https://doi.org/10.28925/2412-2491.2024.2210
UDC 81`33

CORPUS-BASED TECHNOLOGIES IN THE TRANSLATION STUDY OF TONIC DRINK ADVERTISEMENTS: AN ALGORITHM OF ANALYSIS PROCEDURE

Nohovska S.H.
Borys Grinchenko Kyiv Metropolitan University,
ORCID iD https://orcid.org/0000-0001-8608-6172
s.nohovska@kubg.edu.ua

The study is aimed at the procedure used for the analysis of English-language advertising texts for tonic drinks and the peculiarities of their reproduction in Ukrainian using a corpus-based approach. For the research material, 500 advertising texts for the same products of the specified group were selected in English and Ukrainian, since the material of any study involving corpus-based technologies should be chronological and thematically homogeneous.

The proposed analysis algorithm involves a number of sequential steps. Firstly, to ensure a multi-channel method of collecting material. Secondly, the creation of a research parallel corpus of English-language advertising texts for tonic drinks and their translations into Ukrainian using the InterText Editor program. Thirdly, the annotation of corpus from the perspective of its composition and lexical and semantic structure, i.e., the compositional components of advertising texts, their obligatory and variable structural elements, as well as speech figures of expression and nationally marked vocabulary were identified with the help of special tags. Fourthly, using corpus managers AntConc, Textanz, and MS Excel, determination of general quantitative characteristics of the corpus and subcorpora, compilation of frequency lists of word usage, lemmatization, and calculation of partial-word distribution in the source and target text sets and structural elements of advertising messages. Fifthly, conducting a systematic inventory of the units under study, explaining the peculiarities of their formation and functioning, and characterizing the verbal content of advertising texts. Sixthly, an analysis of the morphological, lexical, semantic, and syntactic levels of the language of tonic drink advertisements in the source and translated texts, as well as stylistic devices peculiar to these texts. Seventhly, a systematic consideration of the interaction of verbal, visual, and auditory components, since the transition from one system of signs to another is very important when adapting advertising messages created in one cultural environment to another culture.

Key words: *advertising text, translation analysis, corpus-based approach, parallel research corpus, quantitative characteristics of the research corpus.*

Ноговська С. Г., Корпусні технології у перекладознавчому дослідженні реклами тонізувальних напоїв: алгоритм методики аналізу

Стаття присвячена дослідженню методики аналізу англомовних рекламних текстів тонізувальних напоїв та особливостей їх відтворення українською мовою на основі корпуснобазованого підходу. Для матеріалу дослідження обрано по 500 рекламних текстів до тих самих товарів окресленої групи англійською та українською мовами, що забезпечило хронологічно та тематично однорідний матеріал, релевантний для застосування корпусних технологій.

Запропонований алгоритм аналізу передбачає ряд послідовних кроків. По-перше, забезпечення багатоканального методу збору матеріалу. По-друге, побудову дослідницького паралельного корпусу англомовних рекламних текстів тонізуючих напоїв та їх перекладів українською мовою за допомогою програми InterText Editor. По-третє, маркування корпусу в структурному та лексико-семантичному плані, тобто за допомого спеціальних тег виділення композиційні складові рекламних текстів, їх обов'язкові та варіативні структурні елементи, а також мовленнєві фігури експресивності та національно марковану лексику. По-четверте, встановлення за допомогою корпусних менеджерів AntConc, Textanz та програми MS Excel загальних кількісних характеристик корпусу та підкорпусів, укладання частотних списків слововживань, проведення лематизацію та підрахунок частиномовного розподілу у текстових масивах оригіналу та перекладу та структурних елементах рекламних повідомлень. По-п'яте, проведення планомірної інвентаризації досліджуваних одиниць, пояснення особливостей їх формування та функціонування, характеристики вербального наповнення рекламних текстів. По-шосте, аналіз морфологічного, лексико-семантичного, синтаксичного рівнів мови реклами тонізувальних напоїв в оригіналі та перекладі, а також стилістичних засобів, характерних для цих текстів. По-сьоме, системний розгляд взаємодії вербальних, візуальних та аудіальних складників, адже перехід від однієї системи знаків до іншої дуже важливий під час адаптації рекламних повідомлень, створених в одному культурному середовищі, для іншої культури.

Ключові слова: *рекламний текст, перекладознавчий аналіз, корпусний підхід, паралельний дослідницький корпус, квантитативні характеристики дослідницького корпусу*

Introduction. Modern linguistics is oriented towards the use of computer technologies, which, on the one hand, facilitate the possibility of investigating large (in particular for linguists) bodies of textual information, and on the other hand, make it impossible for the researcher to make subjective assessments. Corpus linguistics enables an objective and systematic analysis of the research object. Corpora can be used to conduct research that is methodologically different from traditional research.

Our research aims to outline the methodology of a corpus-based study of linguistic stylistic and linguistic statistical features of English-language advertisements for tonic drinks (such as tea, coffee, lemonades, soft drinks, energy drinks, etc.) focused on analyzing the peculiarities of their reproduction in the Ukrainian language.

The stated goal involves solving the following specific tasks, such as:

- to construct a research parallel corpus of advertising texts;
- to mark the corpus according to structural and lexical-semantic criteria;
- to determine the quantitative characteristics of the lexical level of the source and target texts;
- to establish the significance/insignificance of the statistical difference between the values of the coefficients for the source and target texts.

Theoretical Background. Corpus-based studies hold an important place in world linguistics. Leading scholars in this field include G. Leech (Leech, 1992), D. Biber (Biber, 1999), J. M. Sinclair (Sinclair, 1991), S. Th. Gries (Gries, 2016), S. Granger (Granger, 2004), T. McEnery (McEnery, 2008), M. Baker (Baker, 1992), A. Hardie, M. Baker P. (Hardie et al., 2006), and others. Among the Ukrainian researchers, notable contributions have been made by S. Buk (Buk, 2009), N. Darchuk (Darchuk, 2010), O. Demska (Demska, 2022), V. Zhukovska (Zhukovska, 2021), V. Shyrokov (Shyrokov et al., 2005), I. Kulchytskyi (Kulchytskyi, 2015), and others.

The key topics of interest in this field can be broadly categorized into several areas, including:

- an analytical review of discussions and foreign publications regarding the place of corpus technologies and corpus linguistics in modern linguistics;
- an overview of corpus linguistics and the history of its formation;
- a discussion of what a corpus of texts involves, its defining features, approaches to the classification of corpora as well as the branches and methods of their use;
- the concept of the national text corpus, its prerequisites and principles of planning and compilation;
- some aspects of the creation and use of specific research corpora, and the technical aspects of preparing texts for further corpus research has been reviewed.

Methods. Modern linguistics, through its interaction with a number of scientific disciplines, operates with a variety of methods and paradigms, which, however, do not negate each other but rather cooperate and integrate. These processes are clearly manifested in the linguistic analysis of advertising, which involves the use of various methodological directions and approaches at different stages of work.

The dominant approaches used to analyze the structure, semantics and functionality of advertising texts are both linguistic-stylistic and linguopragmatic.

The linguistic-stylistic approach involves the study of the language used in an advertising message, in particular, the stylistic figures used to create advertising texts, since advertising not only informs the reader but also forms a vivid, clear advertising image through the system of visual and expressive means of language (Areshenkova, 2014). Therefore, the linguistic-stylistic approach is mainly used to analyze print advertising and verbal components of other types of advertising products, that is, it is focused on the interpretation of advertising as a text rather than a discourse. In other words, in addition to the verbal aspects, attention is also paid to the advertising image and graphic representation of the advertising messages, i.e. composition, font selection and color.

The linguopragmatic approach is based on the theory of speech acts proposed in the middle of the last century by J. Austin (Austin, 1962) and his followers. Most modern researchers of advertising linguopragmatics focus on various aspects of influence, on strategic or axiological parameters of advertising discourse.

Another important area of linguistic research in advertising, which requires a comprehensive analysis based on the synthesis of various methods and techniques, is based on the translation studies approach. Fewer scientific works have been found in this direction than in the strictly linguistic one, with comparative studies prevailing.

Modern studies of advertising translation issues (K. L. Smith (Smith, 2002), I. Torresi (Torresi, 2021), B. P. Faber (Faber, 2012), D. Dobrovolska (Dobrovolska, 2016), etc.) substantiate translation strategies for advertising based on the model of functional translation, since the original text of an advertising message must be adapted to the mental, social, and cultural properties of native speakers of the language into which the text is translated, potential consumers of the advertising product. These works suggest that when translating advertising texts, it is necessary to take into account, first of all, the pragmatic potential of the text, its influential effect, as well as the cultural specifics of the environment where the advertising message will function, such as history, cultural traditions, national stereotypes, etc.

The translation approach to advertising analysis requires the use of a relevant framework. We believe that when comparing the original and the translation of an advertising communication in a particular direction (in our case, English-Ukrainian) and a particular product segment (in our case, tonic drinks), lexical-semantic, linguistic-stylistic and linguistic-statistical analysis of the verbal component of the advertising message using the corpus-based method with further consideration of all visual, audio and other paralinguistic components will help to clarify the features and effectiveness of translation strategies.

Results and Discussion. The aim and objectives necessitate the development of a relevant methodology, which should be based on a detailed description of the research procedure with the specification of the methods used.

The first stage of the research involves the formation of a research corpus. For this purpose, several procedures were performed in the following sequence:

- extraction of authentic advertising messages about tonic drinks from Internet sources in the format of videos, banners, posters, creolized and traditional texts;
- identification advertisement's verbal component;
- systematization and division of advertising texts into groups by type of advertised product (tea, coffee, soft drinks, juices, etc.).

The material was collected by the method of comprehensive sampling from the

- official websites of the companies producing the selected group of advertised products,
- official brand pages on Facebook and Instagram,
- the online application Pinterest, YouTube video hosting, as well as
- publications in specialized online blogs (Your Coca-Cola, Tea-and-Coffee.com, Coffee & Tea Blogs, Coffee Tea Club, etc.)

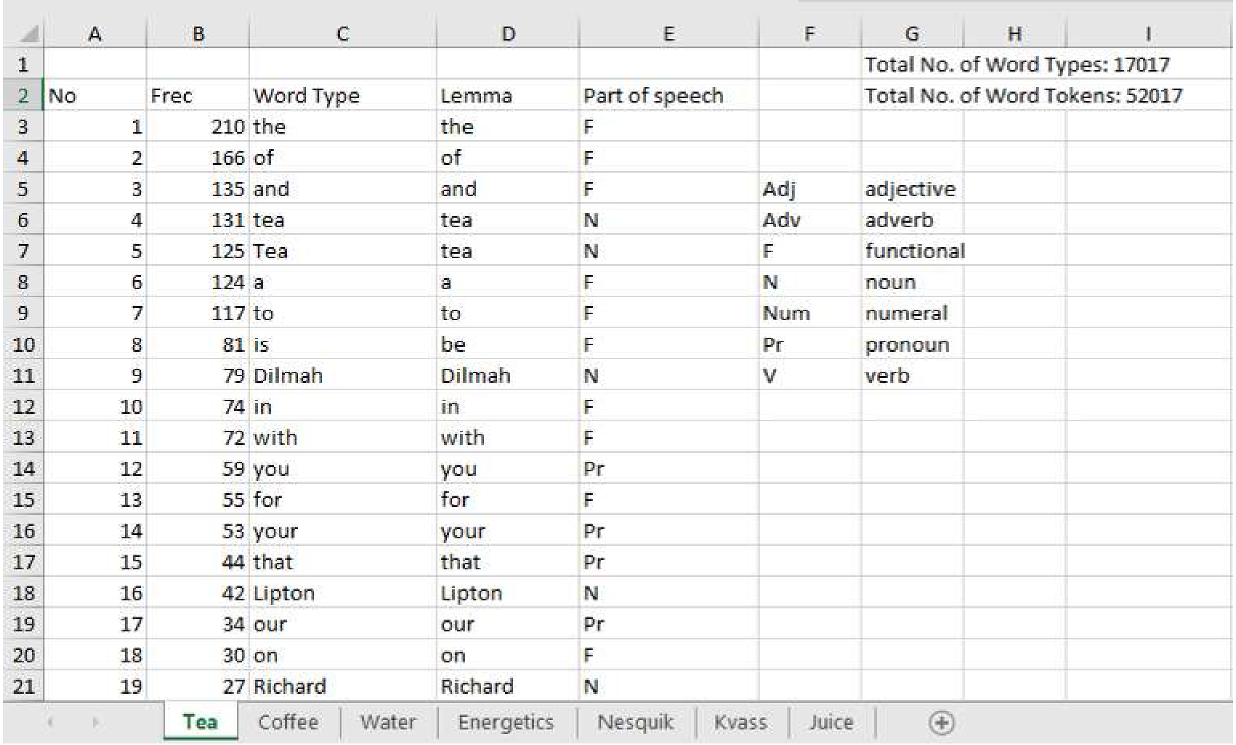
The multichannel method of collecting material is optimal for our research subject because advertising communication involves the integrated use of both verbal and non-verbal elements in advertising messages. The choice of such material allows us to consider both linguistic and extra-linguistic aspects and involves a multimodal approach in its analysis.

For the research material, 500 advertising texts for the same products of the specified group were selected in English and Ukrainian, since the material of any study involving corpus-based technologies should be chronological and thematically homogeneous.

At the second stage of the research, the text data was transformed into a parallel-aligned bilingual corpus using the InterText Editor software. The corpus was annotated from the perspective of its structure and lexical-semantic composition. Special tags were utilized to highlight the compositional components of advertising texts (such as nominative-representative, attractive-appeal, intentional-axiological, descriptive blocks). The obligatory structural elements (such as slogan, headline, and the advertising text itself) and optional elements (if

any – echo phrase, subheading, coda) were identified. In addition, language figures of expressiveness and nationally marked vocabulary were annotated. The following marking system has been used:
<head>…</head> – headline,
<adtx>…</adtx> – information block of the advertising post,
<s>…</s> – slogan,
<eph>…</eph> – echo phrase,
<coda>…</coda> – coda etc.

Using corpus managers like AntConc, Textanz, and the MS Excel program, we established general quantitative characteristics of the corpus and sub-corpora, generated frequency lists of word usage, conducted lemmatization, and calculated the distribution of parts of speech within the original texts and translations, as well as the structural elements of advertising messages. A fragment of lemmatization and the partial speech distribution are presented at *figure 1* and *2*.



	A	B	C	D	E	F	G	H	I
1							Total No. of Word Types: 17017		
2	No	Frec	Word Type	Lemma	Part of speech		Total No. of Word Tokens: 52017		
3	1	210	the	the	F				
4	2	166	of	of	F				
5	3	135	and	and	F	Adj	adjective		
6	4	131	tea	tea	N	Adv	adverb		
7	5	125	Tea	tea	N	F	functional		
8	6	124	a	a	F	N	noun		
9	7	117	to	to	F	Num	numeral		
10	8	81	is	be	F	Pr	pronoun		
11	9	79	Dilmah	Dilmah	N	V	verb		
12	10	74	in	in	F				
13	11	72	with	with	F				
14	12	59	you	you	Pr				
15	13	55	for	for	F				
16	14	53	your	your	Pr				
17	15	44	that	that	Pr				
18	16	42	Lipton	Lipton	N				
19	17	34	our	our	Pr				
20	18	30	on	on	F				
21	19	27	Richard	Richard	N				

Figure 1. A fragment of lemmatization

To determine the reliability of the study results, it is necessary to establish whether the sample size will ensure the reliability of the calculations. In our research, to determine the "reliability" of the sample size (when working with morphological and lexical-semantic characteristics of tonic drinks advertising texts), we resorted to the use of the so-called sampling error formula:

$$\delta = \frac{Z_p}{\sqrt{N \cdot p}} \qquad (1),$$

where Z_p is a constant value, which for a five percent significance level is equal to 1.96; N is the sample size (in absolute terms, for example, in the number of word uses); p is the relative frequency of use of the units of analysis (for example, slogans or tropes).

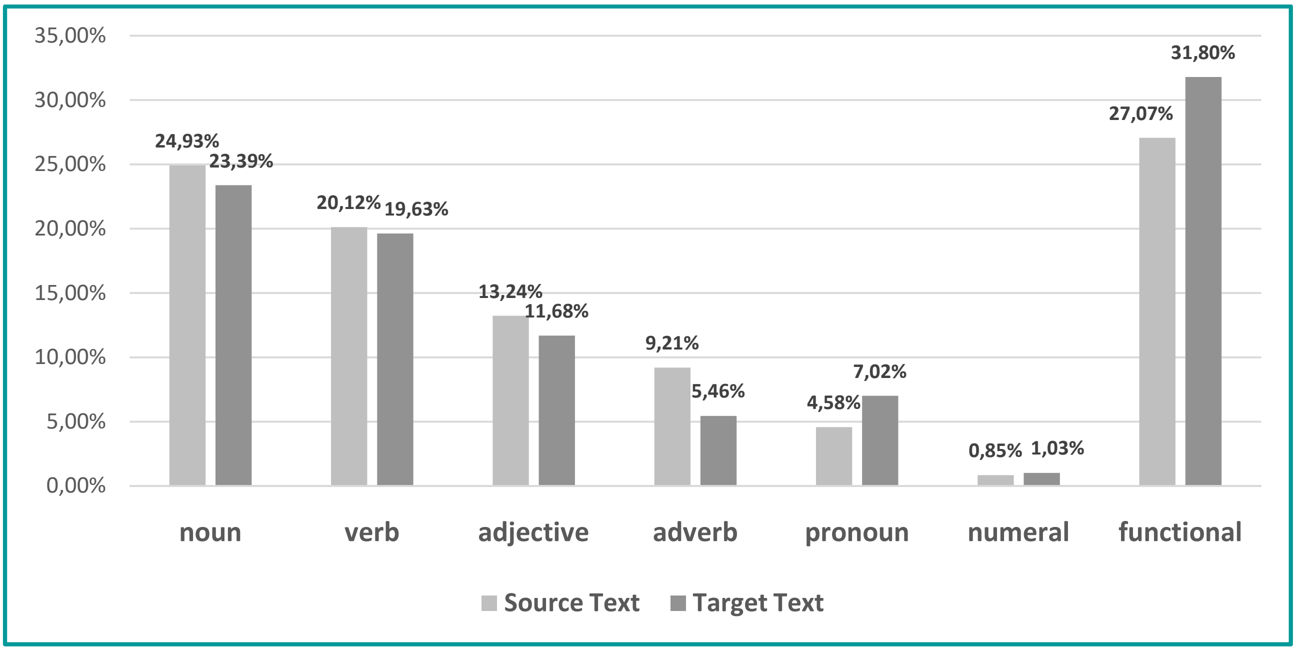


Figure 2. Partial speech distribution

For instance, in the ‘Coffee’ sub-corpus, 36,561 word occurrences were recorded. According to the formula (the formula is not provided on the slide), the error rate for a noun is 0.06%; for an adjective, 0.1%; for a verb, 0.21%; for a pronoun, 2.1%; and for an adverb, 2%. That is, for all the analyzed parts of speech, the sample is sufficient for the reliability of the study results.

At the third stage of the research, several procedures were conducted to systematically inventory the units under study, as well as the explanation of peculiarities of their formation and functioning, and characterize the verbal content of advertising texts. This technique involved:

– identification of the units of analysis;
– classification and interpretation of the selected units.

The units of analysis were the main and optional structural elements of the advertising texts of tonic drinks in the source and target texts, such as: slogan, headline and the advertising text itself (main advertising text), which were further classified. For example:

Lacomba Arabica do Brasil 100% **2)**
Grown at haciendas in the state of (Minas Geraes), Arabica Santos is considered one of the most popular varieties of South America. Strong coffee with a soft rounded aftertaste, thick and rich, will bring you relaxation of the siesta, immerse you in dreams of warm countries, gives you the pleasure to feel the taste.
*Dark as a Brazilian woman's skin. Round like her figure. Soft as her look…***3)**
Arabica do Brasil Santos – the embodiment of Brazil in a cup **1)**.
where **1)** – slogan, **2)** – headline, and **3)** – main advertising text.

After identifying the structural components in all texts, they were classified according to certain characteristics. For instance, slogans were classified according to their formal characteristics, the purpose of expression, communicative intent, and direction/intention; the main advertising text was classified according to its structural characteristics, communicative strategies, and models.

During the fourth stage of the research, the advertising texts for tonic drinks were analyzed on the morphological, lexical-semantic, and syntactic levels both in the source and target versions. In addition, stylistic devices characteristic of these texts were examined. A comparison and contrast of the analysis results were performed for various compositional elements of the advertising text (slogan, headline, main advertising text) in the SL and TL.

The identified predominant parts of speech in the previous stages allow for a lexical-semantic analysis of lexemes in the studied advertising texts and the categorization of lexemes into lexical-semantic fields (such as smell, taste, action/process, feelings/emotions, etc.).

At the fifth stage, a comprehensive approach was applied to the translation studies analysis of advertisements. This involved a systematic examination of the interaction between verbal, visual, and auditory elements because transitioning from one system of signs to another is crucial during the adaptation of advertising messages created in one cultural context for another culture. It was determined that information in advertising for tonic drinks is primarily presented linguistically and visually. Every advertising text utilizes at least one type of visual element, including illustrations, photos, or images, as all analyzed advertising texts contain either images of the advertised product, a logo, or a photograph of a model (a famous person or a consumer verifying the effectiveness of the advertised product).

The sixth stage involved identifying translation strategies and specific tactics used in reproducing both compositional components and lexical-semantic features of advertising texts. A quantitative analysis of differences between both the source and target texts based on various linguistic characteristics was conducted, thus comparing the invariant of SL and TL. Our research methodology relies not on a singular criterion but on a combination of several criteria, aiming to yield reliable and objective results while minimizing the possibility of ambiguous conclusions regarding the quantitative equivalence of the source and target texts.

To research the quantitative correspondence between the source and target texts, criteria were chosen by well-known linguists and translation scholars (such as J. McMenamin (McMenamin, 2003), K. Ponnamperuma, K. Mellish, P. Edwards (Ponnamperuma K. et al., 2012), S.Buk (Buk, 2008), S. Zasiekin (Zasiekin, 2016), I. Kulchytskyi (Kulchytskyi, 2019), L. Tsiokh (Tsiokh, 2019), and others):

1) general quantitative characteristics of the source and target corpora: word usage number, number of word forms, number of lemmas, number of sentences;

2) lexical diversity index;

3) coefficient of syntactic diversity;

4) lexical density coefficient, or the percentage of function words;

5) automatic readability index;

6) coefficient of logical coherence.

Some results for the ‘Coffee’ sub-corpus are presented in the *table 1* and *figure 3*.

Table 1. General characteristics of the research corpus

Index	**ST**	**TT**
Number of word usages	4961	3006
Number of word forms	4041	2789
Number of lemmas	1757	1003
Number of Hapax legomena	709	631
Number of words with frequency >=10	213	167
Number of sentences	618	505
Number of symbols	28100	23281

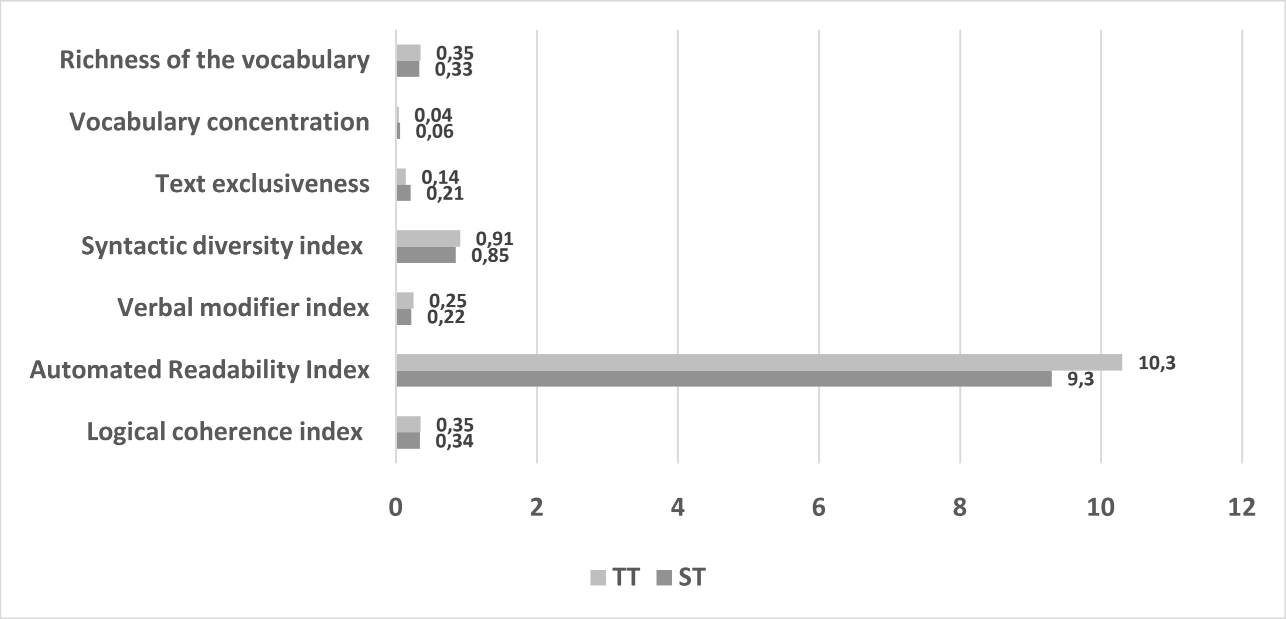


Figure 3. The main quantitative characteristics of the research corpus

It has been found that the deviation between the quantitative characteristics of English and Ukrainian advertising texts for tonic drinks is quite high (ranging from 17.32% to 54.76% for various indexes). This variance can be explained by the application of translation strategies such as localization and transcreation. Thus, only partial quantitative equivalence between the source and target texts can be established.

Conclusions and perspectives. Thus, the proposed research procedure can provide a multifaceted approach and a comprehensive study of tonic advertising texts, their features, functioning and methods of reproduction, taking into account the semantic formations of a particular language and national specificity at the levels of verbal and paralinguistic objectification of cultural and social realities.

The comprehension of linguistic and quantitative characteristics of advertising texts for tonic drinks based on corpus technologies, as well as the peculiarities of their reproduction in translation, shows the complex interrelationships of the elements of the lexical level of this type of social and intercultural communication and outlines the criteria of translation adequacy.

Quantitative analysis is a far from complete linguistic description of advertising discourse, but it makes possible to identify the most frequent elements of the text, show significant correlations between different indicators of text levels, calculate indicators characterizing the informative or emotional saturation of the text, and thus outline a translation strategy that will allow to convey its genre and style specificity with the least losses.

The logical next step in this study will be a qualitative analysis of the data obtained, which can be done within the framework of a lexical and semantic study of the stylistic dominants of advertising communication and their correspondences in translation.

REFERENCES

1. Areshenkova, O. (2014). Reklamnyi tekst yak funktsionalnyi riznovyd movlennya. *Filolohichni studiyi, 10*, 5–11. Doi: https://doi.org/10.31812/filstd.v10i0.407 (in Ukrainian)
2. Austin, J. L. (1962). How to do things with words. Retrieved December 20, 2023, from: https://ia801306.us.archive.org/23/items/HowToDoThingsWithWords.pdf
3. Baker, M. (1995). Corpora in Translation Studies. *Target. International Journal of Translation Studies*, *vol. 7* (2), 223-243. Doi: https://doi.org/10.1075/target.7.2.03bak.
4. Biber, D., Conrad, S., Reppen, R., & Leech, G. (1999). Corpus linguistics : Investigating language structure and use. *International journal of corpus linguistics, Vol. 4* (1), 185-188. Doi: 10.1075/ijcl.4.1.11lee
5. Buk, S. (2008). *Osnovy statystychnoyi linhvistyky*. Vydavnychyi tsentr LNU imeni Ivana Franka. (in Ukrainian)
6. Buk, S. (2009). Strukturne anotuvannia u korpusi tekstiv (na prykladi prozy Ivana Franka). *Ukrainska mova, vol. 3*, 59–71. (in Ukrainian)
7. Darchuk, N. (2010). Doslidnytskyi korpus ukrainskoi movy: Osnovni zasady i perspektyvy. *Visnyk Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. Ser.: Literaturoznavstvo. Movoznavstvo. Folklorystyka*, *vol. 21*, 45-49. (in Ukrainian)
8. Demska, O. (2022) *Tekstovyi korpus: Ideia inshoi formy*. VPTs NaUKMA. (in Ukrainian)
9. Dobrovolska, D. (2016). Metodolohiya doslidzhennya perekladu reklamnoho tekstu: osnovni perekladatski stratehiyi. *Society for Cultural and Scientific Progress in Central and Eastern Europe Budapest. Science and Education a New Dimension. Philology, IV* (21), 42–46. (in Ukrainian)
10. Faber B. P. (2012). The Translation of Advertising Texts in Culturally-Distant Languages: The Case of Spanish and Arabic. *International Journal of Translation*. *Vol. 24*(1-2), 51–64.
11. Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In Connor, U. & Upton, T. (Eds.), *Applied Corpus Linguistics: a Multidimensional Perspective* (p. 123–145). Rodopi.
12. Gries, S. Th. (2016). Quantitative corpus linguistics with R: a practical introduction. Routledge. Doi: https://doi.org/10.4324/9781315746210/
13. Hardie, A., & Baker P. (2006). *Glossary of Corpus Linguistics*. Oxford University Press, Incorporated.
14. Kulchytskyi, I. (2015). Tekhnolohichni aspekty ukladannia korpusiv tekstiv. In Dani tekstovykh korpusiv u linhvistychnykh doslidzhenniakh (p. 29-45). Vyd-vo Lvivskoi politekhniky. (in Ukrainian)
15. Kulchytskyi, I. (2019). Okremi aspekty kvantytatyvnykh doslidzhen ukrayinskoyi movy. *Ukraina Moderna*, *Vol. 27*, 73-96. (in Ukrainian)
16. Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), Directions in corpus linguistics (p. 105-22). Mouton de Gruyter.
17. MCenery, T., Xiao, R. (2008) Parallel and Comparable Corpora : What is Happening? In *Incorporating Corpora. The Linguist and the Translator* (p. 18-31). Clevedon.

18. McMenamin, G. (2003). *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press.

19. Ponnamperuma, K., Mellish, C., Edwards, P. (2012). Using Distributional Similarity for Identifying Vocabulary Differences between Individuals. Computational approaches to the study of dialectal and typological variation. Retrieved March, 12, 2024, from: http://www.sfs.uni-tuebingen.de/~gjaeger/conferences/essli_2012/

20. Shyrokov, V. et al. (2005). *Korpusna linhvistyka*. Dovira. (in Ukrainian)

21. Sinclair, J. (1991). *Corpus Concordance and Collocation (Describing English Language)*. Oxford Univ Pr (Sd).

22. Smith, K. L. (2002). The Translation of Advertising Texts. Retrieved March, 2, 2024, from: https: //etheses.whiterose.ac.uk/3044/2/251329_ VOL1.pdf

23. Torresi I. (2021). *Translating Promotional and Advertising Texts*. Routledge.

24. Tsiokh, L. (2019). Linhvostatystychni metodyky v perekladoznavchomu analizi antyutopiyi. In *Social and Economic Aspects of Education in Modern Society* (p. 30-34). RS Global Sp. z O.O. (in Ukrainian)

25. Zasiekin, S. (2016). Understanding translation universals. *Babel*, *Vol. 62* (1), 122-134. Doi : http://dx.doi.org/10.1075/babel.62.1.07zas

26. Zhukovska V., Mosiiuk O. (2021). Statistical software R in corpus-driven research and machine learning. *Information Technologies and Learning Tools, 86*(6), 1-18. Doi: 10.33407/itlt.v86i6.4627.

Дата надходження статті до редакції: 11.02.2024
Прийнято до друку: 4.03.2024