UDC: 81’27
DOI: 10.32342/2523-4463-2022-2-24-14

YU.I. DEMYANCHUK
PhD in Philology, Lecturer of German language
of Foreign Languages and Translation Studies Department,
Lviv State University of Life Safety

A MODEL OF THE ENGLISH-UKRAINIAN SUB-CORPUS OF TEXTS OF NATO, UN AND WTO OFFICIAL AND BUSINESS DOCUMENTS

Метою наукового дослідження було створення моделі англо-українського підкорпусу текстів офіційно-ділових документів НАТО, ООН, СОТ (далі – АУП) як додаток до Національного корпусу української мови (далі НКУМ), в якому тексти представлені у вигляді ієрархічної структури. Для досягнення цієї мети нам довелося утворити вибірку до якої увійшли одиниці з певними ознаками офіційно-ділових документів НАТО, ООН та СОТ. Лексеми, термінологічні словосполучення, ідіоми були вилучені з текстового континууму офіційно-ділових документів. *Методи* спостереження та експерименту були застосовані для аналізу емпіричного, оглядового й концептуально-аналітичного інструментарію, що уточнює значення описуваних об’єктів і явищ з огляду на їх якісну приналежність до НКУМ. Механізм застосування нашої моделі полягає у визначенні чіткого місця тексту в усіх корпусах НКУМ, що на практиці суттєво спрощує процедуру пошуку та автоматизованого перекладу як тексту, так і, спеціальної термінології зокрема. При цьому автор пропонує збільшення кількості унікальних текстів, що забезпечать репрезентативність АУП в складі Національного корпусу української мови, а також розбиття усього процесу формування АУП на елементарні задачі, що тим самим, підвищить ефективність НКУМ.

Модель англійсько-українського підкорпусу текстів офіційно-ділових документів НАТО, ООН, СОТ – це імплементація паралельного корпусу текстів в НКУМ, що має нормативний вплив на всі підкорпуси текстів, та доповнює підходи до його розгляду. Підкорпус розглядається у вузькому сенсі, який охоплює лише тексти та термінологію міжнародних організацій НАТО, ООН, СОТ.

Результати дослідження показують, що спеціальні терміни можуть функціонувати як маркери комунікації в політичному та економічному житті, законодавстві, адміністративно-господарській діяльності, які постійно змінюються та розвиваються в реальному часі. Формування глосарію термінологічних номінацій, їхнього функціонально-стильового маркування, їхня асиміляція в загальновживаний вокабуляр АУП, згодом НКУМ перебувають на стадії формування та утвердження. Теперішнє дослідження маніфестує уявлення про АУП як підкорпусу НКУМ, що дозволяє найбільш оперативно представляти актуальні тексти офіційно-ділового стилю та охоплювати найширші кола зацікавлених читачів.

Ключові слова: Національний корпус української мови, англійсько-український підкорпус текстів, офіційно-ділові документи, термінологія НАТО, ООН, СОТ, метод спостереження, лінгвістичні експериментальні дослідження.

Для цитування: Demyanchuk, Yu.I. (2022). A model of the English-Ukrainian sub-corpus of texts of NATO, UN and WTO official and business documents. *Вісник Університету імені Альфреда Нобеля. Серія: Філологічні науки / Visnyk Universitetu imeni Alfreda Nobelya. Seriya: Filologicni Nauki*, vol. 2, issue 24, pp. 165-179, DOI: 10.32342/2523-4463-2022-2-24-14

For citation: Demyanchuk, Yu.I. (2022). A model of the English-Ukrainian sub-corpus of texts of NATO, UN and WTO official and business documents. *Alfred Nobel University Journal of Philology / Visnyk Universitetu imeni Alfreda Nobelya. Seriya: Filologicni Nauki*, vol. 2, issue 24, pp. 165-179, DOI: 10.32342/2523-4463-2022-2-24-14

Introduction

The present research is devoted to the substantiation of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO as a part of the Ukrainian National Corpus and to identification and systematization of common patterns of its structure and filling. Timeliness of the chosen topic is enhanced by the intensified scientific interest to the problem of the text corpus and its creation that has a long history of development and is manifested through certain features (machine readability, the authenticity of texts, selectivity, and representativeness), the language units of which are represented by texts of different linguistic status.

Complex and systematic development of the English-Ukrainian sub-corpus of texts is an urgent issue in view of the geopolitical vectors of Ukraine, in particular its accession to the European and world socio-cultural space, being the origin of such key concepts for our research as “text corpus”, “parallel translation”, “special applications”, as well as official and business documents of such international organizations as NATO, UN, WTO.

The obvious timeliness of our research is explained by a lack of a corpus of texts that are parallel with the Ukrainian language and is enhanced by the need to systematize and generalize empirical data that indicate a possible implementation of a sub-corpus of parallel official and business texts of NATO, UN, and WTO within the Ukrainian National Corpus. In 2009, a group of Ukrainian researchers announced an attempt to develop an English-Ukrainian corpus of official and business texts and a corpus of texts on computational linguistics [Kasianenko, Lebediev, Petrenko, 2009, c. 25–28], however, these achievements have not been presented so far [Zhukovska, 2013, c. 187].

Researche Literature review

Modern scientific studies offer a wide range of types of official and business documents, such as ordinances, resolutions, programs, decrees, acts, laws, orders, questionnaires, receipts and documents of international organizations. A high degree of standardization of speech, complete absence of emotionality and visualization, wide use of imperative and impersonal forms are taken into account. The most important classification feature of an official and business document is its content, and in particular affiliation of the information contained in it to a particular subject or area of activity. Review of sources of applied linguistics [Hlushchenko, 2010; Crowley, 2007; Czaykowska-Higgins, 2009] and corpus linguistics [Darchuk, 2013; Shvedova, 2017; Farr, 2019; Pérez-Paredes, 2010] allowed to ascertain the fact that for the purpose of formation of a sub-corpus of texts of official and business documents it is important to make allowance for an exact position of the text in the sub-corpus, aiming not only to determine the content but also to position it in a particular international organization.

The literature review found the ways of origin and reasons for settlement of the Ukrainian National Corpus first of all in corpus linguistics, applied linguistics and later on in philology. The UNC is the first attempt to apply the general theory of corpus linguistics to the Ukrainian language resulting in “a standard general language corpus compared with similar world general language text corpora of the national type” [Baloh, 2008, p. 90–93]. The modern attitude to the formation of the new sub-corpora of texts in the UNC is based on the principles of scientific and theoretical delineation of linguistic researches that allow the development of a parallel corpus of texts. Given the fact that at the level of stylistic differentiation the UNC is embedded in a seven-element system: 1) artistic, 2) scientific, 3) official and business, 4) journalistic, 5) confessional, 6) conversational and 7) epistolary styles, in our opinion namely the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO will become a logical addition to the official business style and a reliable source of international information.

For that reason, the ***purpose*** of the study is to create and prove that the model of the English-Ukrainian sub-corpus of texts of official and business documents can become a full-fledged informational foundation for the formation of a parallel corpus of texts within the Ukrainian National Corpus. To achieve this purpose we provided a possibility to trace and select from the textual continuum units with certain features of official and business documents of NATO, UN, and WTO.

Methodology

The methodological potential of our research is based on the works of well-known Ukrainian and foreign scholars (V. Bohorodytskyi, I. Baudouin de Courtenay, K. Brugman, W. Wundt, B. Delbruck, G. Paul, O. Potebnia, P. Fortunatov, W. Scherer, H. Steinthal, A. Schleicher et al.), who strengthened the scientific interest to the problem of translation and corpus of texts in particular.

Analysis of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO as a construct of the parallel corpus of texts of the UNC was carried out in several stages.

The first stage envisaged description and inventory. Theoretical works on the history of the Ukrainian National Corpus were reviewed and generalized, and the concept of "English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO" was determined. It was found out that the UNC is the object of research in such areas of linguistic studies as the applied and corpus linguistics that construe the corpus of texts as a translator's tool. This encourages us to highlight the advantages and disadvantages of the UNC and justify the expediency of our proposal to develop the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO. Definition of the UNC as a national corpus of texts that contains only Ukrainian texts and does not envisage a parallel translation evidenced the need to create a proposal of a parallel corpus of texts with an updated English-language database.

The selection of lexical items with the official and business semantics is based on documents of international organizations, such as NATO, UN, and WTO. At the same stage, we determine the purpose of the research, in particular, the formation of the English-Ukrainian sub-corpus of texts of official and business documents, which envisages a parallel translation of NATO, UN, WTO texts.

This sequence of procedures allowed us to carry out consideration and interpretation of existing theoretical provisions and inventory of factual research materials.

Theoretical and corpus analysis methods were used in the research.

The second stage envisaged analysis and differentiation. Special aspects of filling-up of the EUS with official and business texts of NATO, UN, and WTO were characterized. Theoretical analysis of the EUS was carried out in order to establish a clear hierarchy of the UNC structure. The main types of participants of the formation of the EUS (students, teaching staff) were differentiated and the behavioral model of every participant in the process was developed. The specifics of presentation of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO in the UNC was established.

This final stage, where **methods of observation and experiment** were used, evidenced determination of an exact place of a text in the EUS (belonging to a certain category of homogeneous texts) that becomes a crucial factor in choosing a translation option. Numerous terminological phrases and expressions in the analyzed official and business documents of UN, NATO, WTO, which act as identifiers in the EUS, were systematized. An increase in the number of unique texts (official and business texts) was noted, which ensures the representativeness of not only the English-Ukrainian sub-corpus of texts of official and business documents, but of the UNC in general. The final stage of the research confirmed the expediency of the division of the whole process of sub-corpus formation into elementary tasks. We generalized the statement that the EUS is a dynamic, context-dependent sub-corpus (only official and business texts of international organizations, such as NATO, UN, WTO) and an optimal algorithm designed to be easily downloaded and used.

Results. Method of observation

Expansion and dominance of empirical research methods evidence the need for relevant statistical information presented in a linguistic framework that not only highlights the latest scientific information but also allows its presentation in an understandable form. Observation as a method of cognition of reality is based on "direct perception of processes, phenomena, and objects" [Vazhynskyi, Shcherbak, 2016, p. 260]. Traditionally, the method of observation is widely used by modern linguists: [Arellano, 2018; Clark, Trousdale, 2013; Crowley, 2007;

Czaykowska-Higgins, 2009; Farr, O'Keeffe, 2019; Kothari, 2004; Lahkar, 2015; McEnery, Xiao, 2010; Pérez-Paredes, 2010; Wilson, 2012]. Scholars regard it as one of the constructs of modern linguistic practices. The solution of each specific problem, combining methods of comparison, measurement, observation and experiment, form a configuration of elements of the empirical research in contrast to ordinary contemplation and have a certain meaning, purpose and tools "helping a knower to move to the subject of research (observed phenomenon) and to the product (result) of the research in a form of a report on the observed phenomena" [Moodle MDU, online].

At this stage, we use the method of observation to outline the content of the proposal of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, WTO and engage relevant methods of analytical tools – diagrams, tables, graphs, charts and figures that specify the meaning of described objects and phenomena given their qualitative affiliation to the UNC.

This choice of the research material, first of all, allows us to take into account modern processing algorithms that have rather complex functionality, and, on the second hand, to offer a model of the EUS, which focuses not on the complication of algorithms, but on their availability. This aim is achieved by involving a large number of people (students, teaching staff and other specialists) in the process of forming a sub-corpus. In addition, the whole process of forming the EUS should be divided into elementary tasks that do not require special knowledge or skills for their solution. For example, students can fill in a sub-corpus (texts) and form dictionaries, while a specialist will perform only the functions of process control. Thus: 1) the whole sub-corpus will be subject to structuring, including the process of sub-corpus creation and filling-in; 2) the data array that can be "processed" by one specialist will greatly increase as compared to existing approaches; 3) in view of the "mass character", the number of unique texts increases, ensuring the representativeness of the UNC; 4) since texts will be processed by people and not by automated computer systems, this will ensure maximum coverage of language phenomena, what is almost impossible to achieve even with the help of complex algorithms; 5) theoretical foundations of the research of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO allow establishing a clear hierarchical structure of the entire UNC.

The place of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO within the UNC structure is shown in Figure 1.

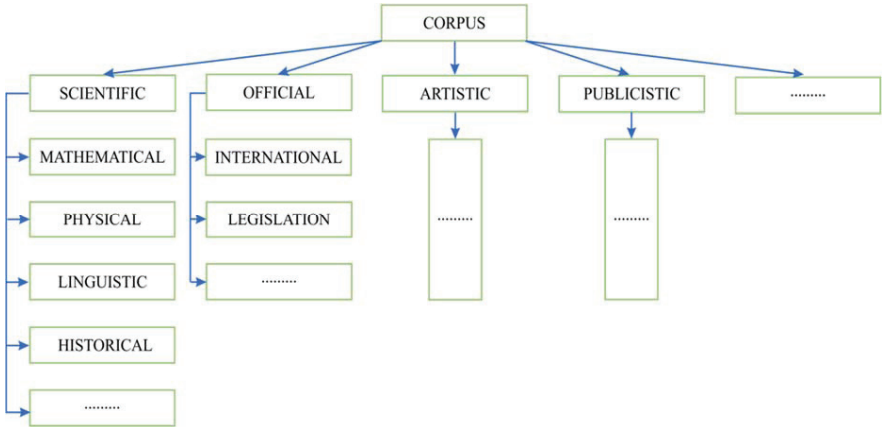

Figure 1. Hierarchical structure of the UNC

Deepening such structuring, each text and dictionary will have its own clearly defined place in the UNC with a "path" to it, what will essentially facilitate the processing of linguistic information by software applications and provide an unambiguous translation with a clearly defined context. The software (mobile application, computer program) will have minimal required functionality and intuitive interface. In this sense, the functionality has a dynamic meaning and carries

information related to text input and translation; editing; formation of dictionaries; collection and processing of statistical data. The software application acquires special importance in case of: 1) automatic determination of the text position in the UNC structure (the same as in the hierarchy proposed above); 2) parameterized request for correct/incorrect definition of this position, what will be discussed further. As a result, a clear hierarchy is established in the process of filling-in and organization of the EUS. For example, a student processes a text, corrects inaccuracies and sends it to a teacher for review; then the teacher "allows" or "does not allow" this text to be included into the EUS; the whole process is considered to be justified because it impacts the formation of the general statistics and understanding of its dynamics. Based on the current situation, a specialist determines the sub-corpus structure, the required direction of moving and current tasks within the EUS. With this approach, a specialist is released from a significant amount of quantitative tasks, which allows one to focus on the quality and improve the process itself. Based on this approach, dictionaries will be compiled "ipso facto" in accordance with the subject of the text and its place in the sub-corpus and in the UNC. In such cases, one can talk about the maximum correlation of dictionaries on the basis of the processed texts, which will inevitably impact the quality of translation as a whole. In fact, by using this approach the software is "self-learning" and is constantly moving towards more accurate translation without the complication of processing algorithms.

The value of the observation method for our research consists in a possibility to trace and select from the textual continuum units with certain features of official and business documents of NATO, UN, and WTO, analysis of which leads to "formation of new knowledge" [Potenko, 2015, c. 116–119], thus evidencing: 1) the importance of a special term as one of the tools for the formation of the EUS and 2) its important role in constructing of a theoretical view on the development and presentation of the terminology of international organizations in one sub-corpus.

In addition, a detailed examination of official and business documents of NATO, UN, and WTO allows to establish a number of regular patterns, namely: 1) the meaning of the term is related to the subject of the document of an organization that it belongs to; 2) the established value significantly impacts the scientific approach and research results.

Thus, we offer to consider the special term (NATO, UN and WTO) as a new object of the corpus analysis, where the key role is played by the hierarchy of construction in the UNC, based on which interpretation of the term is carried out in the EUS in accordance with one of the organizations (NATO, UN, WTO). Expression of such a term under the influence of corpus linguistics is manifested in the context allowing to confirm the meaning of a word or to interpret the analyzed concept. Only an unambiguous term or term combination raises no doubts.

The outline of practical efficiency

The next step of model creation was to demonstrate the practical efficiency of the proposed sub-corpus on the basis of the terminology of official and business documents of NATO, UN, and WTO, including the method of observation. For this purpose, we analyze the ambiguous term "**treatment**" that is also recognizable by fixed expressions: national treatment – national regime (investment) in the WTO document; specialist treatment – special medical treatment in the UN document; degrading treatment – humiliation in the NATO document. The mechanism of use of our model consists in determination of the exact place of the text in the corpus (belonging to a certain category of homogeneous texts), which in practice essentially simplifies the process of automated translation of both a text and special terminology. In other words, we consider it expedient to analyze individual texts, the choice of which maximally fully shows the meaning of special terminology of NATO, UN, and WTO. Use of such analysis in our work is aimed at studying NATO, UN, WTO texts, examination of which allows us to determine the exact version of translation of the term. Moreover, in the construction "national treatment" (national regime (investment regime)) components taken together indicate membership in the WTO, what is confirmed by the following text fragment:

This principle of "national treatment" (giving others the same treatment as one's own nationals) is also found in all the three main WTO agreements (Article 3 of GATT, Article 17 of GATS and Article 3 of TRIPS), although once again the principle is handled slightly differently in each of these [WTO official and business document 1]

Performing the research tasks, we single out the terminological phrase “national treatment” that is used to identify the WTO in the given fragment of the text. The main function of the phrase “national treatment” in the text is an informative one and aims to broaden the understanding of the investment regime for foreigners and country citizens. The corresponding term combination is present in all three main WTO agreements (Article 3 of GATT, Article 17 of GATS and Article 3 of TRIPS) and embodies features of the economic system.

It should be noted here that, we consider those translation options that maximally correspond to the context, and then we form dictionaries (text-oriented approach) already containing features of international organizations, thus narrowing and specifying the meaning of a certain word and terminological phrase.

Having summarized all these options and contexts, we developed a certain structure of the language S (expression, word, etc.) that has several translation options: P1, P2, P3… A graphical model of this process is presented in Figure 2. The sample of its functioning is highlighted with green color.

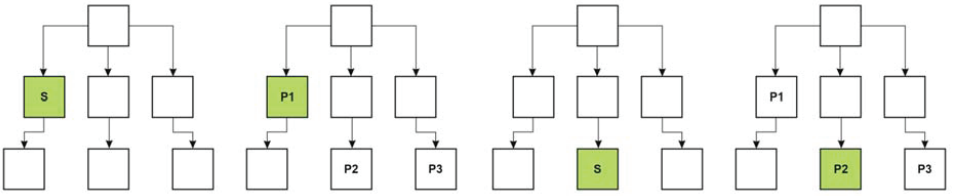

Figure 2. Impact of the place of a text in the corpus on the choice of translation options

Based on the analysis of the illustrative material, our model of the English-Ukrainian sub-corpus of texts of official business and documents of NATO, UN, and WTO is characterized by: 1) exact determination of the place of a text in the corpus; 2) fixed translation version, which is relevant for each separate case. The definition of an exact place of a text in the corpus becomes crucial in the process of choosing a translation option, thus determining its stability and unambiguity within the framework of industry use. According to Figure 2, the translation option P1, and the other two are omitted; respectively, the figure below illustrates option P2, and others are not considered. The specifics of the model are that the expression, preservation and transmission of an unambiguous version of the translation depend on the place of the text in the corpus structure.

At the same time, the received data in computer memory have the structure of a hash table, where the algorithmic complexity of performing operations is almost constant. The necessity to perform some “related” operations points to a complexity that can be described as logarithmic complexity: O (log (x)) and that is at the same time one of the least resource-intensive, what makes it optimal for the multifaceted analysis of large data arrays, as shown in Figure 3.

The obtained values were used for the comparison of the computational complexity of algorithms. The option chosen by us is one of the most optimal (the green line below). Despite the obvious efficiency of the algorithm, exact indicators can be obtained in the process of implementation of this plan in the UNC.

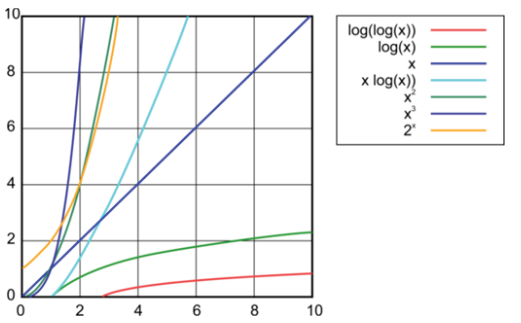

Figure 3. Comparison of algorithms of different complexity

Following that, the division into categories demonstrates: 1) the more iterations of this process, the higher is the efficiency of the translation algorithm; 2) such division must be both extensive and "deep". So, if the structure of the general corpus has the following look as shown in Figure 1 Hierarchical Structure of the UNC, the case examined by us will be placed deeper in the corpus structure. As a result, its place in the corpus can be described with a "path" to it in the following form: <-official and business sub-corpus-> international documents-> documents of international organizations-> NATO documents (UN or WTO as applicable). Identification of a certain document will take place at the algorithmic level by identifying "keywords" that serve as identifiers. Therefore, we propose to examine special terms and phrases of international organizations or title pages of documents to identify an official and business document in the structure of the EUS.

An appropriate presentation of an official and business document (NATO, UN, WTO) is ensured by "keywords", which semantics to some extent explains the meaning of the analyzed document:

1). Resolutions and Decisions adopted by the General Assembly during its seventy-third sessions (United Nations, New York 2019) [UN official and business document 2];

2). Understanding the WTO Agreement on Sanitary and Phytosanitary Measures [WTO official and business document 3];

3). Assistance for the refugee and migrant crisis in the Aegean Sea [NATO official and business document 4].

As illustrated by this example, titles of official and business documents are used in turns to identify the international organization (UN, WTO, or NATO). Apart from that, "keywords" (special terms and expressions) serve as identifiers. They help to make a search and concentrate on the main message. Components of the UN official and business document (Resolutions, General Assembly) perform a function of specification, representing "Resolutions and decisions adopted by the General Assembly at its seventy-third session".

Use of such specialized tokens as refugee and migrant crisis, WTO Agreement to indicate the international organizations of NATO or WTO, evidences the existence of broader opportunities to implement the method of analysis in the EUS, what covers not only the choice of translation options, but also the program control – a result of classification of an official and business document depending on its fixed place in the UNC structure (based on the hierarchy).

The armory of terminological phrases and expressions in the analyzed official and business documents of UN, NATO, and WTO can be divided into the following lexical and semantic groups:

1) protection of human rights and freedoms (human rights, human dignity, fundamental freedom, inter-community trust-building framework, existing legal quotas, Human Rights Monitoring Mission, breakdown of law, human rights challenges, further escalation of tension in the country, humanitarian assistance, foundation of freedom, rule of law, equal rights of men and women, limitation of sovereignty, right to life, liberty and security of person);

2) peacekeeping, humanitarian operations and other operational tasks (global peace, peaceful resolution, preventing conflict, to prevent an unnecessary situation, combat discrimination, national reconciliation, principles of justice and international law, a crippling attack on the kingdom's most important oil installation, threatens to launch a retaliatory campaign, eyewitness reports, indiscriminately targeted schools, largest human exodus, a dire shortage of tents, a multi-ethnic gender-based violence shelter);

3) environmental issues (ecological civilization, relating to the loss of biodiversity, climate change, land degradation, a socio-ecologically centered development pathway, genetic resources, non-communicable diseases, affecting human well-being, low-emission development, forest degradation, to accelerate the transition to a green economy, water-stressed areas, biodiversity loss, healthy and productive ecosystems, private sector business decisions);

4) regulation of international trade (import restrictions, balance-of-payments reasons, production resources, low monetary reserves, a reasonable rate of increase, the programme of economic development, price-based measures, import surcharges, full consultation procedures, incidental protective effects, transparent manner, similar authorization for developing country, inputs needed for production, identical consultation obligations, external trading environment, exchange policies).

Terminological nominations and expressions were obtained from 89 official and business texts of NATO, UN, and WTO and based on the distribution help to search in the UNC and refer to the EUS. Broadening of interpretation of the "keyword" as a defining construct evidences the conceptuality of this document and determines the choice of the necessary context by an average user leading to formation of the vocabulary in the EUS.

The above suggests that the algorithm at first looks for "keywords" and expressions that occur in the text and then determines the place of the text in the corpus structure. Distinguishing the terminology of international organizations, the algorithm chooses the necessary translation version based on the specific place of the text in the corpus. The introduction of the EUS into the UNC structure will provide an opportunity to implement the type of official and business document that is the result of the search.

Experiment

The symbolic nature of the experiment as a scientific method is that it is necessary for a special study of real phenomena with the participation of those subjected to its direct influence. In our experimental research, namely, students are the key link guaranteeing the successful completion of every stage of the formation of the EUS. As a rule, the method of the experiment requires special conditions of implementation provided that a distinction is made between sufficient and insufficient data for confirmation of the hypothesis, allowing to itemize or interpret them. Taking into account that the experiment should be carried out within a short time, we consider it appropriate to analyze individual components as parts of this real-time analysis with minimal cost, in which case the obtained results reflect their quality and efficiency to the fullest extent and confirm the initial hypothesis.

The filling-in of the English-Ukrainian sub-corpus is a creative part of the research. Therefore, it is quite logical to use the method of the experiment at the final stage of our research in order to verify the accuracy of theoretical calculations.

The practice of data selection based on the material of Yu.I. Demyanchuk [Demyanchuk, 2015, c. 132] provides for consideration of the past experience when students were already involved in such experimental processes. The successful experience of the researcher in the field of linguistic experimental research is evidenced by the formation of the Ukrainian-German fire safety dictionary when 24 students were in the focus of the academic research for about 5 months. The research value of the dictionary consists of its effective implementation in the teaching practice and the constructive role of students in its development. We have also derived the formula (c) indicating the number of words processed by one student and formula (T) indicating the time spent. Inserting propositional variables into the formula (c), we determine the meaning of words and express them as a logical structure.

Therefore:

P=3700 words; n=24 students; t=5 months.

The result is as follows:

c=(P/n)/t=(3700/24)/5=30.8≈31 words for one student for one month.

We should note that the number of words depends on the time spent and equals 31 words per month. The formulated value (c = 31) is implicated into the formula (T) and the spent time is calculated.

The general formula for this process will look as follows:

$$T=N/(c\times n),$$

where: T – the time spent; N – the general number of words. Formulas c and T are equivalent, each of them is a logical sequence of the other, i.e. c=T, however only when c→T and T→c.

The mechanism for studying the status of the development contained the results of: 1) processing of measurements; 2) observation of the process of filling-in of the dictionary; 3) completing of the vocabulary sample included in the structure of the method as a part of the experiment.

Namely, the formation of the Ukrainian-German fire safety dictionary and the developed mathematical formulas for its calculation impacted the further development of the EUS. In our

scientific work, we will use the developed formulas, as well as the method of the experiment to confirm our hypothesis (quick and efficient filling-in of the EUS).

The following calculations will show the approximate progress of filling-in of the sub-corpus based on the proposed approach. In the experimental research, we will use graphical methods as technical tools for calculation of the graphical representativeness of data in order to summarize the results of the empirical material and formulate conclusions. Within the framework of the formula shown above, we determine the hyperbolic dependence between the time required to fill in the sub-corpus with the number of words N and the number of students involved in the process, as shown in Figure 4 and Figure 5.

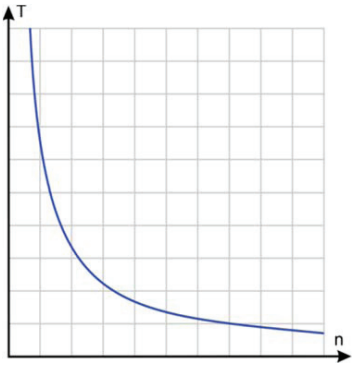

Figure 4. Dependence of the time spent on the number of involved persons

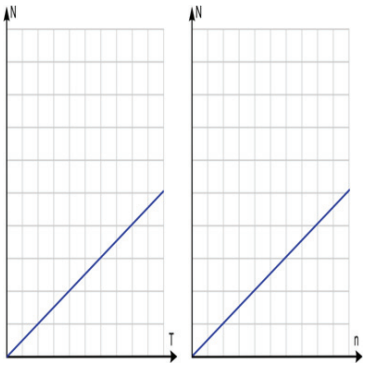

Figure 5. Dependence of the number of words on the time spent

Also we cannot ignore the direct dependence of the number of words in the sub-corpus (N) on the time spent (T) and the number of students involved (n):

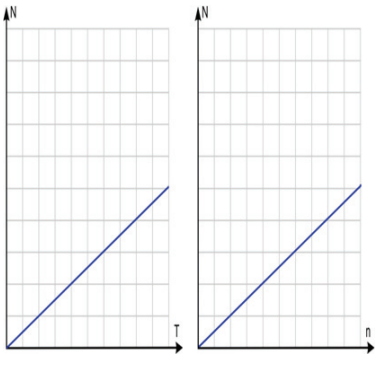

Figure 6. Dependence of the number of words on the number of involved persons

Based on the results of the analysis, involving of 1,000 students in the process of constructing of the sub-corpus results in its filling-in with about 1 million of words/phrases with accurate translation and actualizes the potential of the text dynamics with a view to distinguish important official and business texts of NATO, UN, and WTO. To verify the scientific information, we present the general case of the research results, where the set of values (filling-in of the sub-corpus and the number of spent resources) remains constant, which is reflected by the distribution plane in Figure 7.

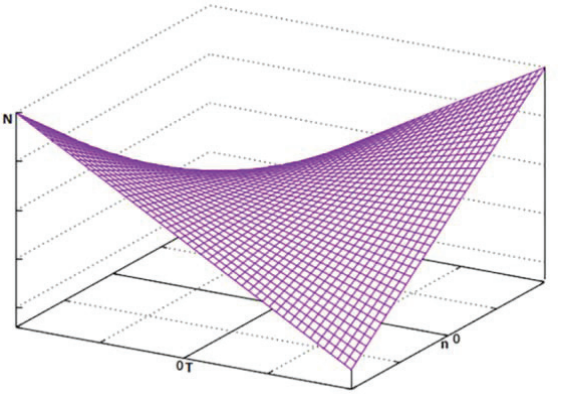

Figure 7. Dependence of the filling-in process on the spent resources (time and persons)

In particular, to improve the quality of translation and ensure the effective filling-in of the EUS, we propose to apply a comprehensive approach, when the general problem is divided into elementary components. Namely in this way the task is represented in the sub-corpus of texts, is objectified by the persons responsible for it, and is reduced to the division of a problem into constituent parts. This division of components provides the transfer from the general structuring of the EUS to the ability to interpret and solve one of the problems that are considered separately (by narrowly focused specialists), where each group of specialists is responsible for its own task and does not interfere into the overall structure of work.

Narrowing the scope of specialist work from the general plane to a narrow specific area will impact the quality of the sub-corpus and its content without complex algorithms further complicating for automated text processing.

In other words, the essence of the proposal is as follows: 1) filling-in and outlining the corpus of texts, forming a sample of official and business text fragments and identification of lexical items, the semantics of which reveal the content of international organizations (NATO, UN, WTO) were carried out by different groups of specialists and each of them was responsible for the set task, and 2) at each stage of formation of the sub-corpus, a focus was set on the improvement of the final result, not just on successful completion of every separate process.

We postulate the division of the whole process of sub-corpus formation into elementary tasks and increase their efficiency as an optimal solution for solving such problems, which is confirmed by mathematical calculations. In our research, the experimental method is also focused on the improvement of a certain process on the one hand, and on the preservation of the integrity of all involved elements on the other hand, which is the most fully embodied in the EUS model. The main function of our approach is a significant increase by 100 per cent with a division into 10 parts by 10 per cent. In such a way, we will increase the efficiency twice (by 100 per cent). The obtained result is checked by using the compound interest formula as shown below:

$1.1^{10}=2.59$ – which thus confirms the efficiency of our approach.

The findings imply that the focus on the graphic illustration and adequate interpretation of the presented information envisage the division of the entire process of the EUS formation into elementary tasks, and the use of the compound interest formula confirms this efficiency. The usual (direct) approach to the formation of the sub-corpus is represented by a red line, and our approach (method of division) is represented by a blue line as systematized in Figure 8.

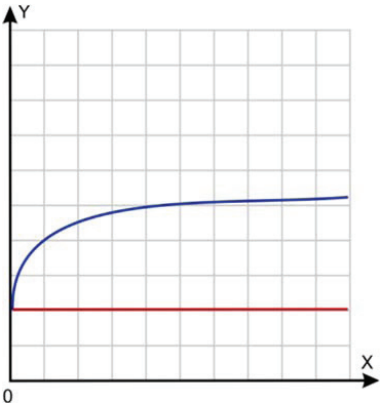

Figure 8. Comparison of the efficiency of two methods

Successful completion of such comparison, i.e. finding the optimal solution for further development and modification of a sub-corpus of texts, is conditioned by the combination of different methods and techniques in one research and presents a holistic addition to the UNC. Our scientific research identified that the approach envisaging division of the whole process of formation of the EUS into elementary tasks is more effective, and therefore it confirms the initial hypothesis and presents an effective research tool. However, this efficiency increase is restricted by a certain upper limit. Therefore, the mathematical pattern is relevant only for minor changes. But it is sufficient in our case.

Theoretical comprehension and empirical calculations require practical application in the UNC, where actualization of our proposal is potentially possible and relevant. The EUS is aimed at marking, recognizing and parallel translation of official and business texts of NATO, UN, and WTO, while there is no such sub-corpus in the UNC.

Based on our findings, comprehensive and systematic research envisages improvement of the final result and is a logical conclusion of our experiment.

Discussions

The study sought to create a model of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO as a supplement to the Ukrainian National Corpus, where texts are presented as a hierarchical structure. According to the author's knowledge, it is impossible to establish which of the official and business texts in the UNC has characteristics of an international organization (NATO, UN or WTO) and the corpus does not have any exact division into categories of homogeneous texts that significantly simplify the search process for both an average user and a linguist. Although the reviewed works reveal the most important tasks solved by the UNC that include the introduction of corpus-based methods of research of the Ukrainian language, development of corpus Ukrainian linguistic studies as a separate direction, as well as ensuring the technological status of the Ukrainian language in conditions of the information society.

It was found that the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, WTO within the UNC is retransmitted through the development of a machine language model, as well as the development of an application for automatic recognition and synthesis of English-Ukrainian speech. Selection of the factual material for the research envisaged the procedural component of the methodology of corpus analysis and found its place in the sub-corpus structure. Despite the complexity of analysis of official and business style (official and business texts, terms, phrases) used in the process of communication in the political and economic life, legislation, administrative and economic activities and constantly changing and evolving in the real time, we consider it expedient to analyze individual texts (first of all NATO, then WTO and UN) as components of this process. The findings imply that the method of observation in our research served as an impetus for further development and the "accelerated" evolution of the UNC. The proposed model of the EUS shows that the sub-corpus itself and the process of its creation and filling-in (not texts, words or phrases) will be subject to structuring.

The possibility of modification of the sub-corpus is confirmed by empirical data, in particular with respect to its filling-in. In this case, the data array that can be "processed" by one specialist will greatly increase as compared to existing approaches. At the same time, the number of unique texts will increase ensuring the representativeness of the sub-corpus, and, in view of the maximum coverage of linguistic phenomena, a clear hierarchical structure of the UNC will be established.

The study agrees with the previous research, particularly with Shvedova [Shvedova, 2017, c. 33–38] who suggests that parameterization of subject areas of the UNC is based on the source base of the Ukrainian lexicography, which genre features are conditioned by stratification of the selected texts. Additionally, scholars Darchuk [Darchuk, 2010, c. 11–19] and Demska-Kulchytska [Demska-Kulchytska, 2011, c. 219] lay the foundation for a theory of statistical analysis of lexical tokens within and outside the corpus of texts. Based on our findings, we also totally agree with Bobkova [Bobkova, 2014] who argued that validation of basic requirements to text sources should consist of: a) adequate representation of the modern Ukrainian language in all major areas; b) coverage of all genres where functional language styles are implemented; c) focusing on the general reader; d) sufficiency as regards quantitative parameters.

Conclusion

According to the results of our study, the model of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO is the embodiment of a parallel corpus of texts in the UNC. The presented model provides for a clear division into categories of homogeneous texts and a quick search for special terminology of international organizations, such as NATO, UN, and WTO, in the EUS. The analyzed English official and business texts allowed to separate numerous words and expressions of identifiers that help to carry out search and refer to a relevant text in the sub-corpus. Compilation of a glossary of terminological nominations, their functional and stylistic labeling and assimilation into the commonly used vocabulary of the EUS and later on in the UNC are currently at the stage of formation and approval. The current research manifests the concept of the EUS as a sub-corpus of the UNC that allows the prompt presentation of relevant texts of the official business style and coverage of the broadest range of interested readers. Further investigation should concern the practical prospect of our model.

Sources

WTO official and business document 1 – Principles of the trading system. Retrieved from https://www.wto.org/english/thewto_e/whatis_e/tif_e/fact2_e.htm

UN official and business document 2 – Resolutions and Decisions adopted by the General Assembly during its seventy-third sessions. Vol. I: Resolutions, 18 September–22 December 2018. New York.

WTO official and business document 3 – Understanding the WTO Agreement on Sanitary and Phytosanitary Measures. Retrieved from https://www.wto.org/english/tratop_e/sps_e/spsund_e.htm

NATO official and business document 4 – Assistance for the refugee and migrant crisis in the Aegean Sea. Retrieved from https://www.nato.int/cps/en/natohq/topics_128746.htm

Bibliography

Балог, В. (2008). Принцип кодування абревіатур у Національному корпусі української мови. *Лексикографічний бюлетень*, 17, 90-93.

Бобкова, Т.В. (2014). Концепція колокації: корпусний підхід. *Науковий вісник Міжнародного гуманітарного університету*. *Серія «Філологія»*, 10, 2, 42-45.

Важинський, С.Е., Щербак, Т.І. (2016). *Методика та організація наукових досліджень*. Суми: СумДПУ імені А.С. Макаренка.

Глущенко, В.А. (2010). Лінгвістичний метод і його структура. *Мовознавство*, 6, 32-44.

Дарчук, Н.П. (2013). Автоматичний синтаксичний аналіз текстів корпусу української мови. *Українське мовознавство*, 43, 11-19.

Демська-Кульчицька, О. (2005). *Основи Національного корпусу української мови*. Київ: Інститут української мови НАНУ.

Дем'янчук, Ю.І. (2015). *Українсько-німецький пожежно-технічний словник.* Львів: ЛДУ БЖД.

Жуковська, В. (2013). *Вступ до корпусної лінгвістики*. Житомир: Видавництво ЖДУ ім. І. Франка.

Касьяненко, М.К., Лебедєв, К.М., Петренко, П.М. (2009). Принципи побудови корпусу текстів різних функціональних стилів. *Науковий вісник Волинського національного університету імені Лесі Українки*, 6, 25-28.

Потенко, Л.О. (2015). Методологія дослідження емотивності німецьких фразеологічних дериватів. *Одеський лінгвістичний вісник*, 5, 2, 116-119.

Шведова, М.О. (2017). Корпусні методи дослідження регіональних відмінностей у межах однієї мови (на матеріалі регіональних корпусів української та російської мов). *Вісник Харківського національного університету імені В.Н. Каразіна. Серія «Філологія»*, 77, 33-38.

Arellano, R. (2018). A corpus linguistics application in the analysis of textbooks as national teaching instruments of English as a second language in Chile. *Actualidades Investigativas en Educación*, 18, 1, 1-19. DOI: 10.15517/aie.v18i1.31807.

Clark, L., Trousdale, G. (2013). Using participant observation and social network analysis. M. Krug, J. Schlüter (Eds.), *Research Methods in Language Variation and Change* (pp. 36-52). Cambridge: Cambridge University Press.

Crowley, T. (2007). *Field linguistics: a beginner's guide*. Oxford: Oxford University Press.

Czaykowska-Higgins, E. (2009) Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous com-munities. *Language documentation & conservation*, 3, 15-50.

Farr, F., O'Keeffe, A. (2019). Using corpora to analyse language. S. Walsh, S. Mann (Eds.), *Routledge handbook of English language teacher education* (pp. 268-282). London: Routledge.

Kothari, C.R. (2004). *Research methodology: methods and techniques*. New Delhi: New Age International.

Lahkar, N. (2015). Linguistic data collection: a field observation. *Language in India*, 15, 10, 216-223.

McEnery, T., Xiao, R., Tono, Y. (2006). Corpus-based language studies: an advanced resource book. London: Published Routledge.

Moodle MDU. Retrieved from http://moodle.mdu.in.ua

Pérez-Paredes, P. (2010). Corpus linguistics and language education in perspective: appropriation and the possibilities scenario. T. Harris, J.M. Moreno (Eds.), *Corpus Linguistics in Language Teaching* (pp. 53-73). Wien: Peter Lang.

Wilson, A. (2012). *Participant observation: a LIP discussion*. Retrieved from http://www.paradisec.org.au/blog/2012/06/participant-observation-a-lip-discussion

A MODEL OF THE ENGLISH-UKRAINIAN SUB-CORPUS OF TEXTS OF NATO, UN AND WTO OFFICIAL AND BUSINESS DOCUMENTS

Yuliya I. Demyanchuk. Lviv State University of Life Safety (Ukraine)
e-mail: y.demianchuk@gmail.com
DOI: 10.32342/2523-4463-2022-2-24-14

Key words: *Ukrainian National Corpus, English-Ukrainian sub-corpus, official and business documents, terminology of NATO, UN, WTO, method of observation, linguistic experimental research.*

This academic article offers a model of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO (hereinafter referred to as the EUS) as a supplement to the Ukrainian National Corpus (hereinafter referred to as the UNC), where texts are presented as a hierarchical structure. In an effort to provide a more comprehensive view of NATO, UN, and WTO official and business documents, a mechanism of use of our model envisages determination of the exact position of a text in all UNC corporas, what essentially simplifies the practical process of search and automated translation of texts and special terminology in particular. In consequence, the author offers to increase the number of unique texts that will ensure the representativeness of the EUS in the Ukrainian National Corpus, and to divide the whole process of the EUS formation into elementary tasks that will increase the UNC efficiency.

The practical value of the research lies in a possibility to use the EUS for texts parallel translation and terminological phrases separation on the basis of such texts.

The methodological potential of our research is based on the works of well-known Ukrainian and foreign scholars (V. Bohorodytskyi, I. Baudouin de Courtenay, K. Brugman, W. Wundt, B. Delbruck, G. Paul, O. Potebnia, P. Fortunatov, W. Scherer, H. Steinthal, A. Schleicher et al.), who strengthened the scientific interest to the problem of translation and corpus of texts in particular.

The object of the research is a model of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO.

The purpose of the research is to prove that the model of the English-Ukrainian sub-corpus of official and business documents texts can become a full-fledged informational foundation for the formation of texts parallel corpus within the Ukrainian National Corpus.

The methods of observation and experiment were applied to analyze empirical, review, conceptual analytical tools that specify the meaning of described objects and phenomena given their qualitative affiliation to the UNC.

The study focuses on a number of *tasks*, in particular:

- to determine a theoretical and linguistic basis of the concept "English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO";
- to determine the mechanism of updating of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, WTO within the UNC;
- to find out the specifics of formation and filling-in of the EUS;
- to substantiate the relevant methodology of examination of official and business texts of NATO, UN, and WTO in the English-Ukrainian sub-corpus of texts of official and business documents;
- to single out and itemize the advantages of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO within the UNC.

Analysis of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO as a construct of the parallel corpus of texts of the UNC was carried out in several stages.

The first stage envisaged description and inventory. Theoretical works on the history of the Ukrainian National Corpus were reviewed and generalized, and the concept of "English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO" was determined. It was found out that the UNC is the object of research in such areas of linguistic studies as the applied and corpus linguistics that construe the corpus of texts as a translator's tool. This encourages us to highlight the advantages and disadvantages of the UNC and justify the expediency of our proposal to develop the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO. Definition of the UNC as a national corpus of texts that contains only Ukrainian texts and does not envisage a parallel translation evidenced the need to create a proposal of a parallel corpus of texts with an updated English-language database.

The selection of lexical items with the official and business semantics is based on documents of international organizations, such as NATO, UN, and WTO. At the same stage, we determine the purpose of the research, in particular, the formation of the English-Ukrainian sub-corpus of texts of official and business documents, which envisages a parallel translation of NATO, UN, WTO texts.

This sequence of procedures allowed us to carry out consideration and interpretation of existing theoretical provisions and inventory of factual research materials.

Theoretical and corpus analysis methods were used in the research.

The second stage envisaged analysis and differentiation. Special aspects of filling-up of the EUS with official and business texts of NATO, UN, and WTO were characterized. Theoretical analysis of the EUS was carried out in order to establish a clear hierarchy of the UNC structure. The main types of participants of the EUS formation (students, teaching staff) were differentiated and the behavioral model of every participant in the process was developed. The specifics of presentation of the English-Ukrainian sub-corpus of texts of official and business documents of NATO, UN, and WTO in the UNC was established.

This final stage, where methods of observation and experiment were used, evidenced determination of an exact place of a text in the EUS (belonging to a certain category of homogeneous texts) that becomes a crucial factor in choosing a translation option. Numerous terminological phrases and expressions in the analyzed official and business documents of UN, NATO, WTO, which act as identifiers in the EUS, were systematized. An increase in the number of unique texts (official and business texts) was noted, which ensures the representativeness of not only the English-Ukrainian sub-corpus of official and business documents texts, but of the UNC in general. The final stage of the research confirmed the expediency of the division of the whole process of sub-corpus formation into elementary tasks. We generalized the statement that the EUS is a dynamic, context-dependent sub-corpus (only official and business texts of international organizations, such as NATO, UN, WTO) and an optimal algorithm designed to be easily downloaded and used.

Conclusion. The analysis conducted in the present paper is meant to illustrate a new approach to forming a modern model of the sub-corpus of texts of official and business documents. The model of the English-Ukrainian sub-corpus of official and business documents texts of NATO, UN, and WTO provides for a clear division into categories of homogeneous texts and a quick search for special terminology of international organizations, such as NATO, UN, and WTO, in the EUS. In the context of the texts analyzed

in the current paper numerous separated words and expressions-identifiers can help to carry out search and refer to a relevant text in the sub-corpus. Compilation of terminological nominations glossary, their functional and stylistic labeling and assimilation into the commonly used vocabulary of the EUS and later on in the UNC are currently at the stage of formation and approval. To this end, the current research manifests the concept of the EUS as a sub-corpus of the UNC that allows the prompt presentation of relevant texts of the official business style and coverage of the broadest range of interested readers.

References

Arellano, R. (2018). A corpus linguistics application in the analysis of textbooks as national teaching instruments of English as a second language in Chile. *Actualidades Investigativas en Educación*, vol.18, issue 1, pp. 1-19. DOI: 10.15517/aie.v18i1.31807.

Baloh, V. (2008). *Pryntsyp koduvannia abreviatur u Natsionalnomu korpusi ukrainskoi movy* [The principle of encoding abbreviations in the national corpus of the Ukrainian language]. *Lexicographic Bulletin: Collection of Scientific Works,* vol. 17, pp. 90-93.

Bobkova, T.V. (2014). *Kontseptsiia kolokatsii: korpusnyi pidkhid*. [Conception of collocation: corpus-based approach]. *Scientific Bulletin of the International Humanities University. Series: Philology*, vol. 10, issue 2, pp. 42-45.

Clark, L., Trousdale, G. (2013). Using participant observation and social network analysis. In M. Krug, J. Schlüter (eds.). Research Methods in Language Variation and Change. Cambridge, Cambridge University Press, pp. 36-52

Crowley, T. (2007). Field linguistics: a beginner's guide. Oxford, Oxford University Press, 216 p.

Czaykowska-Higgins, E. (2009) Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous com-munities. *Language documentation & conservation*, vol. 3, pp. 15-50.

Darchuk, N.P. (2013). *Avtomatychnyi syntaksychnyi analiz tekstiv korpusu ukrainskoi movy* [Automatic syntax analysis of texts of Ukrainian language corpus]. *Ukrainian Linguistics,* vol. 43, pp. 11-19.

Demska-Kulchytska, O. (2005). *Osnovy natsionalnoho korpusu ukrainskoi movy* [Fundamentals of the national corpus of the Ukrainian language]. Kyiv, Institute of the Ukrainian Language of the NAS of Ukraine Publ., 219 p.

Demyanchuk, Y.I. (2015). *Ukrainsko-nimetskyi pozhezhno-tekhnichnyi slovnyk* [Ukrainian-German fire and technical dictionary]. Lviv, Lviv State University of Life Safety Publ., 132 p.

Farr, F., O'Keeffe, A. (2019). Using corpora to analyse language. In S. Walsh, S. Mann (eds.). Routledge handbook of English language teacher education. London, Routledge, pp. 268-282.

Hlushchenko, V.A. *Lingvistychny metod i iogo struktura* [Linguistic method and its structure]. Linguistics, vol. 6, pp. 32-44.

Kasianenko, M.K., Lebediev, K.M., Petrenko, P.M. (2009). *Pryntsypy pobudovy korpusu tekstiv riznykh funktsionalnykh styliv* [Principles of building a corpus of texts of different functional styles]. *Scientific Bulletin of Volyn National University named after Lesya Ukrainka,* vol. 12, pp. 25-28.

Kothari, C.R. (2004). Research methodology: methods and techniques. New Delhi, New Age International Publ., 418 p.

Lahkar, N. (2015). Linguistic data collection: a field observation. *Language in India, vol.* 15, issue 10, pp. 216-223.

Moodle MDU. Available at: http://moodle.mdu.in.ua (Accessed 27 September 2022).

Pérez-Paredes, P. (2010). Corpus linguistics and language education in perspective: appropriation and the possibilities scenario. In T. Harris, J.M. Moreno (eds.). Corpus Linguistics in Language Teaching. Wien, Peter Lang Publ., pp. 53-73.

Potenko, L.O. (2015). *Metodolohiia doslidzhennia emotyvnosti nimetskykh frazeolohichnykh deryvativ* [Methodology for research on the emotionality of German phraseological derivatives]. *Odessa Linguistic Bulletin,* vol. 5, issue 2, pp. 116-119.

Shvedova, M.O. (2017). *Korpusni metody doslidzhennia rehionalnykh vidminnostei u mezhakh odniiei movy (na materiali rehionalnykh korpusiv ukrainskoi ta rosiiskoi mov)* [Corpora methods in studying regional language varieties within one language (exemplified by geographically annotated Ukrainian and Russian corpora)]. *The Journal of V.N. Karazin Kharkiv National University. Series: Philology,* vol. 77, pp. 33-38.

Shypnivska, O. 2005. *Funktsionuvannia mizhmovnykh morfolohichnykh omonimiv v ukrainskykh tekstakh* [Functioning of inter-language morphological homonyms in Ukrainian texts]. *Linguistics,* vol. 6, pp. 70-79.

Vazhynskyi, S.E., Shcherbak, T.I. (2016). *Metodyka ta orhanizatsiia naukovykh doslidzhen* [Methods and organization of scientific research]. Sumy, Sumy State Pedagogical University Publ., 260 p.

Wilson, A. (2012). Participant observation: a LIP discussion. Available at: http://www.paradisec.org.au/blog/2012/06/participant-observation-a-lip-discussion (Accessed 27 September 2022).

Zhukovska, V. (2013). *Vstup do korpusnoi linhvistyky* [Introduction to Corpus Linguistics]. Zhytomyr, ZhSU Publishing House, 187 p.

Одержано 25.10.2022.