Corpus Technologies in Translation Studies: Fiction as Document

Nataliia Hrytsiv, Tetiana Shestakevych and Julia Shyyka

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

Abstract

In the paper we focus on morphological, syntactic, semantic and anaphoric markup. Based on the raw material obtained from Jack London's texts, we come to the following findings: (a) indices of vocabulary richness, exclusivity for the text and the vocabulary, the concentration of the vocabulary do not differ significantly; (b) the most frequent in the target text are functional words; (c) the epithetization index indicates the number of nouns per adjective in the text; (d) the index of verb phrases indicated the number of adverbs per verb; (e) the degree of nominality shows the number of nouns per verb, in the original text there are 1.23 nouns per verb, in the translated text - 1.36 per 1. To define the significance/insignificance of the statistical difference between the values of the coefficients for the source and target texts, $\chi 2$ or criterion of homogeneity in linguostatistics has been calculated. Concluding the quantitative study of the collection "The Children of the Frost", it can be noted that: (a) 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.

Keywords [1]

Translation studies, corpus linguistics, fiction, text mark-up, part-of-speech tagging, AntConc.

1 Introduction

A central notion of corpus linguistics is the concept of "corpus of texts". On the one hand, it is the main goal of corpus linguistics, and on the other hand, it is the object of study in this field of linguistics [1].

The linguistic corpus is characterized by some unique features that distinguish it from ordinary collections of digital texts. The basic features of the text corpus are machine-readable (existence of a digital form of text representation and a special system of encoding text data) and representativeness. There are several similar definitions of corpus, which are based on the features of machine-readability and representativeness as mandatory, determinative, for example: "collection of machine-readable texts selected in such a way as to best represent the language and its diversity"; "corpora" is a large number of natural language texts that have a computer form and are the object of a certain linguistic study, where "natural" means everything that was expressed orally or in written form" [2, 3]. Complementing machine readability and representativeness with the parameter of corpus applicability in linguistic research, N. Dash and B. Chaudhuri define corpus as "a collection of linguistic data composed either from written texts or from transcribed oral texts, the main purpose of which is to examine hypotheses about language".

The main problem in corpus linguistics is the creation of means of automatic (or at least automated) text annotation according to different criteria – morphological, orthoepic, semantic, syntactic, etc. According to V.A. Shyrokov, the main idea of the system engineering of a linguistic corpus (if to omit technical details) is to automatically divide the electronic text of a literary source into "microcontexts" i.e. in text fragments that are "grouped" around the word that is the object of interpretation. Thus, there

MoMLeT+DS 2021: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, 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)

is no need to form and store a traditional linguistic object – a lexical card as a separate physical object; it becomes a virtual object, i.e. implemented as a relationship in the database. The key issue of corpus issues are widely ranged and involve studies of the general theory of corpus linguistics.

2 Theoretical and procedural background

In modern corpus linguistics apart from the creating of regular language corpora, the creation of author corpora is widespread. Of special scholarly interest is a multilingual corpora are comparable and parallel corpora or translation corpora [4].

2.1 Corpus technologies in translation studies

In our research, it is advisable to conduct a contrastive analysis based on the comparison of certain quantitative information of the original text and its translations at the lexical level. The defining of such equivalence (or its absence) between the original text and the translation has been carried out in our study.

The main tool of corpus linguistics, which allows achieving the goal in a particular area, is the corpus, which is viewed as a set of organized digital texts, which are used for certain linguistic purposes.

Compiling a corpus, it is necessary to take into account some criteria: providing the representativeness, authenticity, selectivity, balance, machine readability and markup. The markup process (or annotation) fills the corpus with information that can then be used to investigate specific issues [5, 6]. Annotation involves adding extralinguistic, structural, and linguistic special markers to texts or their components. There are several types of linguistic markup: morphological, syntactic, semantic, anaphoric and prosodic. Also, the following procedures are performed: tokenization, lemmatization, stemming and parsing. Most of the existing corpora are of the morphological or syntactic type. It is worth mentioning that syntactic corpora explicitly or implicitly contain morphological characteristics of lexical units.

2.2 Structuring a case study

Creating a corpus is quite a difficult task. According to Leach, to obtain some results, it is necessary to carry out some preliminary work. Creating a corpus takes twice as long and sometimes ten times more effort than its use. As has been already mentioned, the linguistic corpus contains a markup on at least one linguistic parameter. This feature distinguishes the linguistic corpus from a large number of other linguistic information and tool systems or databases. In other words, the corpus = text + markup. The process of tagging [7, 8] or annotation is to attribute to the texts and their components special tags:

- external, extralinguistic (information about the author and the text: author, title, year and place of publication, genre, subject; information about the author may include not only his/her name, but also age, gender, life-years etc. (this information coding is called meta-markup);
- structural (chapter, paragraph, sentence, word form);
- linguistic, which describe lexical, grammatical and other characteristics of the text elements.

The set of these metadata largely determines the competencies provided by the corpora to researchers. The annotation adds value to the corpus, as it significantly expands the range of research issues that can be investigated on the material of this corpus. When choosing these data it is necessary to take into consideration the purposes of research and needs of linguists as well as the possibility to add to the text some extra features.

It is also necessary to note the difference between annotation and structural markup of the corpus/corpus data. The distinction between these two concepts is based on the definition of annotation offered by G. Leach, according to which "the process of annotation of corpus data is the adding of interpreted linguistic information to the digital corpus of oral and/or written speech.

The term "interpreted linguistic information" means an annotation that is to some extent a product of the understanding of the text by the human mind (for example, an annotation on parts of speech).

Markup provides relatively objective verified information about the parts of the corpus and the structure of each text.

Linguistic annotation in corpus linguistics is traditionally interpreted as:

- any linguistic information about linguistically relevant units of text data presented in a formal code;
- adding formalized linguistic information into the digital text;
- presence of such information in the text.

Linguistic annotation can be performed at different levels and take different forms. Thus, there are the following types of markup:

- morphological markup: in foreign terminology, the term part-of-speech tagging (POS-tagging) is used. Morphological markups include not only the features of a part of speech but also the features of grammatical categories specific to that part of speech. This is the main type of markup because most large corpora are morphologically marked. Besides, morphological analysis is considered as a basis for further forms of analysis – syntactic and semantic. Moreover, advances in computer morphology allow marking large corpora automatically. For example, the British National Corpus is marked like that. The used markup is known as markup C5 (basic) and C7 (supplemented).
- syntactic markup: that results from syntactic analysis or parsing, based on morphological analysis data. This type of markup describes the syntactic relations between lexical units and various syntactic constructions (for example, a subordinate clause, a verb phrase, etc.).
- semantic markup: although there is no single semantic theory for semantics, most often semantic tags denote the semantic categories to which a word or phrase belongs as well as narrower subcategories that specify its meaning.
- anaphoric markup: it fixes referential connections, for example, of the pronoun. Prosodic markup. In prosodic corpora, markups describing stress and intonation are used. In the oral speech corpora, prosodic markup is often accompanied by so-called discourse markup, which is used to indicate pauses, repetitions, warnings, etc. Other types of information can also be encoded in the corpus. For example, sociolinguistic information about such characteristics as gender, age, social status, and place of residence may be presented in colloquial speech corpora.

During the creation of the corpus several procedures and programs are used, such as tokenization, lemmatization, stemming and parsing.

Tokenization is a division of the string of natural language symbols into separate significant units (tokens, word forms).

Lemmatization is a process of forming the initial form of a word, based on its other forms.

Stamming is the process of defining a stem of a word.

Parsing is the process of analyzing the syntactic structure of a text or part of a text, which is based on comparing the linear sequence of tokens (words, tokens) of language with its formal grammar. The construction of automatic parsers for large corpora is one of the most important areas of computational linguistics.

Compiling a corpus and its use can be implemented according to the “model”: Annotation - Abstraction - Analysis. The first stage involves the collection, standardization, segmentation, processing and often manual checking of texts. At the second stage, the research topic is selected, the parameters are determined and the corresponding fragment is removed from the corpus. At the final stage, the hypothesis is constructed and examined, specific rules and systematized structures are searched. During the analysis, automatic and interactive methods, as well as quantitative and qualitative analytical methods are used. The most difficult is to identify specific issues to solve real problems of theoretical, descriptive and applied linguistics.

Thus, fulfilling all the requirements for the creation of a linguistic corpus, a convenient tool for further use in the work of any complexity to solve certain problems can be obtained.

2.3 A case study constrains

The sample of texts for the study has been formed according to the rules of sampling by V. Perebyinis:

- texts should be chronologically limited (the chronological boundaries for the materials of our study are 1900-1902, it was during this period that collections of short stories "The Son of the Wolf" (1900), "The God of His Fathers" (1901) and “Children of the Frost” (1902) were written and published. Fathers (1901);
- texts should be limited in the genre (the sample for our study contains only short stories, so it represents the short prose of the writer);
- texts should be thematically limited (the main topics of the stories are the lives of gold diggers and Indians of the North as well as travelling and adventures in the southern seas);
- texts should be homogeneous in the author's style (all stories belong to one writer, and proving the homogeneity of the author's style is the task of our study).

3 Results and discussion

3.1 Raw material resource base

It has been presumed that statistical population is presented by the Northern stories of J. London and their Ukrainian translations, which were made in the early twentieth century and of which the sample has been formed by the method of purposeful sampling which corresponds to the main condition of its organization i.e. representativeness. The sample size is 82665-word usages.

The following works were selected for further analysis:

Table 1

Source base of the research (original stories)

Story	Number of word usages
The White Silence	3733
The Son of the Wolf	6114
The Men of Forty-Mile	3156
In a Far Country	6239
To the Man on the Trail	3139
The Priestly Prerogative	4103
The Wisdom of the Trail	2988
The Wife of a King	4834
An Odyssey of the North	10669
In the Forests of the North	5970
The Law of Life	2836
Nam-Bok the Unveracious	4500
The Master of Mystery	4085
The Sunlanders	6368
The Sickness of Lone Chief	3632
Keesh, the Son of Keesh	3135
The Death of Ligoun	3610
Li Wan, the Fair	5249
The League of the Old Men	6293
Total:	**91253**

Thus, 19 stories of J. London and 19 translations into Ukrainian have been selected.

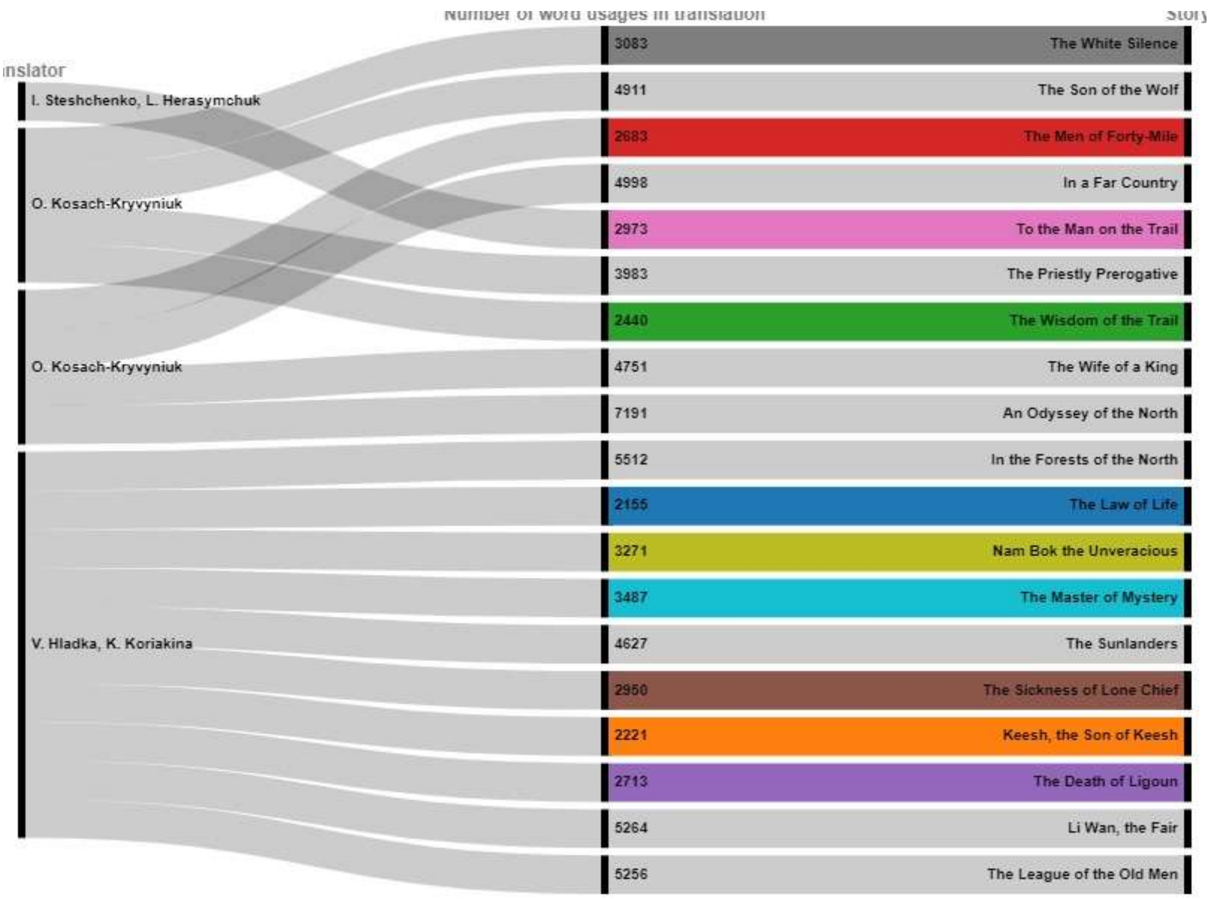


Figure 2: Source base of the research (translations)

3.2 Concordance and AntConc options

The text components include units of all language levels. The most significant unit is the word, the minimum semantic language unit, which is freely reproduced in the language and serves to build expressions (sentences). Stylistic analysis of texts shows that texts differ in units of all language levels, but the stylistic potential of these units is different: the smallest unit is at the phonetic level, the largest unit is at the lexical and syntactic levels. That is why lexical units of text (words) are the object of our research.

For quantitative research, the AntConc program environment has been used [9, 10, 11]. V. Shyrokov claims that the main and quite effective working tool of corpus linguistics is the concordancer, a special program, which quickly processes the corpus of texts and can perform the following functions: search for morphemes, lexical units, phrases in the context; determining the frequency rank of grammatical forms. Dictionaries and concordances can be compiled with the help of specialized programs and tools based on the corpus of texts of any genres.

A distinction should be made between two terms - "concordance" and "concordancer". Concordance is defined as a catalogue that contains all words used in a particular text or the works of a particular author (authors) or a list of all usages of the selected word, presented in the context of its use.

In linguistic research, concordance is used to perform the following tasks: comparing different word usages of one lexical unit, measuring and comparing the frequency of words and phrases, searching and analyzing idioms and paremiological units, creating word lists by selected criteria, analyzing terminological vocabulary.

Concordancer is defined as instrumental software that searches in corpus the usage of the required word, phrase or another specified element with a representation of the contextual environment of this unit".

The advantage of the concordancer over other software for linguistic research is the ability to detect the compatibility of words and their environment. The list of concordancer options is larger than concordance: it can search for words by lemmas and morphological features, search for specific language units and present results in concordance views, display text-type information, output lexical and grammatical statistics, output search results from a predefined number of words (length of word context), save the results in a separate file.

Concordancers can be divided into are index (require pre-indexing of text units), non-index, or network (require Internet connection and are based on its technology, mostly paid) and stationary (can be downloaded to a personal computer, do not require an Internet connection, mostly free).

To conduct our study, we used the AntConc program i.e. a stationary, non-index concordancer. The technical advantages of this program include free access to the Internet, free upgrades and compatibility with the main operating systems - Windows, Mac OS X and Linux. The AntConc program is very useful for linguists who are engaged in linguistic and statistical research because it can perform the following operations:

- to find lexical units and stable expressions of a certain subject area, given by the user;
- to make samples of lexical units which belong to one thematic group;
- to identify usage contexts of given words.

Processing the entered data, AntConc can sort the selected language units according to the criteria set by the user, namely:

- frequency;
- word ending;
- alphabetical order.

The program also helps to choose the number of words to the right and left of the search word displayed in the program window, get the attributes of a given word in alphabetical order, search for collocations by constructing n-grams of different lengths, compare keywords in different text boxes.

So AntConc is a free, multi-platform tool for corpus linguistics research and data learning. This program is developed in Perl using the PerlApp ActiveState compiler to create executable files for different operating systems. It does not require special installation, just double-click on the program icon to run it. AntConc contains seven tools that can be run by clicking on their "bookmarks" in the program window, or by using the function keys F1 - F7.

This tool enables the identification of corpus-specific words. They have the following characteristics:

- Concordance. It presents search results in KWIC format (keyword-in-context) and shows how words and phrases are usually used in the corpus of texts.
- Concordance Plot. This tool shows search results on a barcode chart and shows where the search results appear in the ascending texts.
- File Viewer. This tool shows the text of separate files and allows interrogating the results obtained by other tools of AntConc.
- Clusters (N-Grams). This tool shows clusters based on search terms. It summarizes the results obtained by such tools as Concordance or Concordance Diagram. The N-gram tool scans the entire corpus for the length of the clusters in the "N" number of words (for example, one word, two words, ...). This enables us to find common expressions in the corpus.
- Collocates. This tool shows search word collocations and enables to explore inconsistent patterns in language.
- Words List. This tool counts all the words in the corpus and presents them in an ordered list and allows finding which words are most used in the corpus.
- Keyword List. This tool shows which words are unusually common (or rare) in the corpus.

The screenshots and figures to follow present the program interface in details and the list of the abovementioned tools. Thus, the quantitative distribution of words in originals texts and their translations has been made with the help of the program AntConc.

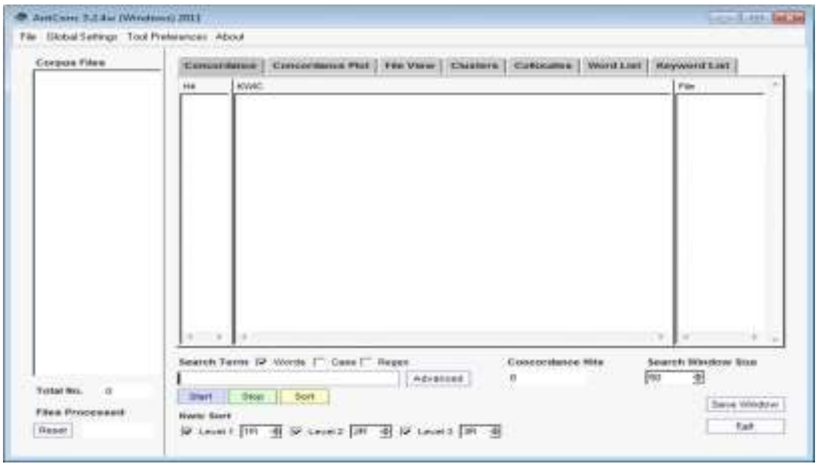
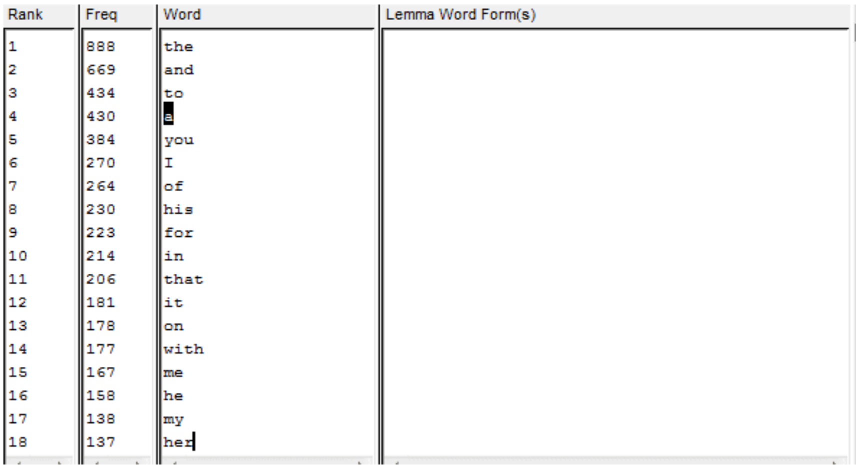
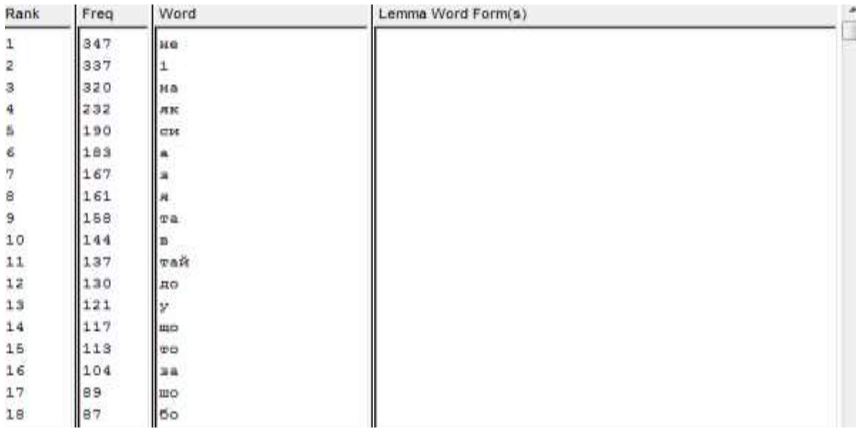


Figure 2: AntConc interface



Rank	Freq	Word	Lemma Word Form(s)
1	888	the	
2	669	and	
3	434	to	
4	430	a	
5	384	you	
6	270	I	
7	264	of	
8	230	his	
9	223	for	
10	214	in	
11	206	that	
12	181	it	
13	178	on	
14	177	with	
15	167	me	
16	158	he	
17	138	my	
18	137	her	

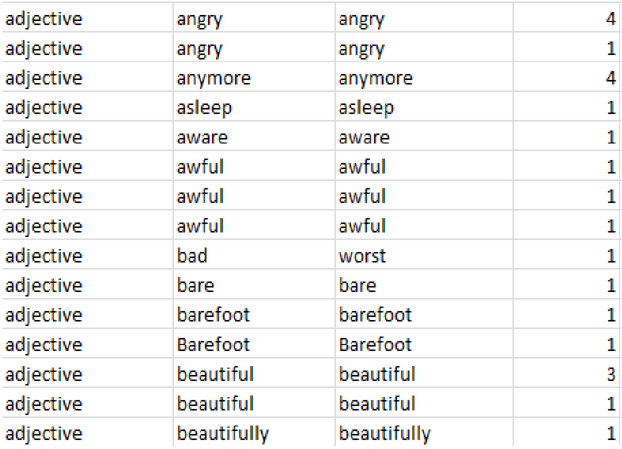
Figure 3: Quantitative word distribution in the original



Rank	Freq	Word	Lemma Word Form(s)
1	347	не	
2	337	і	
3	320	на	
4	232	як	
5	190	си	
6	183	а	
7	167	з	
8	161	я	
9	158	та	
10	144	в	
11	137	тай	
12	130	до	
13	121	у	
14	117	що	
15	113	то	
16	104	за	
17	89	шо	
18	87	бо	

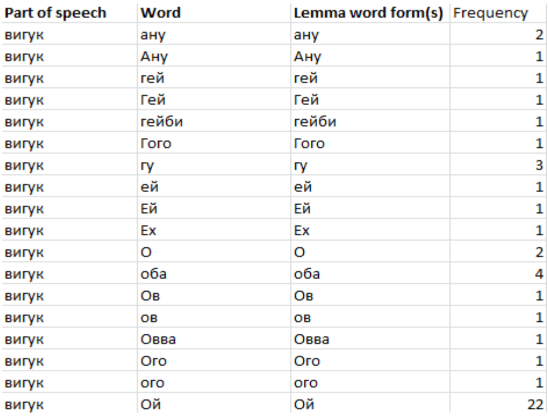
Figure 4: Quantitative word distribution in the translation

At the next stage of the study, the results of preliminary processing of the texts have been transferred to the MS Excel environment and the part of speech of each word as well as its lemma and the number of uses separately for Ukrainian and English texts has been identified



adjective	angry	angry	4
adjective	angry	angry	1
adjective	anymore	anymore	4
adjective	asleep	asleep	1
adjective	aware	aware	1
adjective	awful	awful	1
adjective	awful	awful	1
adjective	awful	awful	1
adjective	bad	worst	1
adjective	bare	bare	1
adjective	barefoot	barefoot	1
adjective	Barefoot	Barefoot	1
adjective	beautiful	beautiful	3
adjective	beautiful	beautiful	1
adjective	beautifully	beautifully	1

Figure 5: Analysis of parts of speech and quantitative distribution of words in the original texts



Part of speech	**Word**	**Lemma word form(s)**	Frequency
вигук	ану	ану	2
вигук	Ану	Ану	1
вигук	гей	гей	1
вигук	Гей	Гей	1
вигук	гейби	гейби	1
вигук	Гого	Гого	1
вигук	гу	гу	3
вигук	ей	ей	1
вигук	Ей	Ей	1
вигук	Ех	Ех	1
вигук	О	О	2
вигук	оба	оба	4
вигук	Ов	Ов	1
вигук	ов	ов	1
вигук	Овва	Овва	1
вигук	Ого	Ого	1
вигук	ого	ого	1
вигук	Ой	Ой	22

Figure 6: Analysis of parts of speech and quantitative distribution of words in translation

In the current study the traditional classification of parts of speech has been used:

- for Ukrainian: content words such as nouns, adjectives, pronouns, verbs (particle, transgressive), adverbs, numerals; function words such as prepositions, conjunctions, particles, interjections;
- for English: content words such as nouns, adjectives, pronouns, verbs, adverbs, numerals; function words such as prepositions, conjunctions, articles, interjections.

The following principles of combining word forms have been used during lemmatization:

- for Ukrainian: noun forms have been reduced to the nominative singular; verbs - to the infinitive; all adjective forms, including degrees of comparison, have been reduced to the nominative singular of the masculine gender; the comparative and superlative adverbs have been reduced to the original forms of adverbs; case forms have been reduced according to the type of declension of pronouns and numerals; phonetic variants of words have been reduced to the original form (the most frequent), where the alternation of first or final letters is caused by euphonism;
- for English: possessive forms and plural forms of the noun have been reduced to the original form; all tense forms of verbs, gerunds, Participle I and Participle II have been reduced to the infinitive; the degrees of comparison of adjectives have been reduced to the original form.

With the help of a specially written computer program, the absolute frequency of each lemma has been automatically calculated.

3.3. Quantitative characteristics of the original: a case study of 'Sun of the Wolf' collection

In our study, priority has been given to the vocabulary of the original text and its translation, and with the help of the automatic processing of the corpus and statistical calculations, several important characteristics have been defined which can form the basis for clarifying the writer's individual style and help to conclude the aesthetic significance of the original texts and its translations.

From the point of view of the quantitative-linguistic analysis of texts [12-15], several problems have been defined which refer to the stylistic aspect of research of vocabulary of the given text, in particular a volume of the text i.e. total number of words in the text (N), volume of vocabulary, number of word forms (Vf), the volume of the vocabulary of lexemes, the number of lemmatized words in the text (V), Hapax legomena (V1) i.e. words that occur in the sample once and their frequency equals 1, the number of words with a frequency ≥ 10 (V10), number of letters (C), number of sentences (S), number of content words (Nps), number of functional words (Npsa) [16, 17]. Based on these statistic data, it is possible to calculate:

- vocabulary richness, diversity index (Id) - the ratio of the volume of the vocabulary of lexemes (V) to the volume of the text (N) is calculated by the formula: Id = V / N;
- the average repeatability of the word in the text (Iwr) - the ratio of the volume of the text (N) to the volume of the vocabulary of lexemes (V) - the value inverse to the diversity index is calculated by the formula: Iwr = N/V;
- exclusivity index is calculated separately for the vocabulary and the text, it characterizes the variability of vocabulary, i.e. the part of the text (vocabulary) which consists of the words that occurred once in the text: the index of exclusivity for the vocabulary (Iev) - the ratio of lexemes with frequency 1 (V1) to the total number of lexemes: Iev =V1/V; text exclusivity index (Ien) - the ratio of the number of lexemes with frequency 1 (V1) to the text volume (N): Ien = V1/N;
- dictionary concentration index (Ivc) - the ratio of the number of words in the vocabulary with an absolute frequency of 10 or more (V10) to the total number of words in the vocabulary: Ivc = V10/V;
- lexical density index (Id) - the ratio of content words (Nps) in the text to the total number of words: Id = Nps / V;
- automatic readability index (ARI) - the degree of readability of texts, the ratio of characters in the word and the number of sentences; it is calculated by the formula: ARI = 4.71 * C / V + 0.5 * V / S - 21.43;
- noun phrases index (Inat), i.e. the epithetization index (the ratio between the number of nouns (Vn) and the number of adjectives (Vadj): Inat = Vn/ Vadj;
- verb phrases index (Ivat) - the ratio of the number of adverbs (Vadv) to the number of verbs (Vv): Ivat = Vadv/Vv;
- degree of nominality (Inom) - the ratio of the number of nouns (Vn) to the number of verbs (Vv): Inom = Vn/Vv.
- verb index (aggressiveness index) (Iv) - the ratio of the number of verbs and verb forms (particles and adverbs) (Vv) to the total number of the words: Iv = Vv / N;
- logical coherence index (Ilk) - the ratio of the total number of functional words (conjunctions and prepositions) (Vpc) to the total number of sentences (S): Ilk = Vpc / S;
- embolism index (Iem) (clogging) of speech – the ratio of the total number of emboli (exclamations and particles) (Vem) to the total number of words: Iem = Vem / N.

According to this scheme, the subcorpus of the original texts of the collection "The Son of the Wolf" has been analyzed. General quantitative characteristics of the stories are given in Table 2.

Morphological markup of the corpus of short stories has been performed, the classical division of words into parts of speech has been applied, and for each part of speech, its frequency in the text and the author's vocabulary (register) has been automatically obtained:

The most frequent words in the original text are functional words: 5.2% of the vocabulary. They function most actively in the text and cover almost a third of it: 30.7%. Pronouns are also frequently used in the text: 3% of the vocabulary and about 13% of the text. Adverbs have approximately the same percentage in the text and the vocabulary (9% and 10%, respectively) and numerals: about 1%. Nouns (22% and 39%), verbs (19% and 26%) and adjectives (8% and 16%) in the text and the writer's vocabulary, respectively, confirm their stylistic function and their ratio proves the nominality of J. London's individual style.

Based on the established general quantitative characteristics of the text and the base of partial linguistic distribution indicators, the indices that characterize the lexical level of the corpus have been calculated. The results of the calculations are presented in Table 3.

Table 2
General quantitative characteristics of the original text

Story	1. Number of word usage	2. Number of word forms	3. Number of words	4. Hapax legomena for word forms	5.Number of word forms used 10 or more times	6. Hapax legomena for lemmas	7. Number of lemmas used 10 or more times	8. The number of characters of the extended alphabet in the text	9. The number of sentences in the text
The White Silence	3733	1326	1072	891	52	682	55	15835	235
The Son of the Wolf	6114	1785	1471	1166	85	904	87	26236	380
The Men of Forty-Mile	3156	1185	949	830	48	615	47	13468	219
In a Far Country	6239	2032	1779	1390	80	1168	76	28009	391
To the Man on the Trail	3139	1223	986	844	39	636	39	13582	190
The Priestly Prerogative	4703	1487	1223	983	80	783	74	19622	367
The Wisdom of the Trail	2988	1027	875	689	48	567	43	12826	166
The Wife of a King	4834	2404	1566	1826	49	943	74	24556	323
An Odyssey of the North	10669	1818	1755	1037	152	1037	152	42601	683
Total:	**45575**	**14287**	**11676**	**9656**	**633**	**7335**	**647**	**196735**	**2954**

Although the indices of epithetization, nominality and verb phrases are not the main characteristics of the stylistic interpretation of the text, they can be considered a significant complement to the qualitative analysis of the text, especially in comparing the original text and the translated text. Nominative and verb (aggressiveness) indices confirm the nominal individual style of J. London's short prose, which is characterized by a high level of epithets, verb phrases, low "clutter" of speech and high logical coherence.

Table 3
Quantitative indicators of the lexical level of the original text

Story	The richness of the vocabulary	The average repetition of the word in the text	Index of exclusivity for the text	Index of exclusivity for the vocabulary	Index of text concentration	Index of vocabulary concentration	Automatic readability index	Lexical density index	Index of nouns phrases	Index of verb phrases	Index of nominality	Verb index	logical coherence index	Speech embolism index
The White Silence	0,29	3,48	0,24	0,64	0,01	0,05	6,49	0,19	2,79	0,3	1,2	0,19	3,02	0,02
The Son of the Wolf	0,24	4,16	0,19	0,61	0,01	0,06	6,83	0,19	3,4	0,35	1,17	0,19	3	0
The Men of Forty-Mile	0,3	3,33	0,26	0,65	0,02	0,05	5,88	0,2	4,66	0,25	1,19	0,21	2,83	0
In a Far Country	0,29	3,51	0,22	0,66	0,01	0,04	7,69	0,21	2,39	0,26	1,06	0,2	3,33	0,01
To the Man on the Trail	0,31	3,18	0,27	0,65	0,01	0,04	7,21	0,19	2,75	0,31	1,31	0,18	3,19	0
The Priestly Prerogative	0,26	3,85	0,21	0,64	0,02	0,06	4,63	0,18	3,96	0,28	1,05	0,22	2,31	0,01
The Wisdom of the Trail	0,29	3,41	0,23	0,65	0,02	0,05	7,79	0,2	3,05	0,38	1,09	0,2	3,55	0
The Wife of a King	0,32	3,14	0,37	0,6	0,01	0,05	9,72	0,21	3,18	0,4	1,51	0,19	3,13	0,03
An Odyssey of the North	0,17	5,87	0,1	0,57	0,01	0,08	5,19	0,23	3,2	0,52	1,43	0,15	3,57	0
Average value in the corpus	0,27	3,81	0,23	0,63	0,01	0,05	6,87	0,20	3,32	0,34	1,23	0,19	3,11	0,01

3.4. Quantitative characteristics of the translations of ‘Sun of the Wolf’ collection

On the next stage of the research the similar calculations for the translated stories of Jack London have been made

Table 4
General quantitative characteristics of translated texts

Story	Number of word usage	Number of word forms	Number of words	Hapax legomena for word forms	Number of word forms used 10 or more times	Hapax legomena for lemmas	Number of lemmas used 10 or more times	The number of characters of the extended alphabet in the text	The number of sentences in the text
The White Silence	3083	1675	1221	1324	32	845	49	1551	266
The Son of the Wolf	4911	2404	1566	1826	49	943	74	24556	385
The Men of Forty-Mile	2683	1399	1014	1060	29	637	40	12986	240
In a Far Country	4998	2612	1978	2078	51	1322	61	25575	397
To the Man on the Trail	2973	1601	1208	1247	27	832	34	14902	200
The Priestly Prerogative	3983	1994	1321	1575	40	823	58	18953	235
The Wisdom of the Trail	2440	1307	964	1017	28	653	37	12170	179
The Wife of a King	4751	2339	1588	1790	53	297	64	24329	331
An Odyssey of the North	7191	2319	2319	1630	81	1630	81	31980	747
Total:	**37013**	**17650**	**13179**	**13547**	**390**	**8657**	**498**	**167002**	**2980**

Based on the morphological corpus markup of translated stories the frequency of each part of speech in the text and in the author’s register has been automatically calculated.

The most frequent words in the target text are the functional parts of speech which have 5.74% of the vocabulary and 25.98% of the texts. High frequency is also shown by pronouns (3.07% of the vocabulary and 12, 34% of the text); adverbs (9.06% of the vocabulary and 11.19% of the text) and numerals (1.3% of the vocabulary).

Nouns (25.23% and 33.13%), verbs (18.64% and 31.10%) and adjectives (7.42% and 14.45%) in the text and in the writer's vocabulary prove their stylistic function and their ratio shows the nominality of J. London’s individual style.

Based on the defined general quantitative characteristics of the text and indicators of distribution by part of speech, the indices that characterize the lexical level of the corpus have been calculated (Table 6).

After a series of calculations, it has been found that the richness of the vocabulary i.e. the ratio of the volume of the vocabulary of tokens on average in the corpus is 0.37. Therefore, the richness of the vocabulary of translated stories can be considered high.

The average repetition of a word in the text is 2.77, i.e. each word in the text is used about three times.

The index of exclusivity is calculated for the vocabulary and the text is 0.65 and 0.39, respectively, which indicates a high artistic level of translated texts.

The opposite of the index of exclusivity is the index of concentration of vocabulary and text, which is 0.04 and 0.01, respectively.

The index of the lexical density of translated stories by J. London is 0.2, so the functional words have about 20%, which indicates a sufficient density of the text.

The automatic readability index is also important for our study, because the higher the ARI, the more difficult to understand the text is. ARI of translated stories by J. London is 8.04, which means that these works correspond to the eighth level of complexity, they are easy to understand, but not primitive.

Indices of epithetization, nominality and verb phrases, which are shown in Fig. 10 indicate a high level of epithet-phrases, verb phrases, low "clutter" of speech and high logical coherence.

3.5. Comparative quantitative characteristics of the source and target texts

The analysis of J. London's collection of short stories "The Son of the Wolf" has been conducted on the basis of the digital marked corpus of source and target texts; it covers a number of general characteristics (ST stands for source text; TT is for target text).

Table 6
General quantitative characteristics of the source and target texts (word usage, word forms, lemmas)

Story	Number of word usage		Number of word forms		Number of words	
	ST	TT	ST	TT	ST	TT
The White Silence	3733	3083	1326	1675	1072	1221
The Son of the Wolf	6114	4911	1785	2404	1471	1566
The Men of Forty-Mile	3156	2683	1185	1399	949	1014
In a Far Country	6239	4998	2032	2612	1779	1978
To the Man on the Trail	3139	2973	1223	1601	986	1208
The Priestly Prerogative	4703	3983	1487	1994	1223	1321
The Wisdom of the Trail	2988	2440	1027	1307	875	964
The Wife of a King	4911	4751	2404	2339	1566	1588
An Odyssey of the North	10669	7191	1818	2319	1818	2319
Totally	**45652**	**37013**	**14287**	**17650**	**11739**	**12179**

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

Table 7
Quantitative indicators of the lexical level of translation

Story	The richness of the vocabulary	The average repetition of the word in the text	Index of exclusivity for the text	Index of exclusivity for the vocabulary	Index of text concentration	Index of vocabulary concentration	Automatic readability index	Lexical density index	Index of nouns phrases	Index of verb phrases	Index of nominality	Verb index	logical coherence index	Speech embolism index
The White Silence	0,4	2,52	0,43	0,69	0,01	0,04	7,36	0,2	3,98	0,49	1,39	0,19	2,37	0,04
The Son of the Wolf	0,32	3,14	0,37	0,6	0,01	0,05	8,5	0,21	3,18	0,4	1,51	0,19	2,63	0,03
The Men of Forty-Mile	0,38	2,65	0,4	0,63	0,01	0,04	6,96	0,19	3,54	0,47	1,26	0,2	2,17	0,05
In a Far Country	0,4	2,53	0,42	0,67	0,01	0,03	8,97	0,21	3	0,49	1,39	0,19	2,59	0,05
To the Man on the Trail	0,41	2,46	0,42	0,69	0,01	0,03	9,61	0,21	3,67	0,46	1,46	0,18	3,06	0,04
The Priestly Prerogative	0,33	3,02	0,4	0,62	0,01	0,04	7,88	0,19	3,8	0,49	1,28	0,19	2,89	0,05
The Wisdom of the Trail	0,4	2,53	0,42	0,68	0,01	0,04	8,88	0,2	3,2	0,5	1,28	0,19	2,79	0,04
The Wife of a King	0,33	2,99	0,38	0,61	0,01	0,04	9,87	0,22	2,8	0,41	1,48	0,18	3,1	0,04
An Odyssey of the North	0,32	3,1	0,23	0,7	0,01	0,03	4,33	0,28	4,15	0,6	1,18	0,18	2,69	0,04
Average value in the corpus	0,37	2,77	0,39	0,65	0,01	0,04	8,04	0,21	3,48	0,48	1,36	0,19	2,70	0,04

The most frequent words in the source and target texts are the functional words (6% of the source text vocabulary and 7.47% of the target text vocabulary). Also, functional words are the most frequently used in texts (29.91% in the source text and 24.11% in target text). Share of nouns and the verbs are 23% and 25% in the source texts and 18% and 19% in the target text, respectively. Pronouns are also frequently used in the texts (3.11% in the source text vocabulary and 3.04% in the target text vocabulary). The share of pronouns in the texts is about 13%. Approximately the same share in the text and the vocabulary goes to adverbs (7.22% and 7.17% in the source text and 10.13% and 10.08% in the target text) and numerals (1.12% and 1.07 in the source text and 1.36% and 1.06% in the target text).

Figure 8

Part of the speech frequency of the vocabulary and the source and target texts

Part of speech	Word usage		Words	
	Source text	Target text	Source text	Target text
Noun	10415	9340	4899	4982
Verb	8397	6901	2940	4502
Adjective	3254	2746	1922	2162
Adverb	3098	3024	930	1604
Pronoun	5505	4566	269	523
Numeral	501	491	501	185
Functional words	14336	9616	468	819
Totally	45506	36684	11929	14777

Nouns, verbs and adjectives are the most frequently used. Their relative number in the vocabulary, on the contrary, exceeds the relative number in both source and target texts.

These parts of speech present the richness of the vocabulary of the source and target texts, and also their ratio confirms that the nominal character of the individual style of the original text was preserved in the translated text.

Quantitative relations between parts of speech have been compared, as they are considered key elements of the statistical characteristics of the text are presented below.

Table 9

Quantitative relations between parts of speech

Index	Mean value on the corpus of the source text	Mean value on the corpus of the target text
Lexical density index	0.20	0.21
Index of nominal phrases	3.32	3.48
Index of verb phrases	0.34	0.48
Nominality index	1.23	1.36
Verb index	0.19	0.19
Logical coherence index	3.11	2.70
Embolism index	0.01	0.04

4 Conclusions

According to these criteria, a collection of short stories by J. London "The Son of the Wolf" and its translation into Ukrainian has been selected for further analysis.

To conduct quantitative research, the AntConc program environment i.e. a stationary, index-free concordance has been used. The technical advantages of this program include free access on the Internet, free upgrades and compatibility with three operating systems - Windows, Mac OS X and

Linux. Processing the entered data AntConc sorts the selected language units according to the criteria set by the user, namely:

- frequency;
- ending of the word;
- alphabetical order.

The program also allows to select the number of words to the right and left of the search word displayed in the program window, get the attributes of a given word in alphabetical order, search for collocations by constructing n-grams of different lengths, and compare keywords in different text corpora.

At the next stage of the study, the results of the preliminary processing of the texts of the stories have been transferred to the MS Excel environment, where each part of speech its lemma and the number of uses in source and target texts have been identified.

In the study, the priority has been given to the lexical content of source and target texts, and with the help of the automatic processing of the corpus and statistical calculations, several important characteristics that are basic for clarifying the idiosyncrasy of the writer and which help to conclude the aesthetic significance and their equivalence have been identified.

The paper presents a statistical analysis of selected works by Jack London and their translation into Ukrainian, which have been compared; some tables and diagrams, which identify features of functional language styles and features of the author's style, have been made.

Summing up the quantitative study of the collection "The Son of the Wolf", it should be noted:

- indices of vocabulary richness, exclusivity for the text and the vocabulary, the concentration of the vocabulary do not differ significantly;
- the most frequent in the target text are functional words (24.91% in the source and 24.11% in the target text). The share of nouns and verbs is approximately 23% and 25% in the source and 18% and 19% in the target texts, respectively. The share of pronouns is comparatively big – about 13% of the text. Approximately the same share in the text and the vocabulary is covered by adverbs (7.22% and 7.17% in source and 10.13% and 10.08% in target text) and numerals (1.12% and 1.07 in source and 1.36% and 1.06% in target text).
- the epithetization index indicates the number of nouns per adjective in the text, i.e. the higher the index, the fewer adjectives per noun. It may be concluded that the index both in source and target text does not differ significantly – 3.32 / 3.48, and therefore the translator managed to preserve the saturation of the text with figurative phrases.
- the index of verb phrases indicated the number of adverbs per verb. The target text has a slightly bigger ratio – 0.48 adverbs per verb, while in the source text the ratio is 0.34 per 1.
- the degree of nominality shows the number of nouns per verb, in the original text there are 1.23 nouns per verb, in the translated text - 1.36 per 1. Therefore, the degree of aggression is equal in source and target text. This confirms the fact that the nominal character of the original text is accurately reproduced in translation.

5 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.

6 References

[1] T.J.M. Sanders, V. Demberg, J. Hoek, M.C.J. Scholman, F.T. Asr, S. Zufferey, J. Evers-Vermeul, Unifying dimensions in coherence relations: How various annotation frameworks are related, Corpus Linguistics and Linguistic Theory, 2021, 17 (1), pp. 1-71.

[2] Mosavi Miangah, Tayebeh. Different Aspects of Exploiting Corpora in Language Learning. Journal of Language Teaching Research, 2012, 3, 1051-1060.

[3] M.N. Wróblewska, Research impact evaluation and academic discourse. Humanities and Social Sciences Communications, 2021, 8 (1), art. no. 58.

[4] E. Aarden, L. Marelli, A. Blasimme, The translational lag narrative in policy discourse in the United States and the European Union: a comparative study. Humanities and Social Sciences Communications, 2021, 8 (1), art. no. 107.

[5] F. Cifuentes-Silva, J.E. Labra Gayo, Legislative document content extraction based on semantic web technologies: A use case about processing the history of the law. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11503 LNCS, 2019, pp. 558-573.

[6] C. Fang, S. Zhang, Geographic information retrieval method for geography mark-Up language data. ISPRS International Journal of Geo-Information, 2018, 7 (3), art. no. 89.

[7] L. Xiao, D. Wei, Linguistic Question-Answering Reasoning Based on Intelligent Perception of Attribute Weight. Journal of Physics: Conference Series, 2021, 1883 (1), art. no. 012137.

[8] C. Chantrapornchai, A. Tunsakul, Information extraction on tourism domain using SpaCy and BERT. ECTI Transactions on Computer and Information Technology, 2021, 15 (1), pp. 108-122.

[9] J.S.N. Rodrigues, K. Ferreguetti, A.S. Pagano, A proposal of coextensiveness between technical term, nominal group, and lexical item in Brazilian Portuguese: A study based on corpus linguistics' software within the framework of systemic-functional theory [Uma proposta de coextensividade entre termo tècnico, grupo nominal e item lexical no Português Brasileiro: Um estudo com base em ferramentas da linguística de corpus sob o arcabouço de teoria sistêmico-funcional] Revista de Estudos da Linguagem, 29 (2), pp. 1325-1379, 2021.

[10] S. Rezaei, D. Kuhi, M. Saeidi, Diachronic corpus analysis of stance markers in research articles: The field of applied linguistics. Cogent Arts and Humanities, 2021, 8 (1), art. no. 1872165.

[11] K. Uzule, Teacher training and education programs in latvia: Are e-competences included? Business Management and Education, 2020, 18 (2), pp. 294-306.

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

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

[14] I. Tsmots, V. Teslyuk, A. Batyuk, V. Khavalko, A. Mladenow, Information-analytical support to medical industry. CEUR Workshop Proceedings, 2019, 2488, pp. 246-257.

[15] V. Vasyliuk, Y. Shyika, T. Shestakevych, Modelling of the Automated Workplace of the Psycholinguist. 2020 IEEE 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2020 - Proceedings, 1, art. no. 9321956, pp. 276-279.

[16] B. Tahir, M.A. Mehmood, Corpulyzer: A Novel Framework for Building Low Resource Language Corpora. IEEE Access, 9, 2021, art. no. 9316706, pp. 8546-8563.

[17] P. Stiles, Beowulf 33a and Hapax Legomena. Neophilologus, 2020, 104 (2), pp. 255-261.