Quantitative Parameters of J. London's Short Stories Collection "Children of the Frost" and its Translation

Mariia Bekhta-Hamanchuk[1], Halyna Oleksiv[2], Tetiana Shestakevych[3] and Yuliia Shyika[4]

[1,2,3,4] *Lviv Polytechnic National University, Stepana Bandery Street, 12, Lviv, 79000, Ukraine*

Abstract

The paper presents the quantitative comparative analysis of Jack London's collection of short stories "Children of the Frost" and Ukrainian translations by V. Hladka and K. Koriakina which has been carried out on the basis of the digital marked corpus of original texts. The novelty of the research lies in the fact that the above-mentioned literary work has not been previously studied from the statistical perspective. The theoretical background of the study is outlined, particularly emphasizing the issues of the corpus, corpus annotation and corpus linguistics software. The source and target texts have been compared according to the following coefficients: text volume, number of different word forms, number of sentences, number of letters, number of content words, number of functional words, hapax legomena and number of words with a frequency of 10 or more. The most frequently used parts of speech both in source and target texts are stated. The quantitative indices of the lexical level, which have been calculated on the basis of the general characteristics of the source and target texts, have been compared. The reproduction of the nominal character of the source text in the target text has been analyzed.

Keywords[1]

Corpus, corpus annotation, corpus linguistics, source text, target text.

1. Introduction

One of the key issues in modern linguistics is natural language processing. Working with large amounts of factual information enables the researcher to avoid subjective selection of facts for confirming or rejecting the hypothesis. Nowadays there is a number of information technologies enabling an automated search with the aim of forming the factual basis of the research, corpora of texts being one of them. The corpus of texts is a central concept in corpus linguistics and its object of study. The issues of corpus linguistics are widely ranged and involve studies of the general theory of corpus linguistics, correlations of corpus linguistics and other linguistic disciplines, corpus typologies and methods of corpus data interpretation, the principles of creating natural languages text corpora (D. Biber [3; 3], J. Sinclair [28], W. Teubert [30], G. Kennedy [16], G. Leech [14; 20], A. Stefanowitsch [29], T. McEnery [10; 14], D. Barth, S. Stefan [2], N.S. Dash, S. Arulmozi [11], G. Desagulier [12], M. Paquot, S. Th. Gries [25], V. Shyrokov [9], O. Demska-Kulchytska [13], A. Baranov [1] etc). Since a language is not a strictly arranged system and has probabilistic and stochastic character, it is advisable to apply statistical methods in order to research it [17]. Research in corpus linguistics is facilitated by special software tools – concordancers and corpus managers – which provide various opportunities to obtain the necessary information from the corpus. Thus, corpora allow addressing the variety of research questions and have been applied in a wide range of linguistic disciplines, including lexicography, grammar, discourse analysis, sociolinguistics, language

[1] COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland.
EMAIL: mariia.p.bekhta@lpnu.ua (M. Bekhta-Hamanchuk); halyna.d.oleksiv@lpnu.ua (H. Oleksiv); tetiana.v.shestakevych@lpnu.ua (T. Shestakevych); yuliia.i.shyika@lpnu.ua (Ju. Shyika). ORCID: 0000-0002-3133-0948 (M. Bekhta-Hamachuk); 0000-0002-8800-6217 (H. Oleksiv); 0000-0002-4898-6927 (T. Shestakevych); 0000-0003-2474-0479 (Ju. Shyika).

teaching, literary studies, translation studies, pragmatics, cognitive linguistics, conceptual studies, etc [26, p.473].

2. Theoretical background

Biber et al. define a corpus as “a large and principled collection of natural texts” [3, p.4]. Generally understood as the collection of texts, the term *corpus* can have different meanings in various disciplines. In fiction studies, it is a collection of particular author’s works. In the field of linguistics, the corpus refers to any collection of data (whether narrative texts or separate sentences) obtained for the purpose of linguistic research, often taking into account a specific research goal [27, p.769; 29, p.22]. But the term is used in a different way in corpus linguistics – “it refers to a collection of samples of language use with the following properties:

- the instances of language use contained in it are authentic;
- the collection is representative of the language or language variety under investigation;
- the collection is large” [27, p.769; 29, p.22].

Additionally, texts in such collections are often commented on in order to enhance their potential for linguistic analysis. In particular, they may contain information about the paralinguistic aspects of the source data (intonation, font style, etc.), linguistic properties of utterances (parts of speech, syntactic structure) and demographic information about speakers / writers [29; p.22]. The volume and content of the corpus may change, but these changes must neither influence its representativeness nor change it reasonably. Search in the data corpus allows a researcher to build a concordance for any word, i.e. to build a list of all usages of the word in the context and with the references to the source. Corpora can be used to obtain a variety of data and statistics on language and language units.

As a rule, the research process within a corpus involves the following stages:

1. Selection of sources of linguistic material.
2. Data entry. Texts in electronic form with the extension *.txt* were included in the corpus.
3. Philological verification and texts editings.
4. Converting and graphematic analysis which includes recoding of nontextual elements or their removal and division of the text into structural parts.
5. Providing texts and their components with additional information, i.e. text markup.
6. Converting of marked texts into the corpus and providing access to it.

To serve as a basis of the scientific research, a corpus should not only have a significant volume or contain data of various types but also it should possess the following features:

- Representativeness. The corpus must represent all the features of a particular area. It can be very large (national corpus) or very small (author corpus). T. McEnery argues that the representativeness of the corpora is caused by two factors: the set of genres that are in the corpus and the selection of texts [10]. Selection is characterized by the limit of real material, selecting certain parts of speech from the language array. However, the largest language corpus can display only a small part of oral and written texts. Representativeness is closely related to volume of the corpus. However, volume of the corpus is determined by two factors: representativeness (sufficiency of texts (words) for accurate representation of the language material) and practicality (accessibility and labour-intensiveness). For example, it is necessary to cover all works of a certain author, or historical texts of a certain period, or texts of a certain subject (for example, radio or TV series, political speeches). In other cases, full representation of language cannot be achieved.
- Balance. Corpus representativeness largely depends on how balanced a corpus is. The acceptable balance of a corpus is determined by its intended uses. A balanced corpus usually covers a wide range of text categories which are supposed to be representative of the language or language variety under consideration. [10] Although balance is indispensable in corpus design, there is no scientific method of measuring it. Nonetheless, text typology is of high relevance if one attempts at corpus balance. To achieve balance, a corpus requires certain characteristics of text selection, which include differences between the book and newspaper, different genres of literature and authors.

- Machine readability is the main criterion for electronic text corpus. Machine readability also requires encoding of corpus data. Corpus computerization has many advantages. It speeds up processing and makes working with data sets much easier. After computer processing of data, the objective and accurate results are obtained. Machine readability enables further automatic processing of data of a particular corpus, and allows the researcher to improve the corpus with all sorts of markup. It is the use of computerized corpora, together with computer programs which facilitate linguistic analysis, that distinguishes modern machine-readable corpora from early corpora [10].

The purpose of the language corpus is to show the functioning of linguistic units in their natural contextual environment. The following prerequisites form the basis for further creation and usage of corpora:

1. substantial (representative) and balanced volume of the corpus guarantees the typicality of the data and provides the whole spectrum of linguistic phenomena;
2. various data, which are included in the corpus, are in their natural contextual form, which creates the possibility of their comprehensive and objective study;
3. once created and prepared data set can be used repeatedly, by different researchers and for different purposes.

In the process of creating a corpus, the certain procedures should be followed, regardless of whether the corpus includes spoken or written language material. Some of the issues that are optimal in building the corpus include: typology of texts, file names and their format, etc. The next important step in building a corpus is marking and annotation. Document markup refers to labeling, similar to HTML code used to indicate features of a document: paragraphs, fonts, sentences, including sentence numbers, author identification, and end-of-text markings. At the basic level, the title can be considered as a type of markup as it provides additional information about the text.

Apart from *corpus*, another key term in corpus linguistics is *corpus annotation,* which is defined by G. Leech as the process of "adding interpretative, linguistic information to an electronic corpus of spoken and / or written language data [20]. The main issue in corpus linguistics is the creating of means of automatic / automated text annotation based on different criteria – morphological, orthoepic, semantic, syntactic, etc. V Shyrokov states that automated division of an electronic literary text into 'microcontexts' is the main idea of linguistic corpus engineering, with microcontexts being text fragments grouped around the object under interpretation [9; p.99].

Corpus annotation can take many forms that can be implemented at different levels:

1. at the phonological level: the corpora can be commented on the constituent boundaries (phonetic / phonemic annotation) or prosodic features (prosodic annotation);
2. at the morphological level: the corpora can be annotated as prefixes, stems and suffixes (morphological annotation);
3. at the lexical level: the corpora can be annotated by parts of speech, lemmas (lemmatization) and semantic fields (semantic annotation);
4. at the syntactic level: the corpora can be annotated to reflect anaphoric connections, pragmatic information such as language acts (pragmatic annotation), or stylistic features such as speech and thought representation (stylistic annotation).

The most common form of corpus annotation includes tags of the parts of speech (POS tagging or grammatical tagging), which mark each word in the corpus as a grammatical category (e.g. noun, adjective, adverb etc.). When corpora began to be annotated, the levels of annotation applied were simple. However, as the tools evolved, more levels of linguistic knowledge started to be incorporated into the texts and corpora [15, p.47]. These tags facilitate settling a number of issues about a simple search for a particular keyword. Many words are ambiguous, but when a word is marked with a part of speech, it eliminates ambiguity and helps focus the search results clearly. Therefore, annotated corpora can be widely applied. Many linguistic analyses depend heavily on POS tagging [10].

To sum up, annotation aims at the addition of extralinguistic, structural, and linguistic special markers to texts. Different types of linguistic markup are distinguished: morphological, syntactic, semantic, anaphoric and prosodic. Also, the following procedures are carried out: tokenization, lemmatization, stemming and parsing. Most corpora belong to the morphological or syntactic type. It should be noted that the latter explicitly or implicitly contain morphological characteristics of lexical units.

Since corpus linguistics uses large and representative samples of natural language texts for the research, there are several types of software that can be used in the study. They are: concordancers (LEXA, MonoConc, MicroConcord, TACT, WordSmith, WordCruncher, Manatee (Bonito), IMS Corpus Workbench (CQP), XAIRA, LEXA, Virtual Corpus Manager (VMC), EXMARaLDA Corpus–Manager (Co–Ma)) and a specific software for comprehensive analysis.

Concordancers are used to make lists of examples (occurrences) of the required token (tokens, lemmas, morphemes etc) in the minimum context. Usually such a context is a fragment of several linguistic units on the left (L) and on the right (R).

The corpus manager refers to the system for managing textual and linguistic data. It is a special search system that uses software to search for data in the corpus, obtain statistical information and provide results to the user in a convenient form. The results of this procedure are presented in the form of horizontal lines with a search word in the middle. This process is called KWIC (Key Word In Context) [18].

Corpus analysis software tools vary in functionalities, but all of them facilitate to search the corpus for a specific set of linguistic units. Most of these software packages have the following features:

1. they create KWIC (keyword in context) concordants, i.e. they display the query in their immediate context, defined in terms of a certain number of words or symbols on the left and right;
2. they identify the collocations of this expression, i.e. the forms of words that occur in a particular position in relation to another word; these words are usually listed in the order in which they occur in the appropriate position;
3. they form lists of frequencies, i.e. lists of all lines of symbols in the corpus, listed in the order of their frequency.

Generally, modern software tools used in corpus linguistics research are fast and rich in features. On the other hand, most of the tools are English-centric in that they only allow access to English corpora. In addition, they all offer a different user-experience, because each tool is created in isolation and thus offers a different user interface, control flow, and functionality [19, p.154]. Nevertheless, corpus software tools are indispensable in corpus-based research projects.

3. Results and discussion

Text corpus, being the main issue of corpus linguistics, is widely applicable in translation studies. This study focuses on the contrastive analysis of the quantitative parameters of the source (English) text and its translation (Ukrainian). Jack London's short stories collection “Children of the Frost” is in the centre of attention. The choice has been made due to the fact that the literary work in question has not been studied from statistical perspective before. In the process of analysis quantitative and qualitative analytical methods have been used.

In this research the analysis of Jack London's collection of short stories “Children of the Frost” has been conducted on the basis of the digitally processed and marked up corpus of original texts and Ukrainian translations by V. Hladka and K. Koriakina [22; 23]. It covers a number of characteristics which are compared in Tables 1-4. Here and after, we propose some denotations, the text volume is N, the number of different word forms is V, the number of sentences is S, the number of letters is C, the number of content words is C_1, the number of functional words is F_1, the number of Hapax legomena is V_1, the number of words in the text with a frequency of 10 or more is N_{10}.

Table 1
Quantitative parameters of source and target texts

Coefficient	Source text	Target text	Ratio
N	45678	32192	1,42
V	11790	16263	0,72
S	3185	3527	0,90
C	210423	199852	1,05

C_1	31418	28755	1,09
F_1	14260	8699	1,64
V_1	6453	7853	0,82
N_{10}	725	575	1,26

The visualization (Fig. 1) of the data from the Table 1 is performed to show the ratio between the quantitative characteristics of the Source and Target texts. Here each quantitative characteristics of the Source text has been divided by the appropriate number that characterizes the Target text. When the result of such division is above 1, it means the appropriate characteristic of the Source text exceeds the Target text.

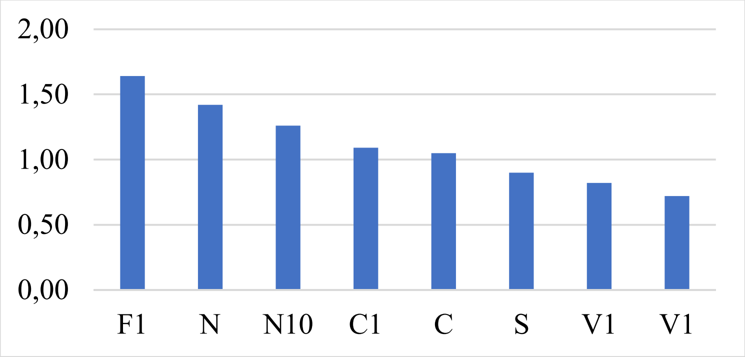

Figure 1. The ratio of Source text and Target text characteristics

As is seen from the Figure 1, in the process of translation, the number of functional words decreased, as well as text volume and Number of words in the text with a frequency of 10 or more. The number of different word forms is higher in the Target text, which is predictable at least because the Ukrainian language has seven cases, as opposed to two cases in English.

Table 2

Quantitative parameters of original stories

Coefficient	In the Fore-sts of the North	The Law of Life	Nam-bok the Unve-racious	The Master of Mystery	The Sun-lan-ders	The Sick-ness of Lon Chief	Keesh, Son of Keesh	The Death of Ligoun	Li Wan, the Fair	The League of the Old Men
N	5970	2836	4500	4085	6368	3632	3135	3610	5249	6293
V	1485	916	1059	1275	1369	906	898	903	1472	1507
S	477	193	379	295	463	180	254	186	413	345
C	28372	11673	22882	17933	32653	14505	15675	14421	26397	25912
C_1	4200	1992	3049	2968	4202	2533	2121	2423	3548	4382
F_1	1770	844	1451	1117	2166	1099	1014	1187	1701	1911
V_1	890	444	654	678	806	372	557	372	981	699
N_{10}	95	41	78	70	102	66	51	54	74	94

Table 3

Quantitative parameters of translated stories

Coefficient	In the Fore-sts of the North	The Law of Life	Nam-bok the Unve-racious	The Master of Mystery	The Sun-lan-ders	The Sick-ness of Lon Chief	Keesh, Son of Keesh	The Death of Ligoun	Li Wan, the Fair	The League of the Old Men
N	5512	2155	3271	3487	4627	2950	2221	2713	5264	5256

V	2487	1175	1212	1655	1632	1323	995	1295	2297	2192
S	529	214	423	324	526	187	277	202	461	384
C	27138	10499	21960	17132	31400	13580	14694	13189	25427	24833
C_1	4382	1705	2311	2736	3287	2310	1642	2105	4159	4118
F_1	1130	450	960	751	1340	640	579	608	1103	1138
V_1	1022	524	861	717	1107	516	715	529	973	889
N_{10}	85	33	50	58	61	55	29	46	75	83

Table 4

Ratio of characteristics of the source text and target text

Coefficient	In the Forests of the North	The Law of Life	Nambok the Unveracious	The Master of Mystery	The Sunlanders	The Sickness of Lon Chief	Keesh, Son of Keesh	The Death of Ligoun	Li Wan, the Fair	The League of the Old Men
N	1,1	1,3	1,4	1,2	1,4	1,2	1,4	1,3	1,0	1,2
V	0,6	0,8	0,9	0,8	0,8	0,7	0,9	0,7	0,6	0,7
S	0,9	0,9	0,9	0,9	0,9	1,0	0,9	0,9	0,9	0,9
C	1,0	1,2	1,3	1,1	1,3	1,1	1,3	1,2	0,9	1,1
C1	1,6	1,9	1,5	1,5	1,6	1,7	1,8	2,0	1,5	1,7
F1	0,9	0,8	0,8	0,9	0,7	0,7	0,8	0,7	1,0	0,8
V1	1,1	1,2	1,6	1,2	1,7	1,2	1,8	1,2	1,0	1,1

The analysis of general characteristics has shown that the number of word usages in the source text exceeds the number of word usages in the target text both in the whole corpus and in separate stories. Altogether, the volume of the source text is 20.69% larger than the volume of the target text. It should be noted that this contradicts the theory of translation S-universals and T-universals, which was put forward by A. Chesterman [7], and involves an increase in the volume of the target text compared to the source text.

The visualization (Fig. 2) of the data from the Table 2 and 3 is performed to show the ratio between the quantitative characteristics of each Source and Target texts. Here each quantitative characteristics of the Source text has been divided by the appropriate number that characterizes the Target text. When the result of such division is above 1, it means the appropriate characteristic of the Source text exceeds the Target text.

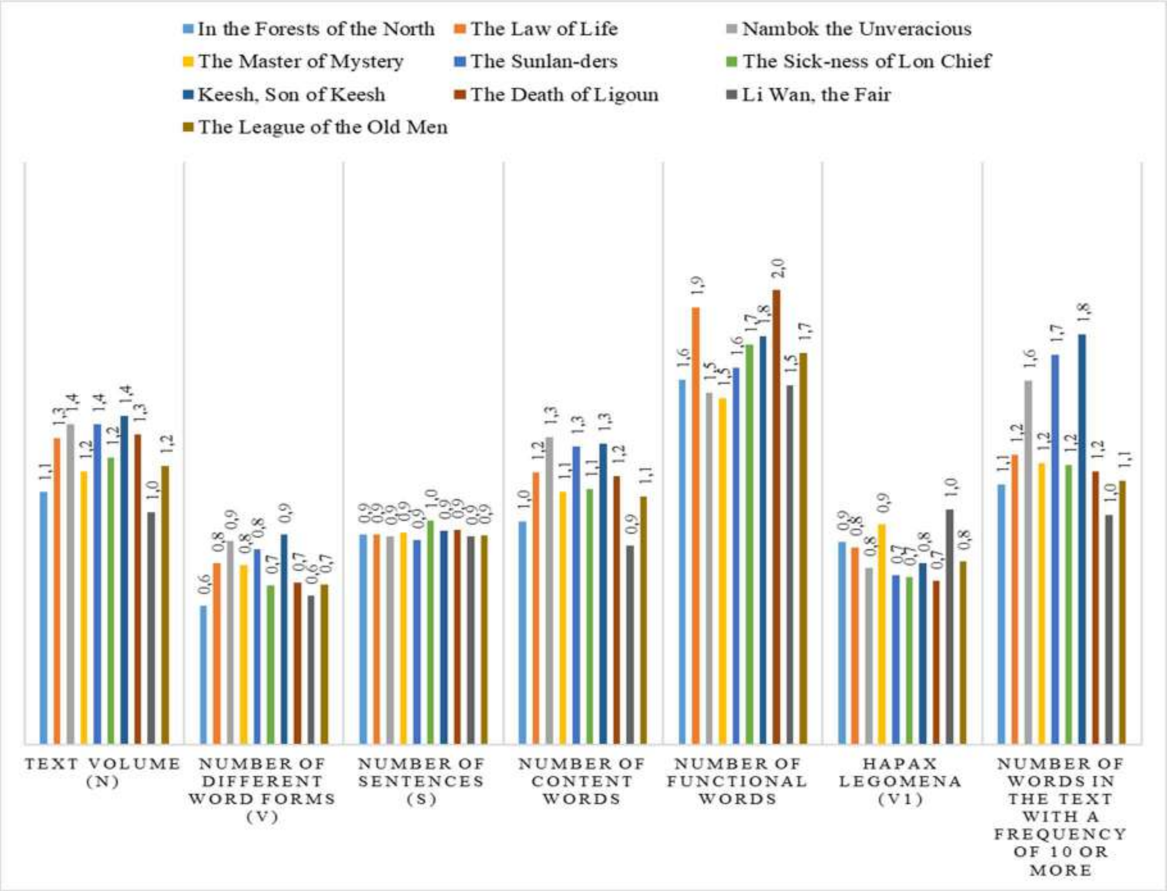

Figure 2. The ratio of each Source text and Target text characteristics

S-universals are observed when comparing the source text with a number of translations in some target language. T-universals appeared as a result of comparing the corpora of target texts and course texts. The following S-universals can be distinguished: increasing the volume of the translated text compared to the original; simplification at the syntactic level; simplification at the lexical level – reduction of lexical diversity and the tendency to use more frequent words in the target language; reduction or avoidance of recurrences in the target language; avoiding the ethnospecific units in translation; standardization (use of typical target language structures); convergence (translated texts show greater linguistic similarities with each other than with the original texts).

As for T-universals, their taxonomy includes:

- simplification (reduction of lexical diversity and density);
- conventionalization (standardization);
- atypical (unstable) lexical patterns [7].

The frequency of each part of speech in the text and the vocabulary of the author (translators) has been compared since the ratio of parts of speech is an important statistical parameter of the individual style of both the author and a particular work (Table 5).

The most frequent in the source and target texts are the functional words (5% of the vocabulary in the source text and 6.37% in the target text). These words function most actively and cover almost a quarter (29.81% in the original text and 23.31% in the translated text) of the text. Pronouns have similar high activity in the text (3.18% of the vocabulary in the source text and 3.24% in the target text). Pronouns cover about 13% of the text. Approximately the same share in the text and the vocabulary is covered by adverbs (7.20% and 8.91% in the source text and 10.13% and 12.16% in the target text) and numerals (0.91% and 1.07 in source text and 1.26% and 1.06% in the target text) (see Table 6). In Figure 3, each quantitative characteristics of the Source text was divided by the appropriate number that characterizes the Target text, as it is calculated in Table 7. When the result of such division is above 1, it means the appropriate characteristic of the Source text exceeds the Target text.

Table 5
Part of speech frequency in the source text

Part of speech	In the Forests of the North	The Law of Life	Nambok the Unveracious	The Master of Mystery	The Sunlanders	The Sickness of Lon Chief	Keesh, Son of Keesh	The Death of Ligoun	Li Wan, the Fair	The League of the Old Men
Noun	1533	570	1070	897	1494	760	767	754	1243	1369
Adjective	556	243	293	325	404	336	258	274	423	557
Pronoun	602	331	524	508	597	476	334	489	574	793
Adverb	565	221	408	352	646	257	255	181	424	457
Verb	887	592	731	844	985	676	478	664	863	1148
Numeral	57	35	23	42	76	28	29	61	21	58
Preposition	819	317	618	471	877	420	454	474	748	734
Conjunction	429	180	325	282	511	326	202	378	414	575
Particle	-	-	-	-	-	-	-	-	-	-
Interjection	19	11	16	17	9	32	24	16	10	14
Article	503	336	492	347	769	321	334	319	529	588

Table 6
Part of speech frequency in the target text

Part of speech	In the Forests of the North	The Law of Life	Nambok the Unveracious	The Master of Mystery	The Sunlanders	The Sickness of Lon Chief	Keesh, Son of Keesh	The Death of Ligoun	Li Wan, the Fair	The League of the Old Men
Noun	1341	529	697	808	1087	673	551	717	122	1251
Adjective	411	192	155	229	248	232	154	175	373	440
Pronoun	977	316	389	605	431	532	268	426	967	875
Adverb	542	225	370	353	514	283	214	205	521	505
Verb	1051	408	674	707	921	558	433	533	1044	980
Numeral	60	36	26	34	86	32	22	49	33	67
Preposition	476	176	339	323	457	267	227	292	468	510
Conjunction	429	177	424	259	693	261	251	237	436	473
Particle	210	94	189	153	177	104	95	78	193	149
Interjection	15	3	8	16	13	8	6	1	6	6
Article	-	-	-	-	-	-	-	-	-	-

Table 7
Ratio of part of speech frequency of source and target texts

Part of speech	**In the Forests of the North**	**The Law of Life**	**Nambok the Unveracious**	**The Master of Mystery**	**The Sunlanders**	**The Sickness of Lon Chief**	**Keesh, Son of Keesh**	**The Death of Ligoun**	**Li Wan, the Fair**	**The League of the Old Men**
Noun	1,1	1,1	1,5	1,1	1,4	1,1	1,4	1,1	10,2	1,1
Adjective	1,4	1,3	1,9	1,4	1,6	1,4	1,7	1,6	1,1	1,3
Pronoun	0,6	1,0	1,3	0,8	1,4	0,9	1,2	1,1	0,6	0,9
Adverb	1,0	1,0	1,1	1,0	1,3	0,9	1,2	0,9	0,8	0,9

Verb	0,8	1,5	1,1	1,2	1,1	1,2	1,1	1,2	0,8	1,2
Numeral	1,0	1,0	0,9	1,2	0,9	0,9	1,3	1,2	0,6	0,9
Preposition	1,7	1,8	1,8	1,5	1,9	1,6	2,0	1,6	1,6	1,4
Conjunction	1,0	1,0	0,8	1,1	0,7	1,2	0,8	1,6	0,9	1,2
Interjection	1,3	3,7	2,0	1,1	0,7	4,0	4,0	16,0	1,7	2,3

Nouns, verbs and adjectives are the most frequent; their relative number in the vocabulary, on the contrary, exceeds the relative number in the text both source and target. These parts of speech represent the vocabulary richness of the source and target texts and their ratio confirms that the nominal character of the individual style of the original text has been preserved in the translation.

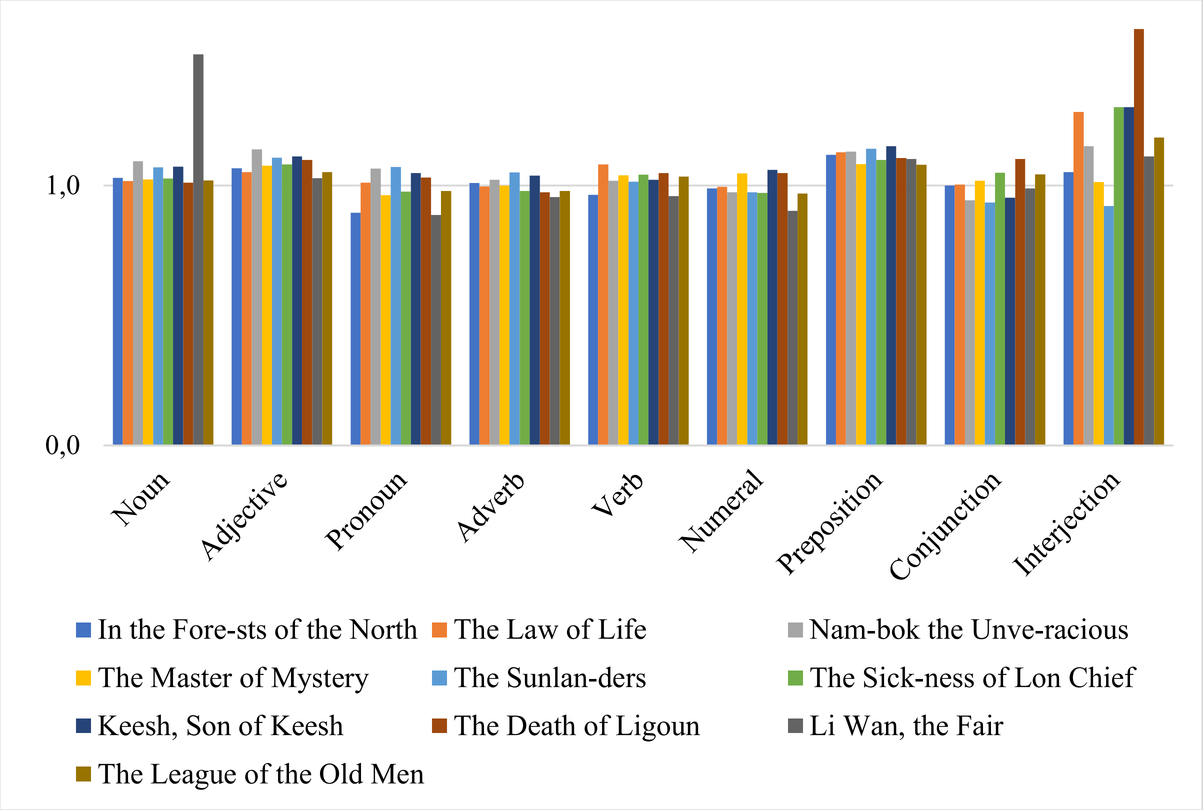

Figure 3. The ratio of each Source text and Target text characteristics (part of speech)

Linguistic and statistical analysis of the corpus under research has been carried out according to the formula developed by S. Buk [5]. The following characteristics of the corpus have been calculated:

- The average word length in source and target texts which is calculated as the total number of letters divided to the total number of words;
- The average frequency of the word in the text (A), which is calculated as the volume of the text (N) divided to the volume of the dictionary of tokens (V). This value is inverse to the index of diversity and is calculated according to the formula (1). In our case, each word of the source texts is repeated at least thrice, and in the target texts – at least twice.

$$A = N / V \qquad (1)$$

- Exclusivity index of the text (Eт) is calculated as a number of words with a frequency of 1 (such words are referred to as hapax legomena) (V1) to the total volume of text (N). The formula is the following:

$$E_{T} = V1 / N \qquad (2)$$

- Exclusivity index of the vocabulary (Ec), i.e. the total number of separate words reduced to the original form (V) is calculated according to the formula:

$$Ec = V1 / V \qquad (3)$$

- The richness of the vocabulary (B) or in other words the index of diversity is calculated as the volume of dictionary of tokens (V) to the volume of text (N). the formula is the following:

$$B = V / N \qquad (4)$$

The higher the index of diversity is, the bigger amount of diverse words the author or the translator used in a particular text. In our case, the index equals 0,264 in the source text and 0,443 in the target text. These indices are high enough, since according to S. Buk, the average index for fiction equals 0.067. [6]

- Concentration index is a value opposite to the index of exclusivity and indicates what share of the text (N) or vocabulary (V) is taken by highly frequency vocabulary (with absolute frequency of 10 or more). Concentration index is calculated according to the formulas: $V10_{T}$ / N is the text concentration index and V10 / V is the vocabulary concentration index.

- Index of lexical density (L) is calculated as the ratio of the number of different words to the total number of words in the text. The algorithm for calculating the index of lexical density includes the following steps: defining an input set of words (either a meaningful text or a part of it, or a random set of words); conversion of each word into its vocabulary form (stemming); deleting all duplicates. The formula for calculating lexical index is

$$L = K / N \qquad (5)$$

where N stands for the number of words after stemming and K stands for a number of words after deleting the duplicates.

- The automated readability index (ARI) is a measure of the complexity of a reader's perception of a text. ARI index is calculated according to the formula:

$$\mathrm{ARI} = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{C} - 21.43, \qquad (6)$$

where C is the number of letters and numbers in the text, W is the number of words in the text and S is the number of sentences in the text. The degree of aggression is the same in the source and target texts and equals 0.19. This confirms the fact that the nominal character of the original text is accurately reproduced in the translation.

- The index of epithetization (Inat), as follows from its definition, indicates the ratio between the total number of nouns in the text (Vn) ant the total number of adjectives (Vadj). The index of epithetization is calculated according to the formula:

$$Inat = Vn / Vadj \qquad (7)$$

The higher the index of epithetization is, the fewer adjectives per noun are present. It can be concluded that this index in source and target texts does not differ significantly: 2.86 / 3.51, and therefore the translator was able to maintain the saturation of the text with figurative phrases.

- The index of verb phrases shows the ratio between adverbs and verbs in the text. The original texts have a slightly bigger percentage: 0.47 adverbs per 1 verb, while in translation – 0.51 per 1.
- Nominality degree shows the ratio between nouns and verbs in the text. In the original texts, there are 1.32 nouns per verb, in translation – 1.22 per 1.
- The average sentence size indicates the peculiarities of verbal intelligence or a radical change of emotional state. There is a negative correlation between the increase of emotionality of speech and the amount of speech. In other words, the more emotional the speaker is, the shorter their statements are.
- The coefficient of aggression represents the ratio between the number of verbs (and participles) and the total number of the words in the text. The coefficient is calculated according the formula:

$$\text{Aggression coefficient} = \text{N verbs / N of all words} \times 100\%, \quad (8)$$

where N – number of appropriate words.

High coefficient of aggression indicates considerable emotional tension of the text, dynamics of events, poor emotional state of the author during text synthesis.

- The coefficient of logical coherence represents the ratio between the total number of function words (prepositions and conjunctions) and the total number of sentences in the text. Values within 1 show a fairly harmonious ratio between function words and syntactic constructions in the text.

$$\text{The coefficient of logical coherence} = \text{N service words / N sentences}, \quad (9)$$

where N – number of appropriate words.

- The coefficient of embolism means pragmatic tagging or clogging of speech and represents the ratio between the total number of emboli (words that do not have semantic meaning) and the total number of words in the text. Such words include interjections, vulgarisms, repetitions, etc. The coefficient of embolism negatively correlates with the indicators of verbal intelligence and the degree of emotional excitement of the speaker / author of the text. The coefficient of embolism is calculated according to the formula:

$$\text{Embolism ratio} = \text{Nembol / All words} \times 100\%, \quad (10)$$

where N – number of appropriate words.

The quantitative indices, which have been calculated on the basis of the general characteristics of the source and target texts, have been compared (Table 8).

Table 8
Quantitative indices in the source and target texts

Coefficient	The average value in the corpus of the source text	The average value in the corpus of the target text	Ratio of average values of source text and target text
Average word length	4,56	5,381	0,8
Average word frequency	3,846	2,294	1,7
Vocabulary exclusivity index	0,538	0,503	1,1
Diversity index	0,264	0,443	0,6
Exclusivity index for text	0,144	0,217	0,7
Vocabulary	0,058	0,0731	0,8

concentration index			
Lexical density index	0,69	0,765	0,9
Automatic readability index	7,229	9,87	0,7
Index of epithetization	2,86	3,511	0,8
Index of verb phrases	0,472	0,51	0,9
Degree of nominality	1,322	1,22	1,1
An average sentence size	14,956	11,012	1,4
Coefficient of aggression	0,194	0,196	1,0
Coefficient of logical coherence	3,113	2,083	1,5
Coefficient of embolism	0,0037	0,0415	0,1

As presented in table 8, the main indicators that characterize the individual style in the source and target texts, do not differ significantly (Figure 4), except of average word frequency, which in source text is almost twice higher, and the coefficient of embolism is ten times higher in a target text, then it is in source text.

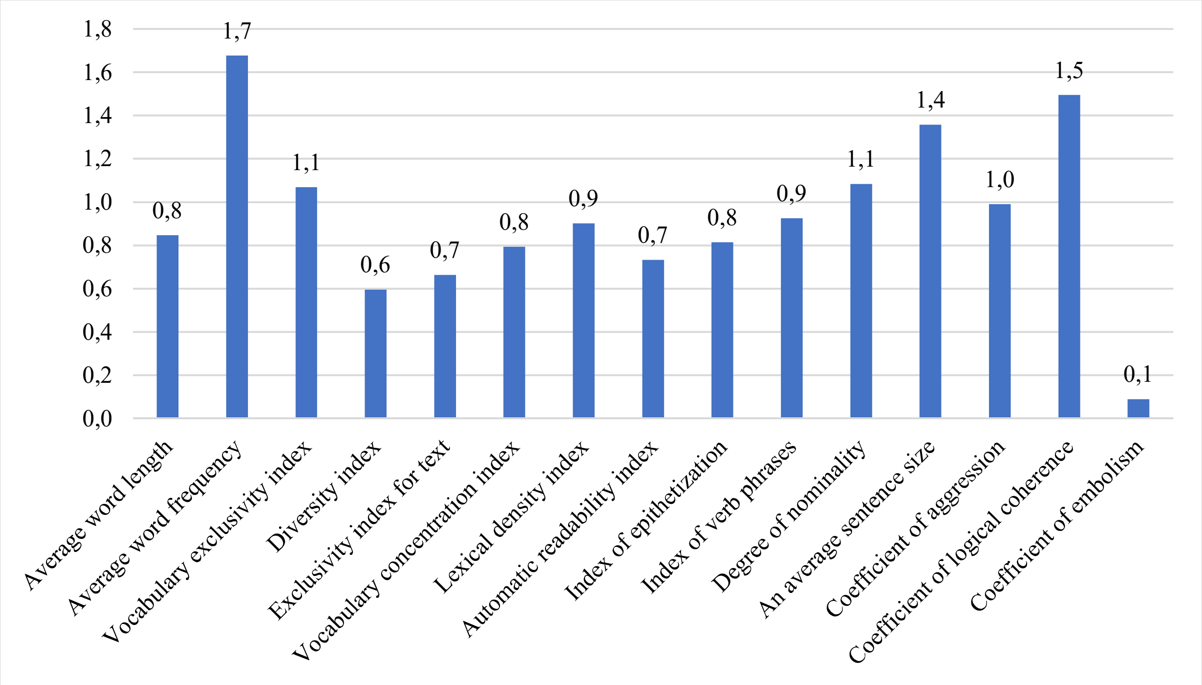

Figure 4. Indicators characterizing the individual style in source and target texts

To determine the significance / insignificance of the statistical difference between the values of the indices, t-criterion has been calculated, using the appropriate functions in Excel. For the given data on our samples, the t- criterion equals 0,69.

To decide whether the t-criterion indicates a significant difference, it is necessary to evaluate it according to the table of critical values of t. This evaluation is carried out by determining the number of degrees of freedom, which in our case f = 15-2 = 13 (the number of indicators subtract the number of samples under comparison). The difference is considered significant if the calculated value of t is greater than the tabular value for a given level of significance. In our case, 0,69 is less than the smallest number in rows. This means that the difference in the statistical indicators of the source and target texts is insignificant and statistically acceptable.

4. Conclusion

All in all the paper presents the quantitative comparative study of the collection “Children of the Frost” by J. London and its Ukrainian translation by V. Hladka and K. Koriakina, which have not been analyzed from the statistical viewpoint before. Concluding the research it can be noted that:

- the number of word usages in the source text exceeds the number of word usages in the target text both in the whole corpus and in separate stories. In general, the volume of the source text is bigger than the volume of the target text by 20.69%;
- indices of vocabulary richness, exclusivity for the text and the vocabulary, the concentration of the vocabulary do not differ significantly;
- the mainly used parts of speech in English and Ukrainian texts are nouns (22.22% and 25.21%), verbs (19.27% and 18.89%), adjectives (7.7% and 7.64%) and adverbs (7.2% and 7.64%);
- the translation preserves the ratio of different parts of speech. The number of pronouns, adverbs and functional words in the vocabulary of the target text has slightly decreased;
- the index of epithetization which indicates the number of nouns per adjective in the text, does not differ significantly in source and target texts – 2.86 / 3.51;
- the index of verb phrases shows the number of adverbs per verb in the text. The index is higher in the source text 0.47 adverbs per 1 verb, while in target text 0.51 per 1;
- degree of nominality shows the number of nouns per verb. In the source text, there are 1.32 nouns per verb, in the target text – 1.22 per 1. Therefore, the degree of aggression, which is calculated as the ratio of the number of verbs and verb forms (particles) to the total number of the words, is identical in source and target text and equals 0.19. This confirms the fact that the nominal character of the source text has been accurately reproduced in the target text.

Various linguistic disciplines will benefit from the research findings. These findings can be applicable in the analysis conducted within the scope of corpus linguistics, translation studies, literary studies, discourse analysis, lexicography etc.

5. References

[1] A. Baranov, Introduction to Applied Linguistics, Nauka, Moscow, 2001.

[2] D. Barth, S. Stefan, Understanding Corpus Linguistics, Routledge, 2022.

[3] D. Biber, S. Conrad, R. Reppen, Corpus Linguistics. Investigating Language Structure and Use, Cambridge University Press, Cambridge, 1998.

[4] D. Biber, Representatives in Corpus Design, in: Literary and Linguistic Computing, volume 8, № 4, 1993, pp. 243–257.

[5] S. Buk Texts’ Quantitative Comparison (based on the 1884 and 1907 editions of Ivan Franko’s novel “Boa Constrictor”, in: Ukrainian Literary Studies, Issue 76, 2012, pp. 179–192.

[6] S Buk, Statistical Characteristics of the Lexis of Main Functional Styles of the Ukrainian Language, in: Lexicographic Bulletin, pp. 166–170.

[7] A Chesterman, Hypotheses about translation universals, in: G. Hansen, K. Mlmkjær, D. Gile (eds.), Claims, Changes and Challenges in Translation Studies, John Benjamins Publishing Company, Amsterdam, 2004, pp. 1–13.

[8] G. Corpas Pastor, R. Mitkov, N. Afzal, L. Garcia Moya, Translation Universals: Do they exist? A corpus-based and NLP approach to convergence, Proceedings of the LREC (2008) Workshop on “Comparable Corpora”, LREC-08, Marrakesh, Morocco, 2008.

[9] Corpus Linguistics, ed. V. Shyrokov, O. Bugakov, T. Hriaznukhina, Dovira, Kyiv, 2005.

[10] Corpus-based Language Studies: An Advanced Resource Book, ed. T. McEnery, R. Xiao, Y. Tono, Routledge, London, 2006.

[11] N. S. Dash, S. Arulmozi, History, Features, and Typology of Language Corpora, Springer, 2018.

[12] G. Desagulier, Corpus Linguistics and Statistics with R, Springer, 2017.

[13] O. Demska-Kulchytska, The Bases of the National Corpus of the Ukrainian Language, Kyiv, 2005.

[14] R. Garside, G. Leech, T. McEnery, Introducing Corpus Annotation, Longman, London, 1997.

[15] A.P. Gomide, Corpus Linguistics Software: Understanding Their Usages and Delivering Two New Tools, Ph.D. thesis, Lancaster University, 2020.

[16] G. Kennedy, The corpus as a research domain, in: Comparing English worldwide: The International Corpus of English, Clarendon Press, Oxford, 1996, pp. 217–226.

[17] I. Khomytska, V. Teslyuk, N. Kryvinska, I. Bazylevych, Software-Based Approach Towards Automated Authorship Acknowledgement – Chi-Square Test on One Consonant Group, in: Electronics, Vol. 7:1138, July 2020.

[18] I. Kulchytskyi, Technical Aspects of Natural Language Information Processing, in: The Journal of Lviv Polytechnic National University, Informational Systems and Networks, № 783, 2014, pp. 344–353.

[19] A. Laurence, A Critical Look at Software Tools in Corpus Linguistics, In: Linguistic Research 30 (2), pp. 141–161.

[20] G. Leech, Introducing Corpus Annotation, in: R. Garside, G. Leech and A. McEnery (eds.) Corpus Annotation, Longman, London, 1997, pp. 1–18.

[21] O. Levchenko, O. Tyshchenko, M. Dilai, Automated Identification of Metaphors in Annotated Corpus (Based on Substance Terms), in: CEUR Workshop Proceedings, 2021, Vol. 2870, Proceedings of the 5th International conference on computational linguistics and intelligent systems (COLINS 2021), Vol. I: main conference, Kharkiv, Ukraine, April 22-23, 2021, pp. 16–31.

[22] J. London, Children of the Frost, URL: https://library.um.edu.mo/ebooks/b28284872.pdf.

[23] J. London, Dity morozu, in: Tvory v 12-kh tomakh, Tom 1, Per. z angl. V. Hladka, K. Koriakina, Dnipro, Kyiv, 1968. (Дж. Лондон, Діти морозу, в: Твори в 12-х томах, Том 1, Пер. з англ. В. Гладка, К. Корякіна, Дніпро, Київ, 1968).

[24] A. Lüdeling, M. Kytö, Corpus Linguistics, An International Handbook, Volume 1, Walter de Gruyter, Berlin, New York, 2008.

[25] M. Paquot, S. Th. Gries, A Practical Handbook of Corpus Linguistics, Springer, 2020.

[26] N. Romanyshyn, Application of Corpus Technologies in Conceptual Studies (based on the Concept Ukraine Actualization in English and Ukrainian Political Media Discourse), in: 4th International Conference on Computational Linguistics and Intelligent Systems, Colins 2020, pp. 472–488.

[27] M. Sebba, S. D. Fligelstone, Corpora, in: Ronald E. Asher & James M.Y. Simpson (eds.), vol. 2, Pergamon, Oxford, 1994, pp. 769–773.

[28] J. Sinclair, Corpus, Concordance, Collocation, Oxford University Press, Oxford, 1991.

[29] A. Stefanowitsch, Corpus Linguistics: A Guide to the Methodology, Language Science Press, Berlin, 2020.

[30] W. Teubert, Corpus Linguistics and Lexicography, in: International Journal of Corpus Linguistics, Vol. 6, Special issue, 2001, pp. 125–153.