The total number of allusions for drama is 83 units, for prose – 293 units and for poetry – 579.

Relying on the data of relative frequencies it can be summarized that the nucleus of drama is represented by theological, literary and personal allusions, the main corpus includes mythological and historical allusions, and to the periphery belong common, folklore and art allusions. Accordingly, in prose the nucleus is presented by literary, theological, historical and personal allusions, the main corpus combines common allusions, and at the periphery are folklore, mythology and art allusion. In poetry, the nucleus contains theological, literary, artistic and mythological allusions, the main corpus also includes historical allusions and periphery covers folklore, art and common allusions. In terms of frequency common nucleus for three compared metasamples is theological, literary and personal allusions. Poetry is differentiated also by mythological allusions, and prose by historical ones.

Table 2. Absolute and relative frequency of allusions

№	Types of allusion	Drama		Prose		Poetry	
		abs. freq.	rel. freq.	abs. freq.	rel. freq.	abs. freq.	rel. freq.
1.	Mythological	*9,00*	*10,84*	9,00	3,07	**89,00**	**15,37**
2.	Theological	**33,00**	**39,76**	**66,00**	**22,53**	**168,00**	**29,02**
3.	Literary	**15,00**	**18,07**	**82,00**	**27,99**	**142,00**	**24,53**
4.	Historical	*9,00*	*7,80*	**49,00**	**16,72**	*42,00*	*7,25*
5.	Folklore	2,00	2,41	*11,00*	*3,75*	*27,00*	*4,66*
6.	Common	4,00	4,82	*21,00*	*7,17*	20,00	3,45
7.	Art-allusions	0,00	0,00	5,00	1,71	9,00	1,55
8.	Personal	*11,00*	*13,25*	**49,00**	**16,72**	**82,00**	**14,16**
Total		83,00	100,00	293,00	100,00	579,00	100,00

Note. In bold type are the frequencies that belong to the nucleus, in italics to the main corpus, and the rest belong to the periphery

Relying on the data of relative frequencies it can be summarized that the nucleus of drama is represented by theological, and literary allusions, the main corpus includes mythological, personal and historical allusions, and to the periphery belong common, folklore and art allusions. Accordingly, in prose the nucleus is presented by literary, theological, historical and personal allusions, the main corpus combines common and folklore allusions, and at the periphery are mythological and art allusions. In poetry, the nucleus contains theological, literary, mythological and personal allusions, the main corpus also includes historical and folklore allusions and periphery covers art and common allusions. In terms of frequency common nucleus for three compared metasamples is theological and literary allusions. Poetry is differentiated also by mythological and personal allusions, and prose by historical and personal ones.

It is interesting to see what types of allusions are characterized by frequency similarity among the three analyzed metasamples. To obtain these data, we need to define threshold zones and separate them by frequency. Follow recommendation of V.I. Perebyinis [8], we decided to take five frequencies – high, above average, below average, average and law. Step of the threshold is determined as follows: the difference between the highest and the lowest frequency of unit is divided by 5 (the number of frequencies). Therefore, for each metasamples the thresholds are different. So, for drama the highest frequency is within theological allusions – 33 units, and the lowest is within art allusions – zero frequency. The difference will be 33. Then for the frequency range 34 is divided by 5 and thus we get 6,6. According to the same scheme we get the frequency ranges for prose and poetry with ranges of 15,4 and 31,8. Allusive types are distributed according to frequency of each metasample and are presented in Table 3.

Table 3. Distribution of allusive types according to frequency

Frequency	**Drama**		**Prose**		**Poetry**	
	Range	Type of allusion	Range	Type of allusion	Range	Type of allusion
High	26,8-33,0	*Theological*	62,0-82,0	**Literary Theological**	127,6-168,0	**Literary**, ***Theological***
Above average	20,1-26,7	-	46,5-61,9	Personal, historical	95,7-127,5	-
Average	13,4-20,0	Literary	31,0-46,4	-	63,8-95,7	Mythological, personal
Below average	6,7-13,3	*Historical*, mythological, personal	15,5-30,9	Common	31,9-63,7	*Historical*
Low	0-6,6	*Art-allusions, folklore, common*	0-15,4	**Art-allusions, folklore**, mythological	0-31,8	***Art-allusions, folklore,*** *common*

Note. In bold type are similar for prose and poetry, in italics – for drama and poetry, underlined – for drama and prose.

Note. In bold type are highlighted allusions similar for prose and poetry, in italics – for drama and poetry, underlined – for drama and prose.

With the coefficient of the degree of similarity of the text corpuses (Q) comparison of three pairs of metasamples can be carried out. The value of Q (%) for a pair of **drama-prose** in the area of low frequency showed similarity of 50% – for the art allusions and folklore allusions, and at high frequencies – 50%. For **drama and poetry** in the area of high frequency similarity is 50%, in the area of frequency below average – 33,3% (historical allusions), in the area of low frequency – 100,0% (art, common and folklore allusions). For a pair of **prose and poetry**, we obtain the following indicators of similarity: zone of high frequency – 100% (literary and theological allusions), zone of low frequency – 50,0% (art and folklore allusions). In other areas the similarities of frequency of these pairs were not found.

Analyzing these figures, we can state that all frequencies between three pairs of metasamples indicate the high similarity of 100% in zones of low frequency for drama-poetry pair and high frequency of prose-poetry pair, average similarity – 50% (for high frequency zone of pair drama-poetry, low-frequency and high-frequency for pair of drama-prose, low-frequency for pair prose and poetry) and below average – 33,3% (for zones of frequency below average of drama-poetry pair). Those that have average similarity and similarity below average are genre-powerful or tend to have similarities. In cases of high similarity which we observe in two cases are not genre-powerful. However, according to preliminary data, in one genre of literature quantitatively theological allusions dominate, in other – literary ones, although both types belong to the nucleus. Moreover, this similarity may be caused by other factors. On stages of above average and average similarities were not found at all. Overall, this confirms the fact that the genre of literature affects the functioning of different types of allusive elements in these texts.

The obtained results allow checking and statistically examining of numerous texts of three general populations and establish relationships between dependent phenomena and factors that affect the appearance of allusive units in these texts. The scientific approach to the study of the use of allusive word forms in the general system of stylistic devices in three genres of literature of modernism period is an important task of identification the dependence of the functioning of these means from the genre form of text and the specific motivations of the author. This would help to find a solution why so many units at a certain segment of the text or in a total corpus is present, and how some types of units is subordinated to a particular genre of literature or author.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Вавилова Г.В. Математическая обработка результатов измерения: учебное пособие / Г.В. Вавилова. – Томск : Изд-во Томского политехнического университета, 2013. – 167 с.
Vavilova G.V. Matematicheskaia obrabotka rezultatov izmereniia: uchebnoe posobie [Mathematical processing of the measurement results] / G.V. Vavilova. – Tomsk: Izd-vo Tomskogo politekhnicheskogo universiteta, 2013. – 167 s.

2. Гальперин И.Р. Информативность единиц языка / И.Р. Гальперин. – М., 1974. – 201 с.
Galperin I.R. Informativnost’ edinits yazyka [Informativeness of language units] / I.R. Galperin. – M., 1974. – 201 s.

3. Головин Б.Н. Язык и статистика / Б.Н. Головин. – М. : Книга по Требованию, 2013. – 193 с.
Golovin B.N. Yazik I statistika [language and statistics] / B.N. Golovin. – M. : Книга по Требованию, 2013. – 193s.

4. Демська-Кульчицька О. Що таке корпус текстів. / О. Демська-Кульчицька // [Е-ресурс]
Demska-Kulchytska O. Scho take korpus tekstiv [What is the corpus of the texts] / O. Demska-Kulchytska // [Online]:
http://kulturamovy.univ.kiev.ua/KM/pdfs/Magazine64-7.pdf

5. Лобжанідзе Б.Д. Аллюзия в произведениях Сидни Шелдона / Б.Д. Лобжанидзе // [Е-ресурс]. Заголовок з екрану.
Lobzhanidze B.D. Alliuziya v proizvedeniyakh Sidni Sheldona [Allusion in the works of Sydney Sheldon] // B.D. Lobzhanidze // [Online]:
http://www.pglu.ru/lib/publications/University_Reading/2011/II/uch_2011_II_00009.pdf

6. Мартыненко Г.Я. Новые информационные технологии систематизации и исследования художественных текстов (на материале русского рассказа XX века) / Г.Я. Мартыненко // [Е-ресурс]. Заголовок з екрану.
Martynenko G.Y. Novie informatsionnie tekhnologiyi sistematizatsiyi i issledovaniya khudozhestvennikh tekstov (na material russkogo rasskaza XX veka) [New informational technologies of systematization and research of fiction] / G.Y. Martynenko // [Online]:
http://www.artinfo.ru/eva/EVA2000M/eva-papers/200007/Martynenko-R.htm

7. Мороховский А.И. К проблеме текста и его категорий / А.М. Мороховский // Текст и его категориальные признаки: Сб. науч. Тр./КГПИИЯ.К., 1989. – С. 3-8
Morokhovskyi A.I. K problem teksta i ego kategoriy [To the problem of the text and its categories]/ A.M. Morokhovskyi // Tekst i ego kategorialnie priznaki: Sb. nauch. tr. / KGPIIY. - .K., 1989. – S. 3-8.

8. Перебийніс В.І. Статистичні методи *для* лінгвістів / В.І. Перебийніс. – Вінниця : Нова книга, 2002. – 168 с.
Perebyinis V.I. Statystychni metody dlia lingvistiv [Statistical methods for the linguists] / V.I. Perebuinis. – Vinnytsia : Nova knyga, 2002. – 168 s.

9. Тураева З.Я. Лингвистика текста / З.Я. Тураева. – М. : Просвещение, 1986. – 213 с.
Turaeva Z.Y. Lingvistika teksta {Linguistics of the text] / Z.Y. Turaeva. – M. : Prosveshchenie, 1986. – 213 s.

10. Юриняк А.Б. Літературні жанри малої форми / А.Б. Юриняк. – К. : Смолоскип, 1996. – 118 с.
Yuryniak A.B. Literaturni zhanry maloyi formy [Literary genres of small form] / A.B. Yuryniak. – K. : Smoloskyp, 1996. – 118 s.

11. Halliday M.A.K., Hasan R. Cohesion in English. London / M.A.K. Halliday, R. Hasan : Longman, 1976. – 374 pp.

Yarema O.B. Применение статистических методов в исследовании аллюзий в литературном тексте

Аннотация. Целью статьи является изучение различных типов аллюзий с помощью статистических методов. В процессе исследования были использованы методы вычисления хи-квадрат критерия и коэффициент степени похожести. Особенность функционирования аллюзий раскрывается через сравнение различных аллюзивных типов в трех жанрах литературы. Процесс сопоставления трех жанровых выборок, идентификация ядерной части каждого жанра и другие признаки аллюзий представлены в статье.

Ключевые слова: *аллюзия, жанр, хи-квадрат, коэффициент сходства*