Journal of Literature and Art Studies, March 2020, Vol. 10, No. 3, 183-192
doi: 10.17265/2159-5836/2020.03.001

DAVID PUBLISHING

# Application of Multivariate Methods in the Investigation of Literary Style: *The Phantom of the Opera*

WU Xiao-fei, HU Dan

School of Foreign Studies, Zhongnan University of Economics and Law, Wuhan 430073, China

This study attempts to construct a comprehensive and multivariate quantitative investigation path of literary style, and verify the applicability of the path with the English text of *The Phantom of the Opera*. Relevant variables include text interactivity, pointed clarity, standardized type-token ratio, moving-average type-token ratio, entropy and relative repeat rate. The verified investigation not only supports the applicability of the path, but also establishes the style of *The Phantom of the Opera* with high colloquialism and small vocabulary change in various aspects.

*Keywords:* multivariate analysis, corpus, literary style, *The Phantom of the Opera*

## 1. Introduction

Traditional literary styles are often based on qualitative analysis, and the description of stylistic style is not clear enough. The conclusions of the study are mostly generalized and vague adjectives (Milic, 1982). With the rapid development of computer technology, quantitative research has been applied to the study of stylistic styles, and three kinds of influential ones, namely Multi-dimensional Analysis, Corpus Stylistics and Stylometry, have gradually formed research paradigm. However, the three have not yet reached agreement on specific variables and calculation methods, and it is difficult to describe and interpret systematically, objectively and verifiably.

Gaston Louis Alfred Leroux, (May 6, 1868-April 15, 1927), a French journalist, detective novelist. His most classic masterpieces are *The Phantom of the Opera* and *The Mystery of the Yellow Room*. *The Phantom of the Opera* has a huge influence and has been adapted into films and stage plays for many times. This study attempts to construct a comprehensive and multivariate quantitative investigation path of literary style, and verify the applicability of the path with the English text of *The Phantom of the Opera*. The aim is to use the new tool of corpus to describe the forms of speech, writing and thought in discourse, and to quantitatively and qualitatively study the stylistic features of the text in an empirical way. In view of this, this study attempts to construct a quantitative path which is comprehensive, multivariate, corpus-based stylistic corpus, and then verify the applicability of the path by using the English text of *The Phantom of the Opera* as the object from text interactivity, pointed clarity, standardized type-token ratio, moving-average type-token ratio, entropy and relative repeat rate. This study attempts to find some variables used to examine and verify the stylistic style in

WU Xiao-fei, MA, Assistant, School of Foreign Studies, Zhongnan University of Economics and Law.
HU Dan (Corresponding author), Ph.D. Professor, School of Foreign Studies, Zhongnan University of Economics and Law.

many aspects among the three quantitative stylistic research models and explores the objective and quantifiable language features in *The Phantom of the Opera.*

## 2. Literature Review

Biber (1988) created a model of multidimensional analysis to fully reveal the morphological variation of spoken and written English. This method studies the macroscopic perspective and emphasizes the multidimensionality of description, which is widely used in the study of stylistic variation. Biber and Finengan (1989) systematically examined the diachronic changes in the style of literary works and found that from the 17th to the 20th century, the evolution of English literary styles gradually showed a more direct and colloquial trend. Biber's follow-up series of studies mostly involve non-literature, especially the study of terminology, such as Biber et al. (2002) found that there are obvious differences between spoken and written language in college students' writing. The former has significant interactivity, and the latter refers to the situation. The latter reflects strong informationality and non-narrative nature. Egbert (2012) examined Dickens in the three dimensions of "presentation and description", "abstract interpretation and concrete action" and "dialogue and narrative".

Mahlberg (2007) starts with the functional classification of five-word clusters (label clusters, verbal clusters, *as if* directed clusters, body part clusters, time and place clusters) and analysis to examine the local text function of Dickens's novels. Finding clusters of words related to body parts is a clue to the development of key storylines. O'Halloran (2007) explores the complex and entangled subconscious mental activities of the protagonist Ivlina through the keyword analysis of *Ivelina* in Joyce's short story collection *Dublin*. The heroine will not follow her boyfriend's departure from Dublin.

In China, Lu Weizhong and Xia Yun (2010) introduces the main research fields and achievements of corpus stylistics, and points out the prospect of further integration of the two disciplines. Ren Yan et al. (2013) bases on the self-built British Gothic fiction corpus and the 18-19th century English novel corpus, revealing the stylistic features of the genre in terms of lexicon. Sandulescu et al. (2015) have examined the stylistic features and the validity of Joyce's novel *Finnegan's Watching the Night* by investigating the parameters, such as rank-frequency distribution, entropy, repetition rate, and rare words (Hapax legomena) and verified the effects of the measurement method in the study of the style of the literary works using the conventional language. Kubát and Cech (2016b) uses the Moving-average type-token ratio, Secondary thematic concentration and Activity to investigate from Washington to Obama. The stylistic style of the inaugural speech of the US president over the past 100 years, was found that the inaugural speech was the most concentrated in the war period, and the vocabulary richness was the highest during the economic depression. Kubát and Cech (2016a) uses 900 texts from different genres as a corpus to test the relationship between the degree of focus and lexical richness of Stylometric features, and finds relative repeat rate and secondary theme concentration. Liu Haitao and Pan Xia Xing (2015) examines the difference and correlation between Chinese new poetry and modern poetry, and verified the "natural" nature of the new poetry text by using Zipf's rank-frequency law, using Zipf-Alexkseev's law.

The above research shows that there are certain applications for the three quantitative style research paradigms at home and abroad, but the boundaries between the paradigms are clear and there is no cross-comprehensive. The quantitative research on literary style in foreign countries is more than domestic. The literary style investigation under the corpus stylistic paradigm is more than the feature recognition of writers or works such as keywords, clusters, collocations, rare and wider breadth and depth or more variables. Liu Haitao points out: “If the measurement method is an effective method of scientific science, then the introduction of measurement methods in linguistics may be a necessary way to scientificize linguistics ” (2012, p. 1). Biber (2011) advocates focusing on integrated applications. The statistical methods of early quantitative stylistic research and the recent corpus stylistic research paradigm explore the meaning of the text.

## 3. Methodology

### 3.1 Corpus Selection

In order to better examine the literary style characteristics of Gaston-Leroux’s novels, this study focuses on the text statistics, word frequency statistics, word collocation, semantic prosody and analysis of index lines from the perspective of corpus stylistics and econometric stylistics. Meanwhile, in order to verify the operability of this comprehensive path, this study compares the English text of *The Phantom of the Opera* with other novels by Gaston-Leroux. To this end, this study has constructed three corpora, namely the “*The Phantom of the Opera*” corpus[*C1*] (*The Phantom of the Opera*) , the Gaston-Leroux corpus[*C2*] (such as *The Double Life, The Double Life, The Floating Prison, The Bride of the Sun.....*) and Library of the Fiction in French[*C3*] (eg. *La Condition Humanie, La Nauseé, La Cantatrice Chauve......La Modification*). The former selects the famous novel *The Phantom of the Opera* by Gaston-Leroux, the latter are some novels that are English version published by Gaston-Leroux from 1903 to 1924, and novels published (English Version) by ten French writers from 1913 to 1958.

### 3.2 Investigation the Establishment of Variables

#### 3.2.1 Colloquial Style

Biber’s (1988) multi-dimensional examination model of linguistic morphological variation, set seven functional dimensions, each of which consists of a set of “co-occurrence” linguistic features, representing a certain style of style, with strong explanatory power. The five dimensions are: Involved vs. Informational Production, Narrative vs. Non-narrative concerns, Explicit vs. Situation-related Reference, Overt Expression of Persuasion, Abstract vs. Non-abstract Information. According to the preliminary calculations of Biber’s model, the three corpora are significantly different in dimension (1) and dimension (3), and there is no significant difference in other dimensions. O’Donnel (1974), Olson (1977), and Chafé (1982) found that compared to spoken language, written genre showed complex structure, detailed expression, and explicit, but two dimensions in involved vs. Informational production and explicit vs. Situation-dependent reference can reflect these characteristics. Therefore, this study selects them as a comparison of the spoken/written feature variables of the three corpora.

Linguistic features of dimensions in involved vs. Informational production (ie, dimension 1) covers the most (Biber, 1988, p. 104) and is the easiest to effectively distinguish the spoken/written features of discourse (Baker & Eggington, 1999, p. 350). A series of positive linguistic features (including personal verbs, ellipsis, abbreviated forms, first and second person pronouns, etc.) on this dimension are characterized by involved and strong personal sentiment, reflecting the colloquial tendency of the discourse. Negative linguistic features (including nouns, prepositions, word lengths, and type-token ratios) focus on the generation and transmission of information, reflecting the informative nature of discourse, its high degree of integration and integration, and its complex language structure. The positive linguistic features (such as *wh*-relational clauses, side-by-side phrases, nominalizations, etc.) on the dimension in explicit vs. Situation-dependent reference (ie, dimension 3) have explicit, descriptive, and complex sentence features, negative Linguistic features (such as adverbs in time, adverbs in place, other adverbs, etc.) reflect the characteristics of spoken language with strong contextual dependence, and have the features of simple sentence structure and referential dependent situation.

### 3.2.2 Lexical Richness

Lexical richness provides a comprehensive indicator for research of authors' writing style and is of interest to stylometry scientists (Smith & Kelly, 2002, p. 411). Biber (2011, p. 15) points out that the relatively mature, sophisticated computational and statistical methods in the early days of copyright ownership and literary style is a useful complement to the recent rapid development of corpus stylistics. Liu Haitao believes that: "Metrology research is the basis of all scientific methods, and it is also a necessary means to explore the structure and evolution of language in the era of big data. It is also a tool for literary research and helps to resolve objectivity-insufficient in literary studies" (2015, p. 40). This study uses the multivariate methods to investigate and select the four categories of standardized type-token ratio, moving-average type-token ratio, entropy and relative repeat rate of vocabulary, and examines the degree of richness in vocabulary use of *The Phantom of the Opera* with the two referential corpora.

The type-token ratio of vocabulary can reflect the richness and variation of vocabulary. It is a relatively simple and convenient way of investigation. The disadvantage is that the result of the investigation is greatly influenced by the length of the text or the capacity of the corpus. The standardized type-token ratio (in 1000 characters) is a lexical diversity indicator that is independent of text length or corpus capacity, so it has a higher degree when the length of the text or corpus capacity is different (Baker, 2000, p. 250). Moving-average type-token ratio is another indicator of the richness of text vocabulary, which has the advantage of being unaffected by text length or corpus capacity (Covington & McFall, 2010). It is calculated by dividing the entire text into several overlapping sub-texts (also called "windows") of the same length, letting them move forward one character at a time, and then calculating the type-token of each "window" in turn. The ratio is used to measure the lexical variability of the text as a whole. Its calculation formula is:

$$\mathrm{MATTR} = \frac{\sum_{i=1}^{N-1} Vi}{L(N-L+1)} = \frac{3+3+2+2+3+3}{3(8-3+1)} = 0.89$$

Where N is the number of overall text symbols, L is the number of subtext characters, and Vi is the number of subtext classes.

In the field of metrology linguistics, entropy is often used to measure the degree of complexity of the language. The higher the entropy, the more complicated the language is, and the simpler it is. The indicator can also be used to reflect the richness of the vocabulary of the text. The larger the text entropy, the richer the vocabulary, and the simpler it is (Pressescu et al., 2011, p. 3). The entropy value H can be obtained by the following formula:

$$\mathrm{H} = -\sum_{i=1}^{v} Pilog_2pi$$

Where Pi = fi/N refers to the frequency of occurrence of words in the text, it represents the total frequency at which the words appear, and N represents the number of symbols.

Repeat rate refers to the concentration of vocabulary, which can be used to reflect the richness of text vocabulary (Kubát & Cech, 2016[a], p. 152) and its calculation formula is:

$$RR = -\sum_{i=1}^{v} Pi^2 = \frac{1}{N^2}\sum_{i=1}^{v} fi^2$$

The repeat rate is negatively correlated with vocabulary richness. The larger the former, the simpler the vocabulary of text, and the more complicated it is. McIntosh (1967) proposed an algorithm for relative repeat rate, namely:

$$RRmc = \frac{1-\sqrt{RR}}{1-1/\sqrt{x}}$$

Where RR represents the repeat rate. The calculation of this indicator is within [0; 1] and is positively correlated with lexical richness and is almost unaffected by the length of the text (Kubát & Cech, 2016a, p. 153).

### 3. 3 Research Tools

The research tools and functions used in this paper include: dimensions in Multi-dimensional Analysis Tagger 1.3 (Nini, 2017), statistical software SPSS 19.0 for sample significance test, corpus search Software WordSmith 5.0 is used for standardized type-token ratio and generation of keyword table. Entropy and relative repeat rate by Quantitative Index Text Analyzer ( Kubát et al., 2014), moving-average type-token ratio of the text by MaWaTaR-aD (Milicka, 2017).

## 4 Results and Discussion

This section examines the colloquial style of *the Phantom of the Opera* using two variables: the interactivity/information dimension and the clear/scenario-dependent dimension. Then, the standardized type-token ratio, moving-average type/token ratio, entropy and four variables with relative repeat rate were used to examine their lexical richness.

### 4.1 Inspection of Colloquial Style

#### 4.1.1 Dimensions in Involved vs. Informational Production

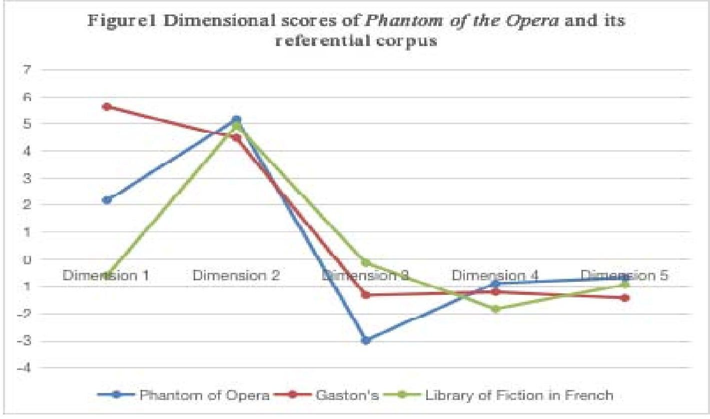

*Figure 1*. Dimensional scores of Phantom of the Opera and its referential corpus.

As shown in Figure 1, the three corpora of this study have the largest differences in the Dimensions in Involved vs. Informational Production. The Dimensions of *the Phantom of the Opera* (2.17) and Callescu (5.61) are positive, and the scores of the Library of Fiction in French (-0.6) are negative, indicating that the first two are focused on using interactivity. The language with stronger color shows its overall colloquial tendency; the latter has a high information density, and the language is more formal, reflecting a higher tendency to write. Further observation shows that the score of Gaston in this dimension is higher than that of *the Phantom of the Opera* by about 61.3%, indicating that the latter is less colloquially than the former, which means that the degree of colloquialism of *the Phantom of the Opera* is not the highest in the Gaston's novel.

### 4.1.2 Dimensions in Explicit vs. Situation-related Reference

On the whole, the scores of the three corpora on the indicated clear/scenario dependent variables (dimension 3) are all less than 0, indicating that they all exhibit the characteristics of the spoken style of the dependent context. Specifically, the dimensions of the three corpora are sorted into the *Phantom of the Opera* > Gaston's Library > Library of Fiction in French, indicating that the former two refer to the situation is higher than the latter, while the latter reflects the strong reference Clarify and interpret detailed textual features. In addition, the absolute value of *C1* is higher than that of *C2* which is about 55%, indicating that Gaston used more time, place and other adverbs in *Phantom of the Opera*.

## 4.2 Inspection of Lexical Richness

### 4.2.1 Standardized Type-token Ratio

As can be seen from Table 1, the values of the three corpora are sorted as *C1* (34.15) < *C2* (35.27) < *C3* (42.15), indicating the smallest variability in the use of the lyrics of *the Phantom of Opera*, followed by the Gaston's, the vocabulary-using in library of Fiction in French is the most versatile. Specifically, the value of *C1* is about 19% lower than that of the *C3*. There is a statistically significant difference, indicating that the former uses more succinct and more frequently used high-frequency words in the vocabulary. In addition, although the value of *C1* is slightly lower than that of the *C2*, the results of the independent sample T-test indicate that there is no significant difference between the two (P = 0.189), which proves that the Gaston's creative style used in the diversity of vocabulary use is consistent.

Table 1

*Variable Statistics on the Vocabulary Richness of Three Corpora*

| Corpus<br>Variable | C1 | C2 | C3 |
|---|---|---|---|
| Standardized Type-token Ratio | 34.15 | 35.27 | 42.15 |
| Moving-average Type-token Ratio | 0.64 | 0.66 | 0.71 |
| Entropy | 8.30 | 8.85 | 9.28 |
| Relative Repeat Rate | 0.8927 | 0.9080 | 0.9149 |

Table 2

*Results of Independent Sample T-test between C1 and C3*

| Number<br>Variable | T-value | df | P-value | MD |
|---|---|---|---|---|
| Standardized Type-token Ratio | 6.991 | 9 | 0.000 | 7.8070 |
| Moving-average Type-token Ratio | 9.558 | 9 | 0.000 | 0.6930 |
| Entropy | 8.678 | 9 | 0.000 | 0.9788 |
| Relative Repeat Rate | 8.733 | 9 | 0.000 | 0.2368 |

**4.2.2 Moving-average Type-Token Ratio**

We set the subtext capacity of the three corpora to 1000 characters, run MaWaTaRaD, and get the moving average type/token ratio of the Phantom of the Opera and the two referential libraries. As can be seen from Table 2, the vocabulary of the C3 (0.71) is the most varied, C2 (0.66) is in the middle, and the lyrics of C1 (0.64) are the most succinct. Specifically, the value of C1 is significantly lower than the C3 (P = 0.000), slightly lower than the C2, but there is no significant difference between the two (P = 0.092), again indicating the creation of the stonian has consistency in the use of vocabulary, and the change over time is not obvious.

**4.2.3 Entropy**

In terms of entropy, it can be seen from Table 1 that the lexical richness of the three corpora is from high to low: C3 (9.28), C2 (8.85) and C1 (8.30). This sorting is consistent with the standardized type-token ratio and the moving-average type/token ratio. It is again explained that the lyrics of the Phantom of Opera are the most succinct, followed by Gaston’s other four novels.

**4.2.4 Relative Repeat Rate**

It can also be seen from Table 1 that from the values of relative repeat rate, the lexical richness of the three corpora is from high to low: C3 (0.9149), C2 (0.9080) and C1 (0.8927), the ranking of the lexical richness of the three is consistent with the results of the first three parameters. The P-value of C1 is significantly lower than the C3 (P = 0.000), slightly lower than the C2 but no significant difference (P = 0.391). The results of this study show once again: Gaston tends to use simple vocabulary in the novel creation, especially the most prominent in C1, and the vocabulary use of other French writers in the similar period is more versatile and more complicated.

**4.2.5 Gaston’s and Library of Fiction in French**

To examine the lexical richness of the other four works of Gaston’s and the French novels of their age, we used SPSS to perform independent sample T-test on the four variables of the two corpora. The results showed (see Table 3) that the four variables in the two corpora were statistically significantly different. At the same time, the specific values in Table 1 also show that the richness of the vocabulary of *C2* is lower than that of the *C3*, indicating that the former has a simple language and small vocabulary changes.

Table 3

*Results of Independent Sample T-test between C2 and C3*

| Number<br>Variable | T-value | df | P-value | MD |
|---|---|---|---|---|
| Standardized Type-token Ratio | -3.469 | 18 | 0.005 | -6.478 |
| Moving-average Type-token Ratio | -0.379 | 18 | 0.020 | -0.049 |
| Entropy | -2.342 | 18 | 0.037 | -0.430 |
| Relative Repeat Rate | -2.252 | 17.99 | 0.044 | -0.006 |

### 4.3 Comparsion of the Stylistic Features with Phantom of Opera and Gaston's

Throughout the study of corpus stylistics applied to literary texts at home and abroad, according to different research perspectives, it can be roughly divided into the following seven research fields: First, the study of the theme and ideological connotation of the works; Second, the language expression of the works Research; Third, writer style research; Fourth, to verify or explain the applicability and validity of a literary criticism theory; Fifth, horizontal comparative analysis of multiple writers' works or study of the characteristics of literary works in a certain period; Sixth, to analyze the genre features and genre differences of literary texts; Seventh, corpus annotation and analysis for the purpose of stylistic study of individual linguistic features. Vocabulary is different from word in a sense. Word refers to different words in the text, that is, word types, not word tokens. The vocabulary in the novels is one of the important factors affecting the difficulty of the material.

In corpus linguistics, vocabulary diversity is represented by the Standardized Type-Token Ratio (STTR). The higher the STTR value in vocabulary selection, the higher the vocabulary diversity. With wordsmith 6.0, the class, character, and STTR of both corpora can be calculated. (Seen from Table 4 for details)

Table 4

*Data Statistics in Two Self-constructed Corpora*

| Type | C1 | C2 |
|---|---|---|
| Tokens | 85574 | 901858 |
| Types | 6408 | 705921 |
| STTR/SD | 41.65/57.23 | 44.31/56.89 |
| Average Word Length/SD | 4.23/2.19 | 4.21/2.08 |
| Average Sentence Length/SD | 15.03/14.35 | 16.47/13.61 |

Note: STTR= Standardized Type-Token Ratio, SD= Standard Deviation

It can be seen from the data in Table 4 that the STRR of the "Opera" corpus is smaller than that of the Gaston-Leroux corpus (41.65 < 44.31), but the difference is not statistically determined by the independent sample t test. The significant significance (t = -1.19, p = 0.236), that is to say, the richness and variation of the words used in the The Phantom of the Opera novel is not different from the other novels of Gaston-Leroux. In addition, it has been found that the average word length and average sentence length of the The Phantom of the Opera novels are significantly higher than the Gaston-Leroux novel corpus (t average word length = 25.32, p <0.01; t average sentence length = 72.63, p < 0.01), indicating that in the other period of the other Gaston-Leroux novels, the words in the The Phantom of the Opera novels are more complicated and the sentences are longer.

## 5 Conclusion

This study draws on and integrates three kinds of quantitative stylistic styles, such as multidimensional analysis model, corpus stylistics and econometrics, to try to construct and verify the comprehensive path and operability of multivariate methods in the investigation of literary style. Specifically, first, the study draws on the interactivity/information dimension of the multidimensional analysis model and the clear/scenario-dependent dimension to quantify the colloquial styles of the English text of *the Phantom of the Opera*. Second, in terms of lexical richness, it not only examines the traditional corpus statistical parameters. That is, the standardized type-token ratio, which also introduces the moving-average type/token ratio, entropy, and relative repeat rate of the measured stylistics. And it found that the corpus-based, systematic, multivariate stylistic style investigation path is feasible. In addition, the previous arguments about the high degree of colloquialism and the low degree of vocabulary of *the Phantom of the Opera* have been supported by comprehensive data. At the same time, the Library of Fiction in French was used as a reference, and the data was used to verify the high consistency of the Gaston's works in the diversity of vocabulary use. The results also validate Milan Kundera's evaluation of Gaston's works, that is, "Vocabulary in Gaston's novel is very limited [...] Gaston's beauty of writing is related to the limits of vocabulary" (2003, pp. 114-121). Third, this study shows that the application of measurement methods in investigation of literary style can make it get rid of the tendency from subjective to objective, from relying on intuition to science, multi-variable cross-examination and mutual authentication can make research findings more persuasive. However, the corpus of this study is too short and succinct, and the six-variable investigation of textual colloquialism and lexical richness is not comprehensive enough. Future research can try to use more variables and longer texts to further validate multivariate quantitative methods applied in the study of literary style.

## References

Bake, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, (2).

Baker, W., & Eggington, W. G. (1999). Eggington. Bilingual creativity, multidimensional analysis, and world Englishes. *World Englishes*, (3).

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future?. *Scientific Study of Literature*, (11).

Biber, D. et al. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 2002, (1).

Biber,D., & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language*, (3).

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literatur. In Tannen, D. (Ed), *In spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio. *Journal of Quantitative Linguistics*, (2).

Egbert, J. (2012). Style in nineteenth century fiction: A multi-dimensional analysis. *Scientific Study of Literature*, (2).

Kubát, M., & Cech, R. (2016a). Thematic concentration and vocabulary richness. In E. Kelih (Ed), *Issues in quantitative linguistics 4*. Lüdenscheid: RAM-Verlag.

Kubát, M., & Cech, R. (2016b). Quantitative analysis of US presidential inagural addresses. *Glottometrics*, (34).

Liu, H. T. (2012). Econometrics: A scientific approach to language research. *Guangming Daily*, 02-15.

Liu, H. T. (2015). Linguistic econometric research in the era of big data. *Journal of Shanxi University* (Philosophy and Social Science Edition), (2).

Liu, H. T., & Pan, X. X. (2015). Quantitative features of Chinese new poems. *Journal of Shanxi University* (Philosophy and Social Science Edition), (2).

Lu, W. Z., & Xia, Y. (2010). Corpus stylistics: A new approach to the study of literary stylistics. *Foreign Languages*, (1).

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora*, (1).

McIntosh, P. (1967). An indicator of diversity and the relation of certain concepts to diversity. *Ecology*, (3).

Milic, L. (1982). The annals of computing: Stylistics. *Computers and the Humanities*, (1).

Milicka, J. MaWaTaTaRaD[OL]. (2017). Retrieved from http://milicka.cz /en/mawatatarad/.

Nini, A. (2017). Multidimensional analysis tagger 1. 1-Manual[OL]. Retrieved from: http://sites.google.com/ site / multi-dimensionaltagger.

O'Donnell, C. (1974). Syntactic differences between speech and writing. *American Speech*, (1/2).

Olson, D. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, (3).

Popescu, I., Cech, R., & Altmann, G. (2011). *The Lambdastructure of texts*. Lüdenscheid: RAM-Verlag.

Ren, Y., Chen, J. S., & Ding, J. (2013). Word clusters in English gothic novels: A corpus-based study of literary stylistics. *Journal of PLA University of Foreign Languages*, (5).

Sandulescu, C. et al. (2014). Quantifying Joyce's Finnegans Wake. *Glottometrics*, (30).

Smith, A., & Kelly, C. (2002). Stylistic constancy and changes across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, (4).