© Studia methodologica, ISSN 2304-1222, No. 42. 2016

LEXICAL DIVERSITY IN SELECTED 18TH CENTURY ENGLISH NOVELS

Lukasz Stolarski

Doctor, Assistant of Professor, Institute of Foreign Philology, Jan Kochanovsky University in Kielce (POLAND)

ABSTRACT

Ten artykuł dotyczy problemu różnorodności leksykalnej mierzonej w wybranych dzieła Samuela Richardsona, Henry Fieldinga i Laurence Sterne'a. Trzej autorzy reprezentują różne style literackie. Ich powieści zostały napisane w tym samym okresie historycznym, więc to porównanie jest uzasadnione. Różnorodność leksykalną mierzono stosując systemy TTR oraz LTR na znormalizowanych próbkach z 6 powieści. Wyniki analizy wskazują, że styl Laurence'a Sterne'a jest najbardziej zróżnicowany pod względem funkcji. W odniesieniu do dwóch innych autorów można odnotować, iż dzieła Samuela Richardsona posiadają nieco wyższy poziom zróżnicowania leksykalnego niż powieści Henrego Fieldinga; Jednak różnica między dwoma autorami jest niewielka. Poza tym wyniki nie wykazały żadnych istotnych rozbieżności między ocenami za pomocą TTR i LTR i konieczności dalszych badań w celu rozwiązania tego problemu.

Słowa kluczowe: różnorodność leksykalna, bogactwo leksykalne, słownictwo, leksykalna zmienność, współczynnik TTR i LTR, wczesne powieści w języku angielskim.

This article deals with the problem of lexical diversity measured in selected works by Samuel Richardson, Henry Fielding and Laurence Sterne. The three authors represent different literary styles. Their novels were written during the same historical period, so this comparison is reasonable. Lexical diversity was measured using TTR (Type Token Ratio) as well as LTR (Lemma Token Ratio) on normalized samples from 6 novels. The results indicate that Laurence Sterne's style is the most diverse in terms of the feature under analysis. Regarding the other two authors, Samuel Richardson's works involve a slightly higher level of lexical diversity than Henry Fielding's; however, the difference between the two authors is small. Additionally, the results do not reveal any crucial proportional dissimilarities between TTR and LTR scores and more research is needed to address this issue.

Keywords: lexical diversity, lexical richness, vocabulary richness, lexical variation, type token ratio, lexeme token ratio, early novels in English

Стаття присвячена проблемі вимірювання лексичного розмаїття в творах Семюела Річардсона, Генрі Філдінга і Лоренса Стерна. Три автори представляють різні літературні стилі. Їх романи були написані в той же історичний період, тому порівняння є можливим. Лексична різноманітність вимірювалась за допомогою систем TTR і LTR на нормованих зразках з 6 романів. Результати показують, що стиль Стерна є найрізноманітнішим з точки зору функції при аналізі. Крім того, твори Річардсона включають дещо вищий рівень лексичного розмаїття, ніж романи Генрі Філдінга, хоча відмінності між цими дво-

ма авторами невеликі. Крім того, результати не виявили будь-яких важливих пропорційних відмінноісей між оцінками за допомогою TTR і LTR і необхідності додаткових досліджень для вирішення цієї проблеми.

Ключові слова: лексичне розмаїття, лексичне багатство, словниковий запас, лексична мінливість, TTR, LTR, ранні романи англійською мовою

Background

The term “lexical diversity” has been understood in several different ways. The most general interpretation holds that it is a measure of how many different words are used in a given text, but there is no consensus as to the way in which that should be calculated. The most obvious solution seems to be counting word types. The immediate problem with this method, however, is that it cannot be used to compare texts of different lengths. Longer passages may have a greater number of word forms than shorter passages, regardless of the overall author’s verbal creativity. In order to overcome such problems, several alternative solutions have been proposed. Some of these focus on standardizing the length of the texts or the duration of the utterances which are being compared. For instance, in [1] the same number of utterances are compared and in [2] recordings over a standard time are analysed. Both of these techniques are criticised in [3, pp. 220–221]. The former does not result in word token samples of equal size and the latter “confounds diversity with volubility and fluency”. It is much more dependable to standardize the number of word tokens in the samples which are being compared. This approach is referred to as “theoretical vocabulary” ([4]–[6]) and involves choosing an equal number of words from each text under comparison. The selection process may include many possible solutions. The word tokens may be chosen at random, in sub-samples, by truncating the beginning or ending of each text, etc.

Other solutions to the problem identified above concentrate on establishing a relationship between the number of types and the number of tokens. One such measure is called “Type Token Ratio” (TTR) ([7], [8]) and it is obtained by dividing the number of word types by the total number of word tokens in a text. Even though the resulting value appears to be more objective than a raw number of token types, it does not overcome the fundamental problem (cf. [9]). Namely, a new token in a text always increases the number of tokens, but not necessarily the number of types. As a consequence, TTR tends to decrease with an increase in number of tokens and, again, only texts of exactly the same token count may be directly compared. In order to overcome this problem various alterations to the basic TTR equation have been proposed. In [10] it is suggested that the number of types be divided by the square root of the number of tokens; in [11] the equation involves the square root of double the number of tokens; in [12] the proposed solution is logT/logN. More advanced formulae that model the relationship between the number of types and tokens are discussed in [13] and [3]. It has also been argued that no single index has the capacity to fully and objectively calculate lexical diversity and the choice of one method over another should be determined by the length of the text itself ([14]).

It is important to emphasize that lexical diversity may be interpreted in a broader sense than just vocabulary range. Alternative names for this phenomenon include “lexical range and balance” ([15]) “lexical richness” ([16]) or “verbal creativity”([17]). Such terms suggest a more general approach to the problem. Furthermore,

in [18] and [19] lexical diversity is directly discussed as a part of multidimensional lexical richness which involves other factors, such as “lexical sophistication” and “lexical density”. Such broader interpretations have resulted in several complex indices (i.e., Michéa’s Constant or Yule’s Characteristic K ([20]).

It has been demonstrated that lexical diversity is affected by various sociolinguistic factors, such as age ([4], [21]) and gender ([22]). Many publications also suggest the effects of text genre ([4], [6], [21], [23]–[25]). Moreover, lexical diversity and similar quantitative measures have been applied in analysing the literary style of specific authors.

The major aim of this paper is to inspect the lexical diversity of 6 novels written by three 18th century writers, each of whom represent a distinct literary style. As explained below, in addition to calculating TTR in “theoretical vocabulary” taken from the novels, a less frequently used measure involving lemmas will be used. It represents a more reliable relationship of new lexical items to all word tokens in a text.

Methods

The materials analysed in this project are 6 novels written by three 18th century English authors Samuel Richardson, Henry Fielding and Laurence Sterne. The writers are frequently characterized as the “fathers” of the English novel, although other authors, such as Jonathan Swift or Daniel Defoe, could also be included in this category. Samuel Richardson (1689-1761) is known for writing epistolary novels and his style involves the realism of presentation ([26]). The works of his analysed in this paper are “Pamela: Or, Virtue Rewarded” (1740) and “Clarissa: Or the History of a Young Lady” (1748). Henry Fielding (1707-1754) represents the realism of assessment and is best known as the author of “The History of Tom Jones, a Foundling” (1749). The other novel he wrote that will be analysed in this paper is “Amelia” (1751). Finally, Laurence Sterne (1713-1768) is famous for his “petty realism” and his novels included in the present analysis are “The Life and Opinions of Tristram Shandy, Gentleman” (1759-1767) and “A Sentimental Journey Through France and Italy” (1768). The three authors represent different literary styles and their works may be directly compared since they were written during the same historical period. A statistical summary of the vocabulary count of all six novels is provided in Table 1.

The word tokens and word types were calculated in Python Programming Language ver. 3.4.3 ([27]). The number of lemmas was also established in Python and the task required writing a script which involved the part of speech tagging available in the NLTK 3.2.1 package ([28]). After the tagging was performed, the number of lemmas was calculated using the lemmatize method of the WordNetLemmatizer class, also available in NLTK.

Although values such as TTR could be established for each of the novels, the results would not be comparable since these measures are directly affected by the number of word tokens in a sample. As discussed earlier, one of the solutions is to normalize the length of the texts under comparison. The strategy is referred to as

	Samuel Richardson		Henry Fielding		Laurence Sterne	
	Pamela: Or, Virtue Rewarded	Clarissa: Or the History of a Young Lady	The History of Tom Jones, a Foundling	Amelia	The Life and Opinions of Tristram Shandy, Gentleman	A Sentimental Journey Through France and Italy
Number of word tokens	221114	933839	346805	211601	178828	39363
Number of word types	9693	28053	14235	9793	20930	6147
Number of lexemes	7920	23913	11229	7603	18361	5344

Table 1: Basic statistics of the texts used in the analysis

"theoretical vocabulary" [4]. In the present research, the normalization of the size of the texts by each author has been performed according to the following procedure:

1 The final sample for each author should contain 200000 word tokens. This limit was set to smallest sample available, the combined number of tokens in the novels by Laurence Sterne, which is only slightly higher.

2 For works by both Samuel Richardson and Henry Fielding a script in Python was written to select exactly 100000 word tokens from each novel. The selection involved 100 sub-samples consisting of 1000 words each. The gaps between each sub-sample were calculated according to the equation $(n_t–n_s)/ss$, where n_t = the number of tokens in the whole novel, n_s = the intended number of tokens in the sample, and ss = the number of sub-samples. This guaranteed that each part of the novels was adequately represented in the resulting samples. The two samples of 100000 word tokens for works by Samuel Richardson were joined into one sample of 200000. The same was done for the two samples of 100000 word tokens taken from the novels by Henry Fielding.

3 The procedure for the two works by Laurence Sterne was similar to the one described above, but not identical. "A Sentimental Journey Through France and Italy" contains only 39363 word tokens. Consequently, all of these were added to the final sample of 200000. The remaining 160637 word tokens were selected from "The Life and Opinions of Tristram Shandy, Gentleman" using the same technique as for the works by Samuel Richardson and Henry Fielding, except that the number of sub-samples was 160.637.

A summary of the way the data was obtained from each novel is presented in Table 2.

In this project, lexical diversity is calculated in two related ways. Firstly, TTR is identified for each author under analysis. Normalization of the samples results in a relatively objective measurement. Secondly, the ratio between lemmas and word tokens, or "Lemma Token Ratio" (LTR), is provided. This measure has not been frequently used in studies on lexical diversity, which is most likely due to the problem with calculating lemmas. Word types are easy to count and several online services (e.g. http://textalyser.net) as well as offline computer programs (e.g. [29]) provide ways to perform the task. Obtaining the number of lemmas in long texts, on the other hand, involves potentially complex algorithms which are not included in many concordancing programs. Nevertheless, the NLTK package for Python contains appropriate tools to perform this task reliably and the advantages of using LTR are obvious. The value of lexical diversity based on types takes into account identical word forms and different inflectional forms of the same word are treated as separate items. LTR, on the other hand, groups such inflectional forms into single units and a more reliable picture of the actual lexical range for a given text is revealed.

	Samuel Richardson		Henry Fielding		Laurence Sterne	
	Pamela: Or, Virtue Rewarded	Clarissa: Or the History of a Young Lady	The History of Tom Jones, a Foundling	Amelia	The Life and Opinions of Tristram Shandy, Gentleman	A Sentimental Journey Through France and Italy
Number of word tokens taken from each novel	100000	100000	100000	100000	160637	39363
Number of sub-samples taken from each novel	100	100	100	100	160.637	the whole text was used
Number of word tokens in each sub-sample	1000	1000	1000	1000	1000	the whole text was used
Size of gaps between sub-samples	1211	8338	2468	1116	113	the whole text was used

Table 2: Data on the way samples were taken from each novel

All the statistical tests in this project were performed with R Programming Language [30].

Results and Conclusion

Table 3 summarizes the results in terms of lexical diversity for the samples taken from the works of Samuel Richardson, Henry Fielding and Laurence Sterne. Since the samples contain an equal number of word tokens (200000), one may directly compare the number of types calculated in each. It is quite clear that the author who used the largest number of them is Laurence Sterne (27514). This suggests that his writing style may involve the highest degree of lexical diversity among the three writers. The differences between the other two authors are less distinct. The sample from Samuel Richardson's works includes 15418 types, while the sample of Henry Fielding's novels contains 14493 types. This may indicate that Richardson's works are slightly more diverse in terms of vocabulary than Fielding's. Nevertheless, the slight difference should be tested statistically before drawing any meaningful conclusions.

A comparison of the number of lemmas reveals similar results. Again, the sample taken from Laurence Sterns' works involves twice as many lemmas as in the other two cases. Moreover, the differences between Samuel Richardson's and Henry Fielding's results are small, with the result in the former case slightly higher (9052) than in the latter case (8166).

In order to statistically test these observations, TTR and LTR scores are more appropriate. Both represent the proportion of elements in a data set, so standard proportion tests may be applied. A 3-sample test for equality of proportions for TTR scores yields a p-value much below 0.0001. This indicates that the differences are statistically significant for at least one pair. Moreover, a pairwise comparison of proportions clearly shows that the differences are valid for all the cases (all p-values are below 0.0001). Therefore, not only is the result of 0.13757 for Laurence Sterne's sample statistically different from the results of the other two authors, but also the differences in values obtained from the samples of Samuel Richardson's and Henry Fielding's texts are statistically significant.

The same statistical tests performed on LTR scores show almost identical results. A 3-sample test for equality of proportions yields a p-value below 0.0001 and a pairwise comparison of proportions indicates that all the differences are statistically significant.

These statistical calculations point to the conclusion that Laurence Sterne's literary style involves the highest level of lexical diversity. His works are clearly more advanced in this respect than the novels of the other two authors under comparison. The difference between Samuel Richardson and Henry Fielding is also statistically relevant, placing Richardson second and Fielding third. Nonetheless, the difference is not large and the literary style of the two writers are relatively similar in terms of lexical diversity.

A possible next step is to investigate whether LTR is potentially more effective than TTR in summarising lexical diversity. The present results do not suggest any crucial differences between the two measures, but a new analysis could be designed in which more samples are compared. There should be proportional dissimilarities between LTR and TTR, but it is possible that these differences are too small to be observed in this project; hence, a much larger study is necessary.

Another problem which could also be examined in future studies is the possi-

ble effect of text length on lexical diversity in novels. As discussed in the introduction, such an influence is obvious in simple TTR calculations and the measure should not be used for comparisons made between samples of unequal size. In order to overcome this problem, various alternations to TTR have been proposed. They model the relationship between the number of types and tokens in such a way as to exclude the influence of text length. Still, it is possible that there is an inherent correlation between the length of the novel and the level of lexical diversity for one and the same author regardless of the way in which calculations are made. An initial analysis performed on the samples used in the present study suggest that there may be such a relationship. A comparison of the samples of 100000 word tokens taken from the two novels by Samuel Richardson indicates that “Pamela: Or, Virtue Rewarded” involves higher lexical diversity (TTR = 0.04384, LTR = 0.02406) than “Clarissa: Or the History of a Young Lady” (TTR = 0.03004, LTR = 0.00712). A similar tendency has been observed for the two novels by Henry Fielding. A sample of 100000 word tokens taken from “Amelia” yielded higher values (TTR = 0.04628, LTR = 0.02493) than the corresponding sample taken from “The History of Tom Jones, a Foundling” (TTR = 0.04104, LTR = 0.01798). Nevertheless, in order to fully investigate this possible trend analysis based on a larger number of samples from many more novels is needed.

REFERENCES

1. Klee, T. Developmental and diagnostic characteristics of quantitative measures of children’s language production. “Topics in Language Disorders” 12(2), 1992, pp. 28–41.
2. Snow, C. E. Change in child language and child linguistics. In: Coleman, H., Cameron, L. (eds.) Change and Language, Clevedon: Multilingual Matters, 1998, pp. 75–88.
3. Durán, P., Malvern, D., Richards, B., Chipere, N. Developmental trends in lexical diversity. “Applied Linguistics” 25(2), 2004, pp. 220–242.
4. Johansson, V. Lexical diversity and lexical density in speech and writing: a developmental perspective. “Working Papers in Linguistics” 53, 2009, pp. 61–79.
5. Broeder, P., Extra, G., van Hout, T. Measuring lexical richness and diversity in second language research. “Polyglot” 8, 1986, pp. 1–16.
6. Johansson, V. Word frequencies in speech and writing: a study of expository discourse. In: Working Papers in Developing Literacy across Genres, Modalities, and Languages, vol. I, Tel Aviv, 1999, pp. 182–198.
7. Lieven, E. V. Conversations between mothers and young children: individual differences and their possible implication for the study of language learning. In: Trott, K., Dobbinson, S., Griffiths, P. (eds.) The Child Language Reader, London and New York: Routlege, 2004, pp. 11-20.
8. Bates, E., Bretherton, I., Snyder, L. From First Words to Grammar: Individual Differences and Dissociable Mechanisms, 20, Cambridge: Cambridge University Press, 1991.
9. Hess, C. W., Haug, H. T., Landry, R. G. The reliability of type-token ratios for the oral language of school age children. «Journal of Speech, Language, and Hearing Research» 32(3), 1989, pp. 536–540.
10. Guiraud, P. *Problèmes et Méthodes de la Statistique Linguistique*. Presses Universitaires de France, 1960.

11. Carroll, J. B. Language and Thought. Englewood Cliffs, NJ: Prentice-Hall, 1964.
12. Herdan, G. Type-token Mathematics, vol. 4., Mouton, 1960.
13. Sichel, H. S. Word frequency distributions and type-token characteristics. “Mathematical Scientist” 11(1), 1986, pp. 45–72.
14. McCarthy, P. M., Jarvis, S. Vocd: a theoretical and empirical evaluation. “Language Testing” 24(4), 2007, pp. 459–488.
15. D. Crystal, Profiling Linguistic Disability, London: Edward Arnold, 1992.
16. Daller, H., van Hout, R., Treffers-Daller, J. Lexical richness in the spontaneous speech of bilinguals. “Applied Linguistics” 24(2), 2003, pp. 197–222.
17. Fradis, A., Mihailescu, L., Jipescu, I. The distribution of major grammatical classes in the vocabulary of Romanian aphasic patients. “Aphasiology” 6(5), 1992, pp. 477–489.
18. Read, J. Assessing Vocabulary. Cambridge: Cambridge University Press, 2000.
19. Malvern, D. D., Richards, B. J., Chipere, N., Durán, P. Lexical Diversity and Language Development. Houndmills, Hampshire UK: Palgrave Macmillan, 2004.
20. Tweedie, F. J., Baayen, R. H. How variable may a constant be? Measures of lexical richness in perspective. “Computers and the Humanities” 32(5), 1998, pp. 323–352.
21. Strömqvist, S., Johansson, V., Kriz, S., Ragnarsdottir, H., Aisenman, R., Ravid, D. Toward a cross-linguistic comparison of lexical quantain speech and writing. “Written Language & Literacy” 5(1), 2002, pp. 45–67.
22. Le Normand, M.-T., Parisse, C., Cohen, H. Lexical diversity and productivity in French preschoolers: developmental, gender and sociocultural factors. “Clinical Linguistics & Phonetics” 22(1), 2008, pp. 47–58.
23. Halliday, M. A. Spoken and Written Language. 1989.
24. Ure, J., Ellis, J. Register in descriptive linguistics and linguistic sociology. In: Uribe-Villegas, O. (ed.) Issues in Sociolinguistics, The Hague: Mouton, 1977, pp. 197–243.
25. Sadeghi, K., Dilmaghani, S. K. The relationship between lexical diversity and genre in Iranian EFL learners’ writings. “Journal of Language Teaching and Research” 4(2), 2013, pp. 328-334.
26. Watt, I. P. The Rise of the Novel: Studies in Defoe, Richardson and Fielding. Berkeley and Los Angeles, California: University of California Press, 2001.
27. Python Software Foundation, Python Language Reference (version 3.4.3) [computer software], 2016.
28. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python. O’Reilly Media, Inc., 2009.
29. Hüning, M. TextSTAT — Simple Text Analysis Tool (version 3.0) [computer software]. Berlin: Freie Universität, 2015.
30. R Development Core Team, R: A Language and Environment for Statistical Computing (version 3.0.3) [computer software]. Vienna, Austria, 2013.