Quantitative Parameters of Lucy Montgomery's Literary Style

Nataliia Hrytsiv, Tetiana Shestakevych and Julia Shyyka

Lviv Polytechnic National University, Stepana Bandery Street, 12, Lviv, 79000, Ukraine

Abstract

This article focuses on the main features of quantitative comparative analysis, as well as on the key features of statistical linguistics used in the consolidated analysis of the English original text along with the translation into Ukrainian. From a statistical viewpoint, in the Anglo-Ukrainian parallel industry, there is no to little relevant research in this area, so the main attention is paid to the works of Canadian writer Lucy Maud Montgomery. The study compares the original document with the corresponding translated text based on software specifically designed for the task, i.e. the combination of XML markup pattern, spreadsheet of Microsoft Excel; also Python programming language. Results obtained present data in ratio, such as coefficient of diversity, average token repeat per text, the exclusivity coefficient, concentration of vocabulary; it also contains findings on absolute and Relevant Frequency Distribution in source text and its translation.

Keywords [1]

linguistic metrology, quantitative comparative analysis, text marking, applied linguistics, translation.

1. Introduction

The current study tends to elucidate dominant features of quantitative comparative analysis. Key characteristics of statistical linguistics as being applied to learning quantitative parameters of the source (English) text and its translation (Ukrainian) are highlighted. In the center of attention is the creativity of a Canadian writer Lucy Maud Montgomery. The choice is made due to the lack of relevant research studies in the aspect of linguistic metrology of the given novel from statistical perspective. Thus, in focus is data in ratio, i.e. coefficient of diversity, average token repeat per text, coefficient of exclusivity, vocabulary concentration; it also contains findings on absolute and Relevant Frequency Distribution in source text and its translation obtained as based on "Anne of Green Gables" novel written by Lucy Maud Montgomery [17] in parallel with the Ukrainian translation done by Anna Vovchenko [16]. The subsection "Preliminary remarks" provides a solid rationale for choosing the material presented and methods and approaches applied, whilst at the same time addressing the typological specificity of the analyzed corpus. The subsection "Characterization of the framework and coefficients" offers an extensive overview of research and practice peculiarities of obtaining statistical data in fiction, tools used for data processing, and obtained results for further incorporation in linguistic analysis of an artistic text. The subsection "Results" dwells upon the differences and similarities between the original and its corresponding translated text with the help of software, exclusively developed for the current task. This software includes the combination of merits from Python programming language possibilities, markup language XML, and Microsoft Excel spreadsheets. Conclusions are derived at on the basis of study findings. The results of the presented research will be of use for text attribution, language learning and translator training as well as statistical documentation of Canadian literary style in Ukrainian translation.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: nhrytsiv@yahoo.com (N. Hrytsiv); Tetiana.v.shestakevych@lpnu.ua (T. Shestakevych); julia-shyika@ukr.net (Ju. Shyyka)
ORCID: 0000-0001-6660-7161 (N. Hrytsiv); 0000-0002-4898-6927 (T. Shestakevych); 0000-0003-2474-0479 (Ju. Shyyka)

2. Preliminary remarks

The widespread use of computer-based ways in modern language studies and natural language processing has become increasingly popular. The combination of both, philological approach in combination with statistically-oriented approaches promises to open new possibilities for in-depth philological and translation analysis.

The research under discussion has certain deliberate limitations: it depicts statistical parameters data (convergent and divergent) of the original text and the translation. Thus, on the basis of a novel by Lucy Montgomery, our research bares the intention to verify absolute and relevant distribution, probability measurement, also: N, max, min, R, Mo, Md, Ẋ, Ϭ, ν, Sẋ, ε.

Besides, we take support from available tools of statistical approaches as borrowed from the dimension of mathematical, applied (computational) linguistics. Such an application comforts the process while contrasting the original text embedded within the source language and the translated text data found within the target language. For the sake of digital text corpora processing – along with their future applications, – many existing mechanisms of this kind seems to have proven the efficiency within versatile domains of applied linguistics and language technology [1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 13].

2.1. Reasoning of the study

In building, on the one hand, and resulting from, on the other, quantitative comparative analysis is effective in producing statistical profiles of the author and translator. A Canadian writer Lucy Maud Montgomery is chosen for the case study from the perspective of parallel translation corpus; the explanation is that it has not been previously studied from the statistical perspective and mathematical linguistics point of view in English-Ukrainian comparison. It is carried out from the viewpoint of statistical linguistics on the content of a digitally processed and marked up corpus of texts.

We make use of quantitative analysis, which is of frequent application in modern linguistic studies. This choice presupposes drawing a holistic statistics profile of a certain author and the translator of the corresponding writing. Preliminary results on selected coefficients have already got published [9]. This article is the continuation and the holistic presentation of findings. We believe the obtained results to eventually become useful in tracing, first and foremost, the deviations of parallel profiling of ST and TT.

2.2. Typological and implicational characteristics of Anne of Green Gables document

- Language data tagging – Tagging of the text related to the subject up to the expression layer.
- Diglot – English and Ukrainian analogous texts.
- Full-length text – Comprise of the entire source and target texts.
- Writer's idiolect – The text collection contains the text by L. M. Montgomery and its rendition.
- Exemplary – It is designed for linguistic statistical comparative analysis of a source and also target texts.
- Collateral – Original text in English compared with the Ukrainian translation.
- Fixed – Does not supply a constant refill of many text collections.
- Written – The text collection includes works in written form.

3. Characterization of the framework and coefficients

3.1. Abbreviations and shortenings

1. Document
- ST – source text.
- TT – target text.
- SL – source language.
- TL – target language.
- TS – translation studies.
- PTC – parallel translation corpus.
- XML –extensible markup language.

2. Frequency distribution
- N – number of meanings
- max – maximal meaning
- min – minimal meaning
- R – range
- Mo – mode
- Md – median
- $\dot{X}$ – mean
- б – standard deviation
- ν – coefficient of variation
- $S\dot{x}$ – standard error
- E – measurement error

3. Ratio
- Kd – coefficient of diversity
- Kwr – average token repeat per text
- Kev – ratio of exclusivity
- Ken – ratio of exclusivity of a text
- Kvc – vocabulary concentration ratio
- Knc – text concentration ratio

3.2. Constructing and organizing the material

1. Construction.

In our parallel translation corpus text documents are taken from the original in English and the Ukrainian translation of this specific text. Text samples of ST and TT are designed to mirror reflect each other on text levels.

2. Tools.

We have prioritized XML, since in XML, tags are used for logical data markup. It made possible to set our own markup rules and adjust (align, augment, restructure) needed data. A special software was created on the basis of Python programming language.

3. Application.

The PTC is used in order to match texts, to compare and contrast ST and TT, to normalize and align the original and the translated texts.

3.3. Theoretical inlay

Among the basic research activity of human-computer interplay concerning language comprehension, “corpus linguistics” is defined as one of the main realms of applied linguistics by authors [9].

As mentioned above, this study of ours is the counterpart of a large-scale project on researching the statistics of literary writings of Canadian authors. Certain elaborations have already been publicized. In the publication [9], we have prior discussed the theoretical prerequisites on the basic of scholarly ideas of [14; 15; 18; 19; 20], the authors present versatile ideas of scholars; on this basis, they clearly confirm the importance of text corpora in the nowadays state of linguistic studies.

As a consequence, the type of corpus, namely PTC is preferred in sphere of inter-lingual research, also in establishment of analogous linguistic collections, which is of ultimate value for translation. In the study, we try to investigate the achievement and advances of PTC with the intension to explore literary style in fiction by means of statistical linguistics. We rely on elaborations of [3; 8; 12; 19], whose research, among other, is dedicated to re-lexicalizations.

Language corpora have the potential to empower translation studies textual analysis. As for the advances, first of all, they consist of texts in electronic form. This is why they can be easily stored, distributed, processed and manipulated from the perspective of enhancing their usefulness [12]. With this in mind, corpus-based translation studies has the potential to open new horizons in terms of decentering and functioning as a dynamic force in translation studies [12].

In addition, considering the theoretical and practical problems in relation to author's idiolect and taking into account the overlap with translator's individual style – profoundly discussed and studied from humanitarian perspective, – the particular difficulty of such approach, among other challenges, is to find a golden middle between philological profiling and multifaceted involvement of mathematical linguistics and applied linguistics tools within human language operations. That such cases are common is of little surprise since there is often a conflicting need to objectively justify the results and support research elaborations with statistics.

3.4. Quantitative parameterizing of author's style

At first, the text materials have been converted into electronic format by means of the ABBYY Fine Reader software. Then, the document was saved in .docx format. The texts were afterwards normalized in the MS Word editor and proofread for the sake of spelling, punctuation, grammar mistakes. Then, structural mark-up was conducted using paragraph mark ( ), sentence tag (<s n = 100> </s>), and page mark (<bp n = 7 />).

Attention is also given to contrasting ST and TT results on *absolute and relevant distribution, N, max, min, R, Mo, Md, $\dot{X}$, $\bar{O}$, v, $S\dot{x}$, ε.*

4. Results

4.1. Statistics in PTC

Table 1
Text length in different units

Unit	ST number	TT number
Letter	434605	403003
Lexeme	102741	82942
Sentence	6678	6666
Paragraph	1762	1770
Chapter	37	38

Thus, the most obvious fluctuation occurs on the level of lexemes with the difference of 19799 tokens lacking in ST.

Table 2
Illustrative comparing of quantitative characteristics of lexical tokens in ST and TT

Statistics	ST	TT
Number of lexical tokens used, (N)	102996	83126
Number of lemmas, (W)	7769	16043
Number of lemmas, used once (Onc)	3627	9800
Number of lemmas, used for ten times and more, (Te)	1058	971

Statistical calculations resulted in the graphs of comparison of the corresponding numbers (sentences, paragraphs, parts) from the ST and TT.

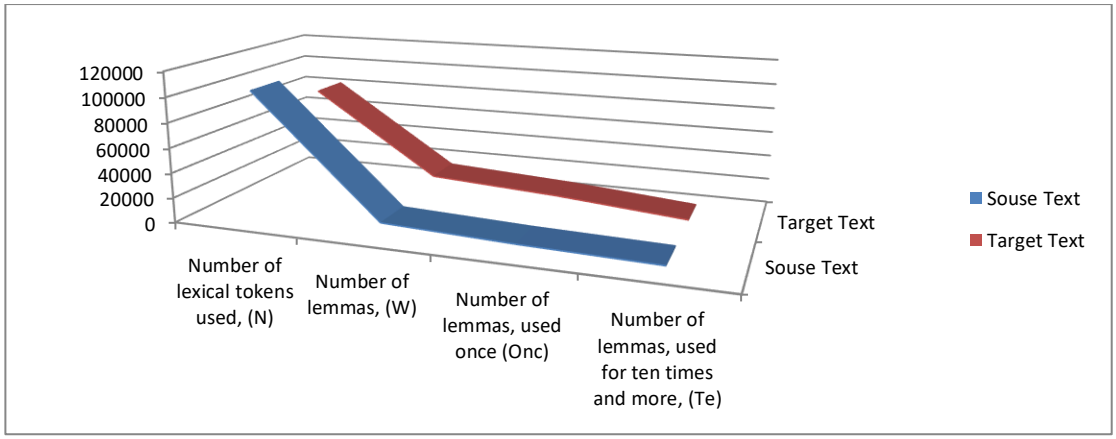

Figure 1: Comparison of ST and TT quantitative characteristics of lexical tokens as well as lemmas

Table 3
Quantitative characteristics of the lexical level of ST and TT

Coefficient	ST	TT
Coefficient of diversity, K_d	0,08	0,19
Coefficient of exclusivity, K_{en}	0,04	0,12
Coefficient of concentration, K_{nc}	0,01	0,01
Average lemma repetition, K_{wr}	13,26	5,18

Calculations show the following: K_d and K_{en} are higher in TT than in ST. Along with this, average lemma repetition is higher in ST and comprises 13, 26. To mention TT with 5,18 of the same nature.

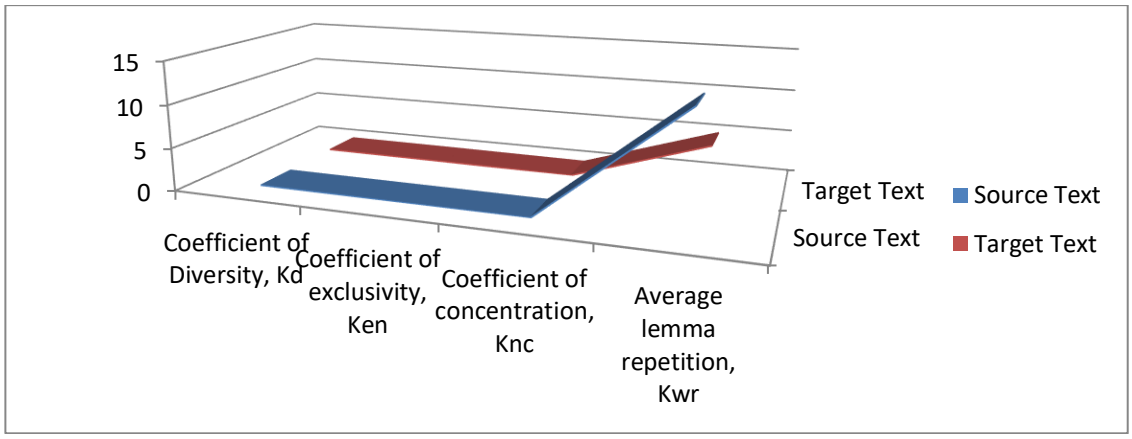

Figure 2: Correlation of ratio in ST and TT

4.2. Absolute and Relevant Frequency Distribution in ST and TT

4.2.1. Distribution of the number of lexical tokens in the paragraphs of the text

The whole ST data. 9 — 39 (2,21%); 13 — 36 (2,04%); 16 — 35 (1,99%); 8 — 31 (1,76%); 14 — 31 (1,76%); 30 — 31 (1,76%); 7 — 30 (1,70%); 12 — 30 (1,70%); 21 — 30 (1,70%); 22 — 29 (1,65%); 29 — 28 (1,59%); 11 — 27 (1,53%); 23 — 27 (1,53%); 27 — 27 (1,53%); 6 — 24 (1,36%); 15 — 24 (1,36%); 24 — 24 (1,36%); 10 — 23 (1,31%); 18 — 23 (1,31%); 20 — 23 (1,31%); 26 — 22 (1,25%); 28 — 22 (1,25%); 37 — 22 (1,25%); 44 — 22 (1,25%); 55 — 22 (1,25%); 33 — 21 (1,19%); 17 — 20 (1,14%); 19 — 20 (1,14%); 34 — 20 (1,14%); 5 — 19 (1,08%); 46 — 19 (1,08%); 54 — 19 (1,08%); 32 — 18 (1,02%); 36 — 18 (1,02%); 25 — 17 (0,96%); 38 — 17 (0,96%); 39 — 17 (0,96%); 41 — 17 (0,96%); 45 — 17 (0,96%); 51 — 17 (0,96%); 43 — 16 (0,91%); 48 — 15 (0,85%); 63 — 15 (0,85%); 64 — 15 (0,85%); 79 — 15 (0,85%); 2 — 14 (0,79%); 35 — 14 (0,79%); 42 — 14 (0,79%); 47 — 14 (0,79%); 49 — 14 (0,79%); 3 — 13 (0,74%); 52 — 13 (0,74%); 66 — 13 (0,74%); 53 — 12 (0,68%); 58 — 12 (0,68%); 71 — 12 (0,68%); 31 — 11 (0,62%); 50 — 11 (0,62%); 56 — 11 (0,62%); 65 — 11 (0,62%); 72 — 11 (0,62%); 76 — 11 (0,62%); 84 — 11 (0,62%); 57 — 10 (0,57%); 60 — 10 (0,57%); 68 — 10 (0,57%); 87 — 10 (0,57%); 101 — 10 (0,57%); 61 — 9 (0,51%); 67 — 9 (0,51%); 70 — 9 (0,51%); 74 — 9 (0,51%); 80 — 9 (0,51%); 85 — 9 (0,51%); 96 — 9 (0,51%); 40 — 8 (0,45%); 75 — 8 (0,45%); 81 — 8 (0,45%); 95 — 8 (0,45%); 97 — 8 (0,45%); 73 — 7 (0,40%); 78 — 7 (0,40%); 82 — 7 (0,40%); 93 — 7 (0,40%); 105 — 7 (0,40%); 4 — 6 (0,34%); 77 — 6 (0,34%); 88 — 6 (0,34%); 90 — 6 (0,34%); 94 — 6 (0,34%); 108 — 6 (0,34%); 111 — 6 (0,34%); 112 — 6 (0,34%); 141 — 6 (0,34%); 83 — 5 (0,28%); 98 — 5 (0,28%); 119 — 5 (0,28%); 133 — 5 (0,28%); 134 — 5 (0,28%); 59 — 4 (0,23%); 62 — 4 (0,23%); 91 — 4 (0,23%); 92 — 4 (0,23%); 104 — 4 (0,23%); 106 — 4 (0,23%); 110 — 4 (0,23%); 121 — 4 (0,23%); 123 — 4 (0,23%); 130 — 4 (0,23%); 132 — 4 (0,23%); 160 — 4 (0,23%); 166 — 4 (0,23%); 181 — 4 (0,23%); 69 — 3 (0,17%); 86 — 3 (0,17%); 99 — 3 (0,17%); 102 — 3 (0,17%); 107 — 3 (0,17%); 109 — 3 (0,17%); 117 — 3 (0,17%); 125 — 3 (0,17%); 137 — 3 (0,17%); 140 — 3 (0,17%); 144 — 3 (0,17%); 147 — 3 (0,17%); 156 — 3 (0,17%); 157 — 3 (0,17%); 169 — 3 (0,17%); 177 — 3 (0,17%); 213 — 3 (0,17%); 1 — 2 (0,11%); 89 — 2 (0,11%); 100 — 2 (0,11%); 113 — 2 (0,11%); 115 — 2 (0,11%); 116 — 2 (0,11%); 118 — 2 (0,11%); 122 — 2 (0,11%); 127 — 2 (0,11%); 129 — 2 (0,11%); 131 — 2 (0,11%); 135 — 2 (0,11%); 138 — 2 (0,11%); 142 — 2 (0,11%); 148 — 2 (0,11%); 150 — 2 (0,11%); 153 — 2 (0,11%); 155 — 2 (0,11%); 159 — 2 (0,11%); 171 — 2 (0,11%); 174 — 2 (0,11%); 182 — 2 (0,11%); 184 — 2 (0,11%); 190 — 2 (0,11%); 196 — 2 (0,11%); 207 — 2 (0,11%); 214 — 2 (0,11%); 222 — 2 (0,11%); 294 — 2 (0,11%); 452 — 2 (0,11%); 103 — 1 (0,06%); 114 — 1 (0,06%); 120 — 1 (0,06%); 124 — 1 (0,06%); 126 — 1 (0,06%); 139 — 1 (0,06%); 143 — 1 (0,06%); 145 — 1 (0,06%); 146 — 1 (0,06%); 149 — 1 (0,06%); 152 — 1 (0,06%); 154 — 1 (0,06%); 162 — 1 (0,06%); 163 — 1 (0,06%); 165 — 1 (0,06%); 167 — 1 (0,06%); 170 — 1 (0,06%); 172 — 1 (0,06%); 173 — 1 (0,06%); 178 — 1 (0,06%); 179 — 1 (0,06%); 180 — 1 (0,06%); 183 — 1 (0,06%); 185 — 1 (0,06%); 188 — 1 (0,06%); 192 — 1 (0,06%); 194 — 1 (0,06%); 195 — 1 (0,06%); 197 — 1 (0,06%); 199 — 1 (0,06%); 203 — 1 (0,06%); 204 — 1 (0,06%); 205 — 1 (0,06%); 208 — 1 (0,06%); 212 — 1 (0,06%); 215 — 1 (0,06%); 216 — 1 (0,06%); 220 — 1 (0,06%); 225 — 1 (0,06%); 229 — 1 (0,06%); 232 — 1 (0,06%); 236 — 1 (0,06%); 239 — 1 (0,06%); 240 — 1 (0,06%); 245 — 1 (0,06%); 250 — 1 (0,06%); 252 — 1 (0,06%); 263 — 1 (0,06%); 271 — 1 (0,06%); 275 — 1 (0,06%); 279 — 1 (0,06%); 283 — 1 (0,06%); 298 — 1 (0,06%); 302 — 1 (0,06%); 303 — 1 (0,06%); 304 — 1 (0,06%); 323 — 1 (0,06%); 328 — 1 (0,06%); 353 — 1 (0,06%); 366 — 1 (0,06%); 370 — 1 (0,06%); 371 — 1 (0,06%); 379 — 1 (0,06%); 394 — 1 (0,06%); 495 — 1 (0,06%); 522 — 1 (0,06%); 533 — 1 (0,06%); 574 — 1 (0,06%); 582 — 1 (0,06%); 637 — 1 (0,06%); 655 — 1 (0,06%); 690 — 1 (0,06%);.

The whole TT data. 7 — 47 (2,66%); 10 — 45 (2,54%); 13 — 43 (2,43%); 21 — 40 (2,26%); 8 — 39 (2,20%); 12 — 39 (2,20%); 9 — 37 (2,09%); 16 — 37 (2,09%); 19 — 37 (2,09%); 15 — 34 (1,92%); 5 — 33 (1,86%); 6 — 33 (1,86%); 11 — 32 (1,81%); 17 — 32 (1,81%); 18 — 31 (1,75%); 30 — 30 (1,69%); 25 — 28 (1,58%); 20 — 27 (1,53%); 31 — 27 (1,53%); 14 — 26 (1,47%); 24 — 26 (1,47%); 26 — 25 (1,41%); 28 — 25 (1,41%); 37 — 25 (1,41%); 4 — 23 (1,30%); 22 — 23

(1,30%); 23 — 23 (1,30%); 34 — 23 (1,30%); 39 — 23 (1,30%); 32 — 21 (1,19%); 40 — 20 (1,13%); 36 — 19 (1,07%); 47 — 19 (1,07%); 41 — 18 (1,02%); 50 — 18 (1,02%); 57 — 18 (1,02%); 27 — 17 (0,96%); 33 — 17 (0,96%); 35 — 17 (0,96%); 38 — 17 (0,96%); 42 — 17 (0,96%); 44 — 16 (0,90%); 3 — 15 (0,85%); 29 — 15 (0,85%); 45 — 14 (0,79%); 63 — 14 (0,79%); 64 — 13 (0,73%); 46 — 12 (0,68%); 48 — 12 (0,68%); 56 — 12 (0,68%); 59 — 12 (0,68%); 67 — 12 (0,68%); 2 — 11 (0,62%); 54 — 11 (0,62%); 65 — 11 (0,62%); 68 — 11 (0,62%); 49 — 10 (0,56%); 53 — 10 (0,56%); 58 — 10 (0,56%); 60 — 10 (0,56%); 71 — 10 (0,56%); 77 — 10 (0,56%); 80 — 10 (0,56%); 51 — 9 (0,51%); 70 — 9 (0,51%); 76 — 9 (0,51%); 83 — 9 (0,51%); 43 — 8 (0,45%); 62 — 8 (0,45%); 72 — 8 (0,45%); 90 — 8 (0,45%); 94 — 8 (0,45%); 115 — 8 (0,45%); 52 — 7 (0,40%); 66 — 7 (0,40%); 69 — 7 (0,40%); 73 — 7 (0,40%); 78 — 7 (0,40%); 79 — 7 (0,40%); 91 — 7 (0,40%); 74 — 6 (0,34%); 75 — 6 (0,34%); 84 — 6 (0,34%); 86 — 6 (0,34%); 98 — 6 (0,34%); 61 — 5 (0,28%); 82 — 5 (0,28%); 110 — 5 (0,28%); 118 — 5 (0,28%); 130 — 5 (0,28%); 132 — 5 (0,28%); 157 — 5 (0,28%); 55 — 4 (0,23%); 92 — 4 (0,23%); 101 — 4 (0,23%); 105 — 4 (0,23%); 111 — 4 (0,23%); 117 — 4 (0,23%); 124 — 4 (0,23%); 137 — 4 (0,23%); 146 — 4 (0,23%); 1 — 3 (0,17%); 81 — 3 (0,17%); 85 — 3 (0,17%); 93 — 3 (0,17%); 96 — 3 (0,17%); 97 — 3 (0,17%); 102 — 3 (0,17%); 104 — 3 (0,17%); 108 — 3 (0,17%); 121 — 3 (0,17%); 209 — 3 (0,17%); 87 — 2 (0,11%); 88 — 2 (0,11%); 89 — 2 (0,11%); 95 — 2 (0,11%); 99 — 2 (0,11%); 106 — 2 (0,11%); 109 — 2 (0,11%); 116 — 2 (0,11%); 119 — 2 (0,11%); 123 — 2 (0,11%); 127 — 2 (0,11%); 133 — 2 (0,11%); 136 — 2 (0,11%); 139 — 2 (0,11%); 148 — 2 (0,11%); 155 — 2 (0,11%); 158 — 2 (0,11%); 160 — 2 (0,11%); 161 — 2 (0,11%); 173 — 2 (0,11%); 180 — 2 (0,11%); 193 — 2 (0,11%); 252 — 2 (0,11%); 263 — 2 (0,11%); 271 — 2 (0,11%); 100 — 1 (0,06%); 103 — 1 (0,06%); 112 — 1 (0,06%); 113 — 1 (0,06%); 120 — 1 (0,06%); 125 — 1 (0,06%); 128 — 1 (0,06%); 129 — 1 (0,06%); 131 — 1 (0,06%); 134 — 1 (0,06%); 135 — 1 (0,06%); 140 — 1 (0,06%); 141 — 1 (0,06%); 142 — 1 (0,06%); 145 — 1 (0,06%); 149 — 1 (0,06%); 151 — 1 (0,06%); 154 — 1 (0,06%); 163 — 1 (0,06%); 164 — 1 (0,06%); 165 — 1 (0,06%); 166 — 1 (0,06%); 167 — 1 (0,06%); 168 — 1 (0,06%); 169 — 1 (0,06%); 174 — 1 (0,06%); 178 — 1 (0,06%); 179 — 1 (0,06%); 181 — 1 (0,06%); 182 — 1 (0,06%); 186 — 1 (0,06%); 187 — 1 (0,06%); 189 — 1 (0,06%); 190 — 1 (0,06%); 198 — 1 (0,06%); 204 — 1 (0,06%); 211 — 1 (0,06%); 213 — 1 (0,06%); 220 — 1 (0,06%); 229 — 1 (0,06%); 232 — 1 (0,06%); 234 — 1 (0,06%); 250 — 1 (0,06%); 264 — 1 (0,06%); 270 — 1 (0,06%); 274 — 1 (0,06%); 285 — 1 (0,06%); 294 — 1 (0,06%); 298 — 1 (0,06%); 305 — 1 (0,06%); 332 — 1 (0,06%); 340 — 1 (0,06%); 357 — 1 (0,06%); 390 — 1 (0,06%); 417 — 1 (0,06%); 441 — 1 (0,06%); 477 — 1 (0,06%); 491 — 1 (0,06%); 495 — 1 (0,06%); 512 — 1 (0,06%); 564 — 1 (0,06%);.

Description: from the above presented distribution data we can see the results of the number of lexical tokens found in the paragraphs of the text. This gives up a clear picture of the overall statistics. Let us, for example, consider the first three figures of whole ST: 9 — 39 (2,21%); 13 — 36 (2,04%); 16 — 35 (1,99%) and TT data: 7 — 47 (2,66%); 10 — 45 (2,54%); 13 — 43 (2,43%). It becomes obvious that the number of tokens in the paragraphs of the original text exceeds the number of tokens in the paragraphs of the translation text. It gives the grounds for further philological research on the lexical level while taking into account the context of the compared texts. In other words, why the translator shortened the translation (the article on this research is forthcoming).

To facilitate the perception of the abovementioned data, was selected top-30 frequencies of lexical tokens in both ST and TT paragraphs. At Fig. 3 it is seen, that the numbers of source lexical tokens are mostly lower than the appropriate tokens in TT. Also, there are some extreme cases, for example, in TT, there are 4 lexical tokens in 23 paragraphs, while there are no paragraphs with such number of lexical tokens in ST.

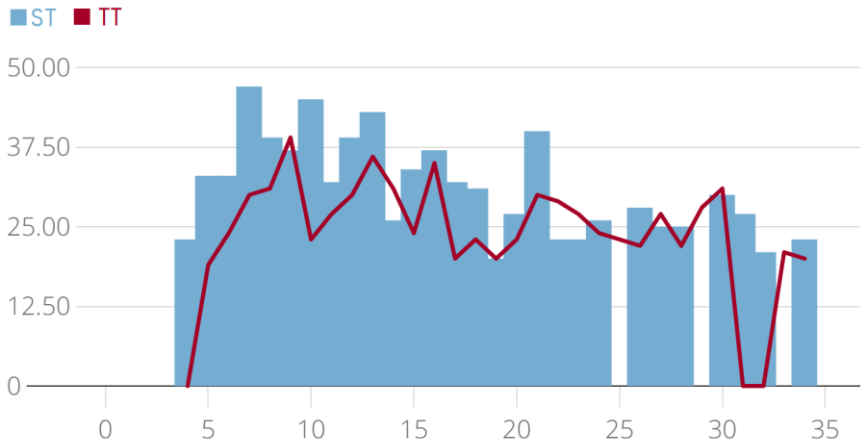

Figure 3: The correlation between number of lexical tokes in ST and TT paragraphs (the horizontal axis represents the number of lexical tokens, and the vertical axis stands for the number of paragraphs)

Numeric characteristics of data

ST numeric characteristics. Number of meanings (N) — 1762; maximal meaning (max) — 690; minimal meaning (min) — 1; range (R) — 689; mode (Mo) — 9; median (Md) — 116,5; mean ($\dot{X}$) — 58,31; standard deviation (б) — 65,74; coefficient of variation (ν) — 1,1275; standard error ($S\dot{x}$) — 1,5662; measurement error (ε) — 0,0526.

TT numeric characteristics. Number of meanings (N) — 1770; maximal meaning (max) — 564; ; minimal meaning (min) — 1; range (R) — 563; mode (Mo) — 7; median (Md) — 99,5; mean ($\dot{X}$) — 46,86; standard deviation (б) — 53,87; coefficient of variation (ν) — 1,1495; standard error ($S\dot{x}$) — 1,2803; measurement error (ε) — 0,0536.

ST and TT data contrast. The Table 4 below summarizes and contrasts the findings of the whole text concerning distribution of lexical tokens in the paragraphs of the text**.**

Table 4

Contrast of ST and TT the findings on lexical tokens in the paragraphs of the text

Unit	ST	TT
N	1762	1770
max	690	564
min	1	1
R	689	563
Mo	9	7
Md	116,5	99,5
$\dot{X}$	58,31	46,86
б	65,74	53,87
ν	1,1275	1,1495
$S\dot{x}$	1,5662	1,2803
ε	0,0526	0,0536

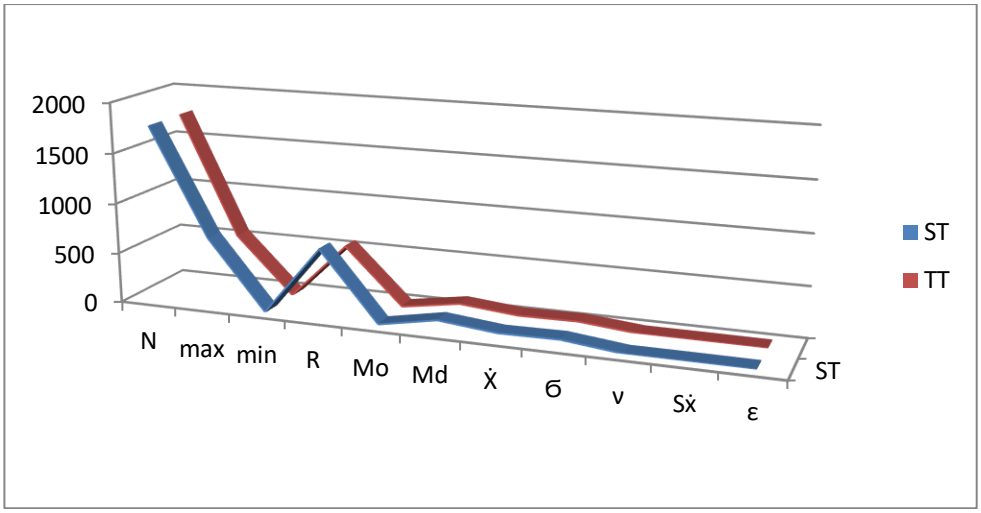

Figure 4: Interrelation of ST and TT coefficients in the paragraphs of the text

4.2.2. Distribution of a number of tokens (lexical) in the sentences of the text within text chapters

Numeric characteristics of data obtained

Below is the interpretation of data of the highest and lowest number within book chapters both in ST and TT, respectfully.

ST numeric characteristics.

ST highest number: § 01.02: Number of meanings (N) — 341; maximal meaning (max) — 72; minimal meaning (min) — 1; range (R) — 71; mode (Mo) — 9; median (Md) — 25,0; mean (Ẋ) — 14,51; standard deviation (б) — 10,88; coefficient of variation (v) — 0,7498; standard error (Sẋ) — 0,5893; measurement error (ε) — 0,0796.

ST lowest number: § 01.21: Number of meanings (N) — 89; maximal meaning (max) — 65; minimal meaning (min) — 4; range (R) — 61; mode (Mo) — 11; median (Md) — 20,5; mean (Ẋ) — 16,70; standard deviation (б) — 11,72; coefficient of variation (v) — 0,7019; standard error (Sẋ) — 1,2422; measurement error (ε) — 0,1458.

TT numeric characteristics.

TT highest number: § 01.15: Number of meanings (N) — 350; maximal meaning (max) — 45; minimal meaning (min) — 1; range (R) — 44; mode (Mo) — 6; median (Md) — 19,5; mean (Ẋ) — 12,43; standard deviation (б) — 8,63; coefficient of variation (v) — 0,6945; standard error (Sẋ) — 0,4612; measurement error (ε) — 0,0728.

TT lowest number: : § 01.35: Number of meanings (N) — 74; maximal meaning (max) — 48; minimal meaning (min) — 5; range (R) — 43; mode (Mo) — 9; median (Md) — 19,5; mean (Ẋ) — 17,81; standard deviation (б) — 9,35; coefficient of variation (v) — 0,5248; standard error (Sẋ) — 1,0866; measurement error (ε) — 0,1196.

4.2.3. Absolute and relevant distribution of the number of lexical tokens in the sentences of the text

The whole Source Text. 8 — 422 (6,32%); 9 — 419 (6,27%); 6 — 392 (5,87%); 7 — 379 (5,68%); 10 — 378 (5,66%); 11 — 346 (5,18%); 12 — 330 (4,94%); 5 — 305 (4,57%); 14 — 276 (4,13%); 13 — 263 (3,94%); 15 — 251 (3,76%); 16 — 239 (3,58%); 17 — 219 (3,28%); 18 — 185 (2,77%); 19 — 183 (2,74%); 4 — 177 (2,65%); 20 — 163 (2,44%); 21 — 150 (2,25%); 22 — 139 (2,08%); 23 — 126 (1,89%); 3 — 123 (1,84%); 24 — 110 (1,65%); 25 — 87 (1,30%); 27 — 83 (1,24%); 26 — 81 (1,21%); 28 — 71 (1,06%); 29 — 64 (0,96%); 2 — 63 (0,94%); 30 — 62 (0,93%); 32 — 46 (0,69%); 33 — 44 (0,66%); 34 — 42 (0,63%); 31 — 38 (0,57%); 36 — 33 (0,49%); 35 — 31 (0,46%); 1 — 29 (0,43%); 42 — 26 (0,39%); 37 — 24 (0,36%); 43 — 24 (0,36%); 39 — 23 (0,34%); 40 — 23

(0,34%); 41 — 22 (0,33%); 38 — 16 (0,24%); 45 — 16 (0,24%); 48 — 14 (0,21%); 47 — 13 (0,19%); 44 — 12 (0,18%); 49 — 11 (0,16%); 52 — 11 (0,16%); 54 — 10 (0,15%); 58 — 9 (0,13%); 46 — 7 (0,10%); 50 — 6 (0,09%); 51 — 5 (0,07%); 55 — 5 (0,07%); 56 — 5 (0,07%); 59 — 5 (0,07%); 53 — 4 (0,06%); 62 — 3 (0,04%); 63 — 3 (0,04%); 65 — 3 (0,04%); 57 — 2 (0,03%); 60 — 2 (0,03%); 66 — 2 (0,03%); 68 — 2 (0,03%); 70 — 2 (0,03%); 74 — 2 (0,03%); 78 — 2 (0,03%); 79 — 2 (0,03%); 61 — 1 (0,01%); 64 — 1 (0,01%); 67 — 1 (0,01%); 71 — 1 (0,01%); 72 — 1 (0,01%); 73 — 1 (0,01%); 77 — 1 (0,01%); 80 — 1 (0,01%); 92 — 1 (0,01%); 101 — 1 (0,01%); 121 — 1 (0,01%); 148 — 1 (0,01%); 166 — 1 (0,01%);.

The whole Target Text. 7 — 521 (7,82%); 5 — 460 (6,90%); 8 — 453 (6,80%); 6 — 452 (6,78%); 9 — 407 (6,11%); 10 — 382 (5,73%); 11 — 343 (5,15%); 13 — 336 (5,04%); 12 — 331 (4,97%); 4 — 326 (4,89%); 16 — 231 (3,47%); 15 — 228 (3,42%); 3 — 222 (3,33%); 14 — 220 (3,30%); 17 — 215 (3,23%); 18 — 156 (2,34%); 19 — 146 (2,19%); 21 — 137 (2,06%); 20 — 134 (2,01%); 2 — 101 (1,52%); 22 — 100 (1,50%); 23 — 87 (1,31%); 24 — 72 (1,08%); 25 — 62 (0,93%); 1 — 59 (0,89%); 28 — 51 (0,77%); 30 — 51 (0,77%); 26 — 49 (0,74%); 29 — 43 (0,65%); 27 — 39 (0,59%); 31 — 34 (0,51%); 33 — 23 (0,35%); 32 — 21 (0,32%); 36 — 19 (0,29%); 34 — 17 (0,26%); 35 — 15 (0,23%); 37 — 15 (0,23%); 39 — 15 (0,23%); 38 — 11 (0,17%); 42 — 9 (0,14%); 44 — 9 (0,14%); 40 — 8 (0,12%); 41 — 8 (0,12%); 43 — 6 (0,09%); 45 — 6 (0,09%); 46 — 5 (0,08%); 48 — 5 (0,08%); 54 — 4 (0,06%); 47 — 2 (0,03%); 49 — 2 (0,03%); 51 — 2 (0,03%); 52 — 2 (0,03%); 55 — 2 (0,03%); 50 — 1 (0,02%); 53 — 1 (0,02%); 56 — 1 (0,02%); 58 — 1 (0,02%); 60 — 1 (0,02%); 65 — 1 (0,02%); 66 — 1 (0,02%); 72 — 1 (0,02%); 73 — 1 (0,02%); 83 — 1 (0,02%); 87 — 1 (0,02%); 134 — 1 (0,02%).

Description: distribution data on the number of lexical tokens in the sentences of the text (within text chapters) shows the whole picture of statistics. To mention here, for example, three initial figures of whole ST: 8 — 422 (6,32%); 9 — 419 (6,27%); 6 — 392 (5,87%) and TT data: 7 — 521 (7,82%); 5 — 460 (6,90%); 8 — 453 (6,80%) we now see the tendency of the number of tokens in the sentences within the original, as expected, also exceeds the number of tokens in the sentences of the translation text. To specify, the translated variant eventually lacks the figures (8 — 422 (6,32%); 9 — 419 (6,27%) of the original which makes it possible to further discuss the loss and distortion of the author's style from translatological perspective (to be researched in the forthcoming article).

Top-30 frequencies of lexical tokens in both ST and TT sentences were selected (Fig. 5). It is seen, that the lexical tokens distribution curves are corresponding but slightly shifted along the horizontal axis. The max number of lexical tokens in ST sentences is 422 (for 8 lexical tokens), while the appropriate number in TT is 521 (for 7 lexical tokens).

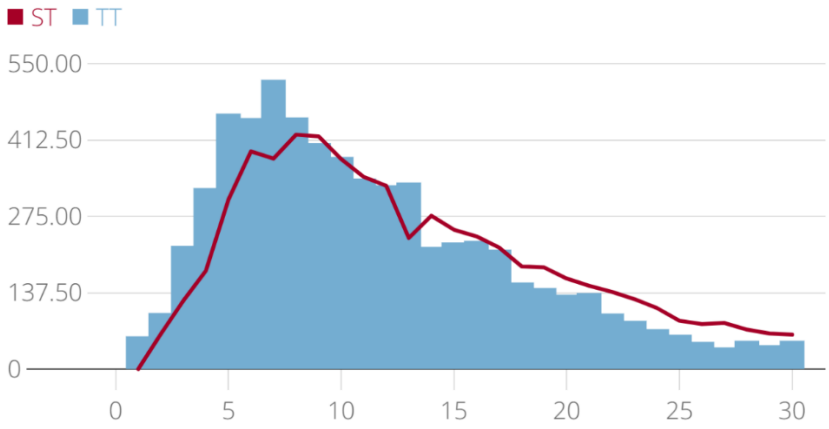

Figure 5: The correlation between number of (lexical) tokens in ST and TT sentences (the horizontal axis corresponds to the number of lexical tokens, and the vertical axis corresponds to the number of sentences).

All in all, the findings have proven that most frequent token usage within the sentence in ST equals to 8 which is 6, 32 % as related to the whole text. It occurs in 422 sentences. While in TT most

frequent word usage within the sentence sums up to 7 making up 7,82%; this feature is witnessed in 521 cases.

Numeric characteristics of token usage within the sentence

Numeric characteristics of corpus in ST. Number of meanings (N) — 6678; maximal meaning (max) — 166; minimal meaning (min) — 1; range (R) — 165; mode (Mo) — 8; median (Md) — 41,5; mean (Ẋ) — 15,38; standard deviation (б) — 11,00; coefficient of variation (ν) — 0,7147; standard error (Sẋ) — 0,1346; measurement error (ε) — 0,0171.

Numeric characteristics of corpus in TT. Number of meanings (N) — 6666; maximal meaning (max) — 134; minimal meaning (min) — 1; range (R) — 133; mode (Mo) — 7; median (Md) — 33,0; mean (Ẋ) — 12,44; standard deviation (б) — 8,46; coefficient of variation (ν) — 0,6797; standard error (Sẋ) — 0,1036; measurement error (ε) — 0,0163.

ST and TT data contrast. The Table 5 below summarizes and contrasts the findings of numeric characteristics of parallel corpus concerning the distribution of lexical tokens in the sentences of the text.

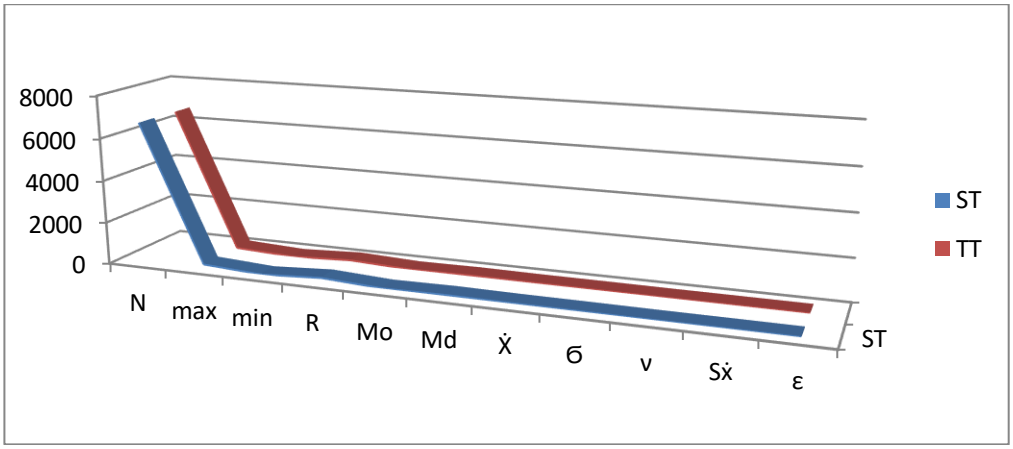

Figure 6: Interconnection of coefficients in ST and TT in the sentences of the text

Table 5

Contrast of ST and TT the findings on lexical tokens in the sentences of the text

Unit	ST	TT
N	6678	6666
max	166	134
min	1	1
R	165	133
Mo	8	7
Md	41,5	33,0
Ẋ	15,38	12,44
б	11	8,46
ν	0,6797	0,6797
Sẋ	0,1036	0,1036
ε	0,0163	0,01636

4.2.4. Distribution of the number of lexical tokens in chapters of the book

Interpretation of data obtained on absolute and relevant distribution of lexical tokens in book chapters

Whole ST. 1203 — 1 (2,70%); 1486 — 1 (2,70%); 1539 — 1 (2,70%); 1626 — 1 (2,70%); 1776 — 1 (2,70%); 1840 — 1 (2,70%); 1928 — 1 (2,70%); 1930 — 1 (2,70%); 2009 — 1 (2,70%); 2026 — 1 (2,70%); 2120 — 1 (2,70%); 2237 — 1 (2,70%); 2272 — 1 (2,70%); 2330 — 1 (2,70%);

2348 — 1 (2,70%); 2395 — 1 (2,70%); 2476 — 1 (2,70%); 2517 — 1 (2,70%); 2532 — 1 (2,70%); 2663 — 1 (2,70%); 2703 — 1 (2,70%); 2851 — 1 (2,70%); 2905 — 1 (2,70%); 2927 — 1 (2,70%); 2928 — 1 (2,70%); 3048 — 1 (2,70%); 3207 — 1 (2,70%); 3333 — 1 (2,70%); 3440 — 1 (2,70%); 3491 — 1 (2,70%); 3652 — 1 (2,70%); 4003 — 1 (2,70%); 4025 — 1 (2,70%); 4305 — 1 (2,70%); 4449 — 1 (2,70%); 4949 — 1 (2,70%); 5272 — 1 (2,70%);.

Whole TT. 898 — 1 (2,63%); 1145 — 1 (2,63%); 1188 — 1 (2,63%); 1318 — 1 (2,63%); 1459 — 1 (2,63%); 1478 — 1 (2,63%); 1487 — 1 (2,63%); 1499 — 1 (2,63%); 1703 — 1 (2,63%); 1731 — 1 (2,63%); 1742 — 1 (2,63%); 1747 — 1 (2,63%); 1836 — 1 (2,63%); 1844 — 1 (2,63%); 1859 — 1 (2,63%); 1870 — 1 (2,63%); 1893 — 1 (2,63%); 1928 — 1 (2,63%); 1962 — 1 (2,63%); 2010 — 1 (2,63%); 2030 — 1 (2,63%); 2114 — 1 (2,63%); 2151 — 1 (2,63%); 2220 — 1 (2,63%); 2278 — 1 (2,63%); 2315 — 1 (2,63%); 2388 — 1 (2,63%); 2478 — 1 (2,63%); 2600 — 1 (2,63%); 2717 — 1 (2,63%); 2782 — 1 (2,63%); 2810 — 1 (2,63%); 3050 — 1 (2,63%); 3241 — 1 (2,63%); 3320 — 1 (2,63%); 3345 — 1 (2,63%); 4157 — 1 (2,63%); 4349 — 1 (2,63%).

Description: depicted above results on absolute and relevant distribution of lexical tokens in book chapters prove that the original data exceeds the number of tokens in correlation to its translation, compare: 2,70% in the ST and 2,63% in the TT.

The number of lexical tokens in both ST and TT chapters are at Fig. 7.

Numeric characteristics of token usage in the chapters of the book

Numeric characteristics of corpus in ST. Number of meanings (N) — 37; maximal meaning (max) — 5272; minimal meaning (min) — 1203; range (R) — 4069; mode (Mo) — 1203; median (Md) — 2532,0; mean ($\dot{X}$) — 2776,78; standard deviation (б) — 968,44; coefficient of variation (ν) — 0,3488; standard error ($S\dot{x}$) — 159,2108; measurement error (ε) — 0,1124.

Numeric characteristics of corpus in TT. Number of meanings (N) — 38; maximal meaning (max) — 4349; minimal meaning (min) — 898; range (R) — 3451; mode (Mo) — 898; median (Md) — 1986,0; mean ($\dot{X}$) — 2182,68; standard deviation (б) — 768,20; coefficient of variation (ν) — 0,3520; standard error ($S\dot{x}$) — 124,6190; measurement error (ε) — 0,1119.

ST and TT data contrast. The Table 6 below summarizes and contrasts the findings of numeric characteristics of parallel corpus concerning the distribution of lexical tokens in the chapters of the book.

Table 6

Contrast of ST and TT the findings on lexical tokens in the chapters of the book

Unit	ST	TT
N	37	38
max	5272	4349
min	1203	898
R	4069	3451
Mo	1203	898
Md	2532,0	1986,0
$\dot{X}$	2776,78	2182,68
б	968,44	768,20
ν	0,3488	0,3520
$S\dot{x}$	159,2108	124,6190
ε	0,1124	0,1119

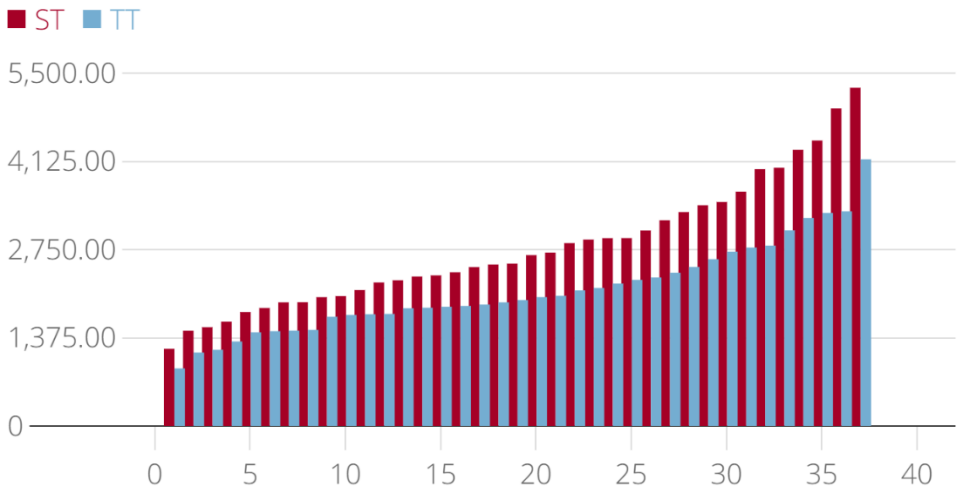

Figure 7: The number of lexical tokens in ST and TT chapters (the horizontal axis relates to the number of chapters, and the vertical axis corresponds to the number of lexical tokens)

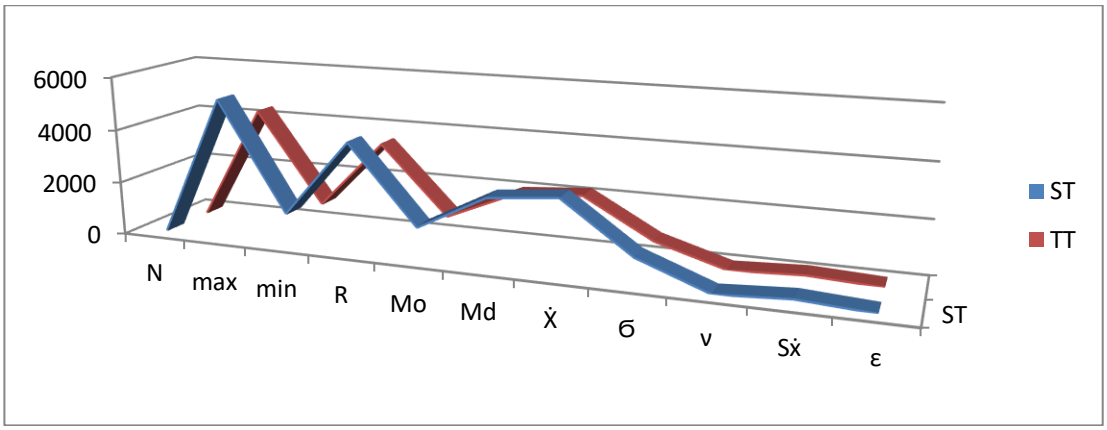

Figure 8: Interdependence of coefficients in ST and TT in the chapters of the book

5. Conclusions

Presented in the paper are the results of absolute and relevant distribution, probability measurement, as well as: N, max, min, R, Mo, Md, Ẋ, б, v, Sẋ, ε in the sentence, paragraphs and chapters of both texts: the original and the translation under analysis.

At this stage, the study is of practical and applied value. We are convinced that further elaborations in this realm of statistically oriented translation studies will prove promising and result in formulating and supporting scientific theoretical hypothesis. Thus, the prospect of the study is to further explore Montgomery's individual style and its rendition into the target language. Our next step is part of speech tagging.

6. Acknowledgements

The project has been carried out within the complex academic topic "Application of modern technologies for optimization of information processes in natural language" at Lviv Polytechnic National University. At the initial stage the project underwent the consultancy of Ihor Kulchytskyy, to whom we express our gratitude.

7. References

[1] 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019): Proceedings of the Workshop, Varna, Bulgaria, September 5-6, 2019.

[2] L. Akhtyamova, M. Alexandrov, J. Cardiff, O. Koshulko, Opinion Mining on Small and Noisy Samples of Health-Related Texts. In: Shakhovska N., Medykovskyy M. (eds) Advances inIntelligent Systems and Computing III. CSIT 2018. Advances in IntelligentSystems and Computing, vol 871. Springer, Cham. (2019) https://doi.org/10.1007/978-3-030-01069-0_27.

[3] I. Bekhta, U. Tykha, Ludic Linguistic Challenges in the Transtextual Dimensions of David Lodge's Deaf Sentence. Journal of Narrative and Language Studies, 8(14), 2020, pp. 50–63. URL: http://nalans.com/nalans/article/view/229

[4] K.H. Chen, H.H. Chen, Aligning bilingual corpora especially for language pairs from different families. Information Sciences Applications, 42, 1995, pp. 57–81

[5] Corpus-based Language Studies: An Advanced Resource Book, (Eds.) T. McEnery, R. Xiao, Y. Tono, Routledge, 2006

[6] Corpus-based Translation Studies: Theory, Findings, Applications, ed. S. Laviosa, Rodopy, 2002

[7] N. S. Dash, S. Arulmozi, History, Features, and Typology of Language Corpora, Springer: Springer Nature Singapore, 2018.

[8] M. Dilai, O. Levchenko, Discourses, Surrounding Feminism in Ukraine: A Sentiment Analysis of Twitter Data, in: Proceedings 2018 IEEE 13thInternational Scientific and Technical Conference on Computer Sciences andInformation Technologies, CSIT 2018, 2, art. no. 8526694, pp. 47–50 (2018).

[9] N. Hrytsiv, I. Kulchytskyy, O. Rohach, Quantitative Comparative Analysis in Parallel Translation Corpus: building author's and translator's statistical profiles: (a case study of Lucy Maud Montgomery), in: Proceedings 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), pp. 255–258 (2020) doi: 10.1109/CSIT49958.2020.9321893

[10] S. Hunston, Corpora in Applied Linguistics, Cambridge University Press, Cambridge, 2002.

[11] O. Flys, Woman Lingual Cultural Type Analysis Using Cognitive Modeling and Graph Theory. In: Shakhovska N., Medykovskyy M.O. (Eds.) Advances in Intelligent Systems and Computing IV. CSIT 2019. Advances inIntelligent Systems and Computing, vol 1080. Springer, Cham. (2020)

[12] D. Kenny, Lexis and Creativity in Translation. A corpus-based study. Routledge, 2001.

[13] O. Levchenko, O. Tyshchenko, M. Dilai, L. Gajarsky, A Model of the Information System of the Associative Verbal Network Presentation. In: Shakhovska N., Medykovskyy M. (Eds.) Advances in Intelligent Systems and Computing V. CSIT 2020. Advances in Intelligent Systems and Computing, vol1293. Springer, Cham, 2021.

[14] O. Lozynska, V. Savchuk, V. Pasichnyk, Individual Sign Translator Component of Tourist Information System. In: Shakhovska N., Medykovskyy M.O. (Eds.) Advances in Intelligent Systems and Computing IV. CSIT2019. Advances in Intelligent Systems and Computing, vol. 1080. Springer, Cham, 2020.

[15] V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko, O. Kanishcheva (Eds.), in: Proceedings Computational Linguistics and Intelligent Systems. 4th Int. Conf. COLINS 2020. Volume I: Workshop. Lviv, Ukraine, April 23-24, 2020, CEUR-WS.org, online.

[16] L. Montgomeri, Enn iz Zeleny`x daxiv. 2-he vy`d. Per. z angl. Vovchenko A., Urbino, L`viv, 2019.

[17] L. Montgomery, Ann of Green Gables, Film & TV Tie-in Ed edition, Puffin, 1988.

[18] G. Szymanski, P. Lipinski, Model of the effectiveness of Google Adwords advertising activities, in: Proceedings 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018, 2, art. no. 8526633, 2019, pp. 98–101.

[19] F. Zanettin, Parallel corpora in translation studies: Issues in corpus design and analysis. In Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects (Eds.) M. Olohan, St. Jerome, Manchester, 2000, pp. 105–118.

[20] O. Zuban, Lexicographical Database of Frequency Dictionaries of Morphemes Developed on the Basis of the Corpus of Ukrainian Language. In: N. Shakhovska, M. Medykovskyy (Eds.), Advances in Intelligent Systems and Computing IV. CSIT 2019. Advances in Intelligent Systems and Computing, vol. 1080. Springer, Cham, 2020, doi: 10.1007/978-3-030-33695-0_37