Frequency Dictionaries to the Instructions to Medical Products

Roksolana-Yustyna Perkhach[0000-0002-6466-6122] and Yuliia Shyika[0000-0003-2474-0479]

Lviv Polytechnic University, 79005 Lviv , Ukraine
roksoliana-iustyna.t.perkhach@lpnu.ua,
yuliia.i.shyjka@lpnu.ua

Abstract. The actuality of the research stems from the necessity to investigate the linguistic cognitive and linguistic cultural characteristics of the terms in the instructions to medical products. Since the number of imported drugs has grown rapidly in recent years and the need for the correct translation of the instructions has also increased. The methodological basis for the study has been formed from researches of contemporary scientists. Thus, the compilation of frequency dictionaries of modern medical terminology will facilitate the study of the terms used in the instructions to medical products. The tokenization of the instructions texts has been made using the KWIC Concordance corpus manager. 195 instructions to medical products have been analyzed and sorted according to Anatomical Therapeutic Chemical Classification System. As a result, the frequency dictionaries have been created: Ukrainian language dictionary includes 1000 words, Polish language dictionary includes 1000 words and the frequency dictionary of German language includes 1000 words.

Keywords: research corpus, frequency dictionary, medical terminology, KWIC Concordance, text corpora.

1 Introduction

Rapid development of medicine has resulted in the growth of the number of medical terms which are difficult to process without appropriate technical means. One of the most common forms of linguistic information systematization is the linguistic corpora, which helps to create a text resource. Afterwards such data can be used in various scientific researches.

Nowadays there are numerous medical dictionaries of high quality [1, 2, 3], however there is no frequency dictionary for medical terminology.

The purpose of the study is to create frequency dictionaries using corpus-based technologies as well as to introduce corpus technologies of representation of various texts of contemporary Ukrainian, Polish and German languages.

The set goal requires the solution of the following tasks:

1. to describe and systematize existing scientific researches of Ukrainian and foreign scholars on the texts corpus and the role of corpus based researches in modern science;
2. to study classifications of linguistic text corpora most popular among scholars;

3. to create and describe the factual bases of the research (for Ukrainian, Polish and German instructions to medical products);
4. to describe the method of frequency dictionary compilation on the basis of the text corpus of the instructions to medical products.
5. to identify and describe the benefits of frequency dictionaries regarding the specific features of modern medical terminology.

From a cognitive point of view, such a study is of high importance because it will enable us to identify the peculiarities of Ukrainian, Polish and German terminology, and to standardize Ukrainian terminology.

The terminological and methodological basis for the study has been formed from researches of O. Baranov [4], M. Baker [5], A. Hardie, P. Baker, T. McEnery [6], S. Buk., Y. Levus, Y. Yavorskyi, [7], Y. Karpilovska [8] O. Maksymiv [9] et al.

2 Problem and proposed method

2.1 Related works and researches

To compile a frequency dictionary for medical terminology, it is necessary to create a text resource of the linguistic information storage – a linguistic corpus which consists of three subcorpora: Ukrainian, Polish, and German instructions to medical products.

The first attempts to create an electronic text resource for conducting linguistic research were made by «Centro Automazione Analisi Linguistica» (CAAL) (Italy, 1956) [10, p. 11].

Later many studies in the field of automated data processing have been conducted, which resulted in the formation of several meanings of the notion «linguistic corpus» in linguistics. These meanings are the following: electronic archive, electronic libraries, collection of texts, array of texts, full text database, etc.

O. Demska offers the following definition of the corpus: «Organized electronic collection of written and oral texts of any natural language, which has obligatory characteristics and which is assigned for scientific study of a language» [11, p. 89].

However, M. Svidzinskyi believes that a corpus is structured and philologically competent linguistic data which are represented in an automated form, intended to solve specific linguistic problems [12, p. 27].

A. Baranov, P. Baker, Ye. Karpilovska, G. Sinclair and others suggest classifying linguistic corpora according to:

- type of linguistic data (written, oral, mixed);
- «parallelism» (monolingual, bilingual, multilingual);
- «literacy» (literary, dialect, colloquial, terminological, mixed);
- the purpose of creation (multi-purpose and specialized);
- genre (literary, folk, drama, journalistic);
- accessibility (available for free, commercial and private);
- purpose (research and illustrative);

- dynamism (dynamic and static);
- markup (marked and not marked);
- markup type (morphological, semantic, syntactic, prosodic, and others);
- text volume (full text and "fragmented") [4; 8; 13].

Some researchers assume that it is necessary to reduce the number of types of linguistic corpora and to distinguish the following ones: specialized, reference, multilingual, parallel, educational, diachronic, and instructive. O. Demska-Kulchytska claims that there are the following corpora types:

- full text (texts in the corpus are full) and fragmentary (there are fragments of texts in corpus);
- research (are used in linguistic studies to formulate new theories and concepts);
- illustrative (are used to prove theorems or hypotheses about the language);
- monitoring / dynamic (provide the possibility to monitor changes in language, taking into account the aspect of diachrony; statistic corpora verify the state of language in a certain synchronous period);
- diachronic (represent the language over time period, and the synchronous corpora represent language or type of the text in a fixed time period);
- general language (represent the general, national language, and specialized corpora, are aimed to solve specific and research tasks) [11, p. 156-157].

Multilingual and parallel corpora as well as comparative corpora form a separate field of linguistics, which is highly important for the theory and practice of translation [5, p. 234].

The main features of the corpus are:

- representativeness: the ability of the corpus to reflect all the properties of the subject area, that is the level of implementation of the language system, which contains linguistic descriptive phenomena;
- authenticity: involves the selection of the written or oral text(s) or passage(s) of the text(s) created by the native speaker(s) in the process of real communication. Meeting the requirement of authenticity is one of the components of empirical research of the actual corpus material;
- selection demands the limitation of the actual material by selecting certain fragments of the language;
- balance: the proportional amount of text resources in the corpus [11, p. 102].

Scientists use four basic parameters to characterize linguistic corpora: firstly, it should be sufficiently large; secondly, the corpus should be structured or tagged; thirdly, the texts, components of a particular corpus, should be in e-form; fourthly, «electronic corpus» should include special software for further corpus processing [4; 5; 6].

V. Shyrokov notes that during the creation of the Ukrainian National Corpus the following criteria for the texts selection have been used:

- diachronic aspect (which texts and of which time period should be selected);

- stylistic (texts should also represent substyles of the national language);
- territorial (texts should represent the specificity of the literary language depending on the region of Ukraine, as well as represent examples of oral or written texts created abroad);
- quantitative (clearly specifies the number of words in the text or passage, which are included in the corpus, the number of texts and / or passages). Generally, in corpora studies the criteria for selecting text fragments are separate problem, within which linguistic and technical criteria are developed, and in the scientific literature on this problem they are generalized as the theory of criteria for text material selection for corpora of different types [10, p. 11].

2.2 Experimental results

One of the tasks of our research is to create a research corpus of instructions for medical products. We have considered the linguistic corpus as a text collection stored in electronic databases, which consists of three subcorpora: Ukrainian, Polish, and German [14, 15, 16].

There are a lot of programs for text processing:

1. programs of analysis and linguistic processing of texts;
2. text conversion programs;
3. psycholinguistic programs;
4. texts generators and "talking programs";
5. systems of natural language processing and others [13, 18].

The program for analysis and linguistic processing of texts KWIC Concordance for Windows [17] was used in our research.

The program possesses the following features:

- building up the concordance;
- searching both separate words and combinations;
- sorting the list of words according to several criteria set by a user;
- the ability to display the found word forms in the context;
- saving the results;
- quick processing of the request;
- supporting text file format (txt).

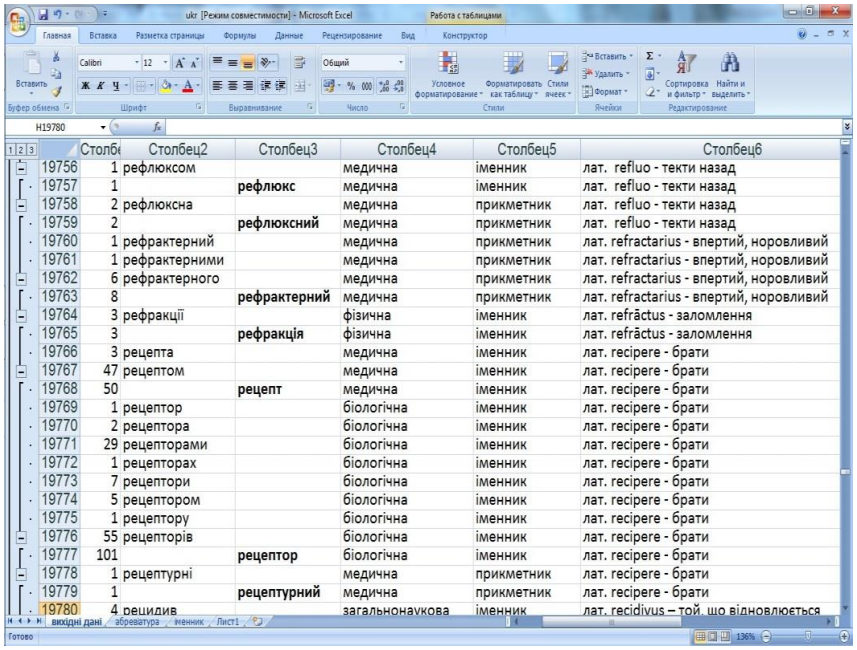
Computer program for linguistic data processing KWIC Concordance for Windows is attached to the Windows operating system, has lists of combinations to the reference element specified by the token set (Collocation tool).

Fragments of factorial databases of research corpora on the material of the languages under research are given in Fig. 1, 2, 3. [21, p. 27].

Frequency dictionary is «a type of a dictionary, which shows the number of uses, that is, the frequency of a certain part of speech in the texts under survey. As a rule, it consists of several lists: lists of words, usually in a root form, listed according to the

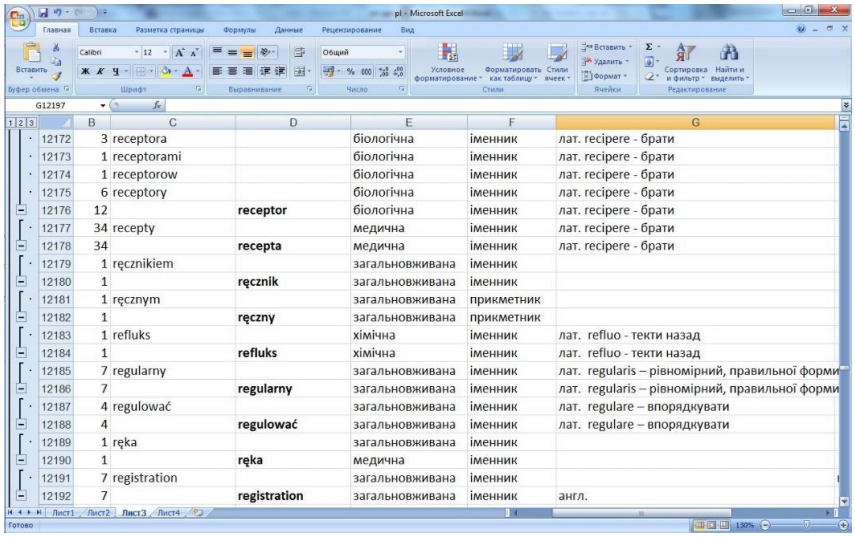
frequencies reduction and in alphabetical order and the similar lists for word forms» [19, p. 292].

The importance of frequency dictionaries data has been repeatedly emphasized by modern scholars: «Statistics also indicates that the first 100 of the most frequent words covers about half the average text. However, the frequency dictionary data objectively reflect vocabulary of the language only to the extent of the objective selection of sources for this vocabulary. Their reliability depends on the quantitative and qualitative characteristics of the lexical array for counting» [9, p. 109].



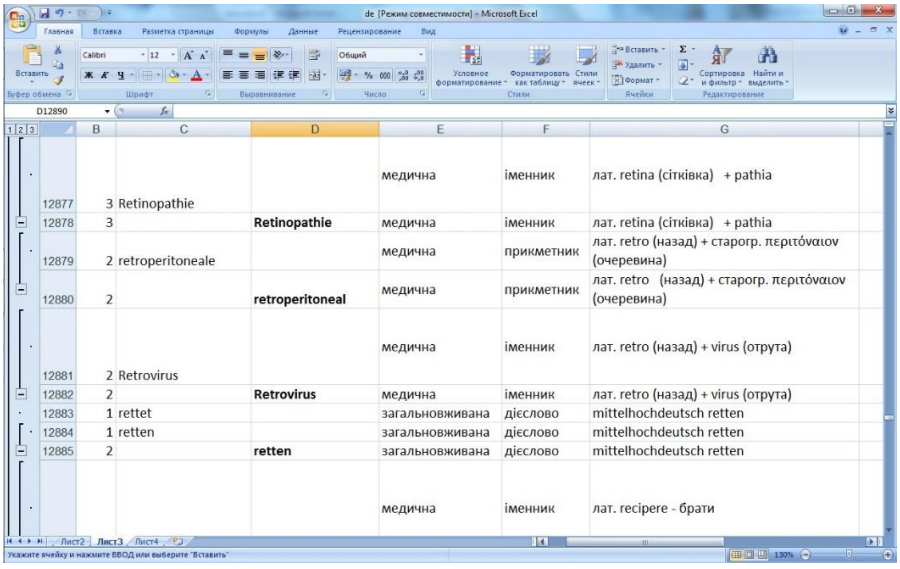
	Столбец	Столбец2	Столбец3	Столбец4	Столбец5	Столбец6
19756	1	рефлюксом		медична	іменник	лат. refluo - текти назад
19757	1		**рефлюкс**	медична	іменник	лат. refluo - текти назад
19758	2	рефлюксна		медична	прикметник	лат. refluo - текти назад
19759	2		**рефлюксний**	медична	прикметник	лат. refluo - текти назад
19760	1	рефрактерний		медична	прикметник	лат. refractarius - впертий, норовливий
19761	1	рефрактерними		медична	прикметник	лат. refractarius - впертий, норовливий
19762	6	рефрактерного		медична	прикметник	лат. refractarius - впертий, норовливий
19763	8		**рефрактерний**	медична	прикметник	лат. refractarius - впертий, норовливий
19764	3	рефракції		фізична	іменник	лат. refrāctus - заломлення
19765	3		**рефракція**	фізична	іменник	лат. refrāctus - заломлення
19766	3	рецепта		медична	іменник	лат. recipere - брати
19767	47	рецептом		медична	іменник	лат. recipere - брати
19768	50		**рецепт**	медична	іменник	лат. recipere - брати
19769	1	рецептор		біологічна	іменник	лат. recipere - брати
19770	2	рецептора		біологічна	іменник	лат. recipere - брати
19771	29	рецепторами		біологічна	іменник	лат. recipere - брати
19772	1	рецепторах		біологічна	іменник	лат. recipere - брати
19773	7	рецептори		біологічна	іменник	лат. recipere - брати
19774	5	рецептором		біологічна	іменник	лат. recipere - брати
19775	1	рецептору		біологічна	іменник	лат. recipere - брати
19776	55	рецепторів		біологічна	іменник	лат. recipere - брати
19777	101		**рецептор**	біологічна	іменник	лат. recipere - брати
19778	1	рецептурні		медична	прикметник	лат. recipere - брати
19779	1		**рецептурний**	медична	прикметник	лат. recipere - брати
19780	4	рецидив		загальнонаукова	іменник	лат. recidivus – той, що відновлюється

Fig. 1. Ukrainian database



	B	C	D	E	F	G
12172	3	receptora		біологічна	іменник	лат. recipere - брати
12173	1	receptorami		біологічна	іменник	лат. recipere - брати
12174	1	receptorow		біологічна	іменник	лат. recipere - брати
12175	6	receptory		біологічна	іменник	лат. recipere - брати
12176	12		**receptor**	біологічна	іменник	лат. recipere - брати
12177	34	recepty		медична	іменник	лат. recipere - брати
12178	34		**recepta**	медична	іменник	лат. recipere - брати
12179	1	ręcznikiem		загальновживана	іменник	
12180	1		**ręcznik**	загальновживана	іменник	
12181	1	ręcznym		загальновживана	прикметник	
12182	1		**ręczny**	загальновживана	прикметник	
12183	1	refluks		хімічна	іменник	лат. refluo - текти назад
12184	1		**refluks**	хімічна	іменник	лат. refluo - текти назад
12185	7	regularny		загальновживана	іменник	лат. regularis – рівномірний, правильної форми
12186	7		**regularny**	загальновживана	іменник	лат. regularis – рівномірний, правильної форми
12187	4	regulować		загальновживана	іменник	лат. regulare – впорядкувати
12188	4		**regulować**	загальновживана	іменник	лат. regulare – впорядкувати
12189	1	ręka		загальновживана	іменник	
12190	1		**ręka**	медична	іменник	
12191	7	registration		загальновживана	іменник	
12192	7		**registration**	загальновживана	іменник	англ.

Fig. 2. Polish database



	B	C	D	E	F	G
12877	3	Retinopathie		медична	іменник	лат. retina (сітківка) + pathia
12878	3		**Retinopathie**	медична	іменник	лат. retina (сітківка) + pathia
12879	2	retroperitoneale		медична	прикметник	лат. retro (назад) + старогр. περιτόναιον (очеревина)
12880	2		**retroperitoneal**	медична	прикметник	лат. retro (назад) + старогр. περιτόναιον (очеревина)
12881	2	Retrovirus		медична	іменник	лат. retro (назад) + virus (отрута)
12882	2		**Retrovirus**	медична	іменник	лат. retro (назад) + virus (отрута)
12883	1	rettet		загальновживана	дієслово	mittelhochdeutsch retten
12884	1	retten		загальновживана	дієслово	mittelhochdeutsch retten
12885	2		**retten**	загальновживана	дієслово	mittelhochdeutsch retten
				медична	іменник	лат. recipere - брати

Fig. 3. German database

The process of frequency dictionary compilation for the instructions to medical products included the following steps:

1. The absolute frequency of each lemma is calculated by the semi-automatic method.
2. Relative frequency of the words is calculated automatically.
3. Words are listed according to their frequency.

4. The ranks in the dictionary are singled out.
5. The lexical minimum of the vocabulary of words for the instructions to medical preparations is distinguished on the basis of separated ranks [20, p. 71].

So, the frequency dictionaries not only reflect the modern languages, but also contribute to linguistic research of terminology and its standardization.

2.3 Our methods

For a comparative analysis of the instructions to medical products, it is necessary to create a text resource of the linguistic information storage – a linguistic corpus which consists of three subcorpora: Ukrainian, Polish, and German instructions to medical products. [21, p.27]

The word forms are reduced to the base form. It indicates to which part of speech belongs each of the word form. In addition, the traditional approach to lemmatization, which is particularly important «for the quantitative study of lexemes in texts written in the synthetic languages (to which the Ukrainian belongs)» [18, p. 90] is used:

- Noun – all grammatical cases are reduced to the nominative case singular, for example: *medytsynoiy* – *medytsyni* – *medytsyna*. Plural nouns are reduced to the nominative case plural.
- Adjective – case singular and plural forms are given as the root form of the masculine gender. Adjectives in the comparative and the superlative degrees as well as suppletive adjectives also follow this rule.
- Verb – the synthetic forms of tense (present, past, future) are reduced to the infinitive form of the verb. According to S. Buk [19, p. 298] analytical tense forms are reviewed as syntactic forms and each component is considered as a separately registered word.
- Participle – case forms of singular and plural are given as the root form of the masculine gender.

Numerals, pronouns, prepositions, particles and interjections are not given in the dictionary, because they are not terminological units. Instead, meaningful commonly used words, such as nouns, adjectives, verbs, adjectives, adverbs and participles, are listed.

Our research has been based on the instructions sorted according to Anatomical Therapeutic Chemical Classification System; 195 instructions have been analyzed.

Thus, the frequency dictionary of Ukrainian language includes 1000 words (see Tab.1), but we shared in this article their English equivalents, the frequency dictionary of Polish language includes 1000 words and the frequency dictionary of German language includes 1000 words [21, p. 26].

Table 1. Ukrainian frequency dictionary.

Rank	Word	Absolute frequency	Relative frequency
1	patient	1604	1,879012
2	preparation	1514	1,773581
3	application	1462	1,712666

4	dose	1349	1,580291
5	treatment	1076	1,260485
6	mg	1062	1,244084
7	to be	1043	1,221827
8	to be able	752	0,880933
9	research	718	0,841104
10	trace	665	0,779017
11	reaction	564	0,6607
12	therapy	552	0,646643
13	impairment	525	0,615013
14	patient	510	0,597442
15	ml	494	0,578698
16	level	492	0,576355
17	introduction	488	0,57167
18	clinical	483	0,565812
19	concentration	481	0,563469
20	to apply	430	0,503725

It is important that the dictionaries based on the corpora bring the users facts about the real functioning of the language, since the corpus-based studies of the language through the volume of the analyzed material and the technical capabilities of the corpus toolkit can reveal such linguistic realia that scientists have not even suspected to exist.

Since terms are coined in a professional environment and are mainly used in professional texts, in particular in the instructions to medical products, it is possible to introduce special meaningful components which have specific meanings and thus can perform the classification function.

Undoubtedly, in addition to one-component terms, the systematic nature is given to medical terminology by complex terms, which are formed by borrowed components (which can be located both in the preposition and in the postposition) with the nominal basis.

The main trends in term coining, found in the frequency dictionaries to the instructions to medical products, are one-component terms and composite terms.

Composite terms have an advantage over term combinations, since they are more economical and they serve as a word-building base for the derivative words: ukr. hinekolohiia – hinekoloh – hinekolohichnui. Components, used in the language as independent words (akusher-hinekoloh), are used to coin complex and compound terms, as well as components that are not used independently, such as: lat. -algia, auto-, bio-, gastro-. Thereby, it is necessary to describe not only independent words in the dictionary, but also elements of the terms. Since, focusing on international elements of term and morphemes, derived from Latin words, the user of the dictionary can better remember, or even understand the name, still unknown to him. For

example, lat. rhil - 'inclination', lat. -lyt- 'destroying', lat. -haem- 'blood', lat. -troph- 'nutrition', lat. -gen- 'producing', lat.'- oid-' similar'.

It is obvious that international term elements and morphemes function only in the Ukrainian and German instructions. This is especially noticeable after comparing terms given in the frequency dictionaries to the instructions to medical products. For example: a suffix of Latin origin lat. itis – ukr. -it-(-yt-) represents the meaning 'inflammation': ukr. apendytsyt, pol. zapalenie wyrostka robaczkowego, germ. die Appendizitis 'inflammation of the appendix'. While compilers of Ukrainian and German instructions and dictionaries use the international Ukrainian suffix -it-(- yt-) or German – itis, in Polish sources there is a term derived on a national basis.

A certain problem is caused by the transfer of terms-eponyms in the Ukrainian language, since under the influence of extralinguistic factors, the compilers of Ukrainian-language instructions use Russian transliteration.

As for the German umlauts, we follow the instructions of the Ukrainian-Latin-English Medical Encyclopedic Dictionary compilers [1], which emphasize on the necessity to reproduce the sound image, since the German umlauts do not have exact equivalent in the Ukrainian phonetic system. Thus, the most similar sounds must be chosen: o - labial [e] – ukr., [ü], u – labial – [i] – ukr.: Gete, Shenbakh, Shreder, Minkhen (not Miunkhen). When original sound image is chosen during translation, we get the most accurate pronunciation of the word.

The terms-eponyms, for example: ukr. – Bazedova khvoroba - pol. choroba Basedowa – germ. Basedowkrankheit / Morbus Basedow (in German-language instructions, terms-eponyms are represented not only by term combinations, but also by composite terms). Obviously, the use of terms-eponyms is practical for specialists in the medical and pharmaceutical industries, but embarrasses the non-professionals. That is why it is necessary to indicate a synonym for such a term, or to explain it.

Due to the quantitative indicators in frequency dictionaries, it was discovered that the terms-eponyms are used differently by the compilers of the Ukrainian-language instructions: S. Stivensa or S. Styvensa (according to the system of Russian transliteration). Existence of two variants of the writing of one term-eponym indicates the need to standardize the Ukrainian rules of transliteration.

The surname components have two- or more - stems and also include a root word that is the general name: the syndrome and proper name (surname or surnames) of scientists, which semantically specifies the term. Dual transliteration, both on the basis of graphical (visual) and sound images in the text of instructions to medical products, lead to inaccurate reproduction of terms.

Thus, frequency dictionaries provide information on the frequency of the use of terminological synonyms (doublets) and variating forms. Data on the frequency and consistency of term combination as well as on the functioning of international term elements are important. To identify lexical minimum and to compile terminology dictionaries, information about the kernel and the periphery is highly important, because such data indicate the 'archaization' / 'neologization' of the system.

Compiling terminological dictionaries the peculiarities of the term formation of each particular terminology should be taken into account, therefore the compilation of

frequency dictionaries is important because they represent not only professional terminology but also reflect the present state of language.

3 Conclusions

The source of the research is based on the instructions sorted according to Anatomical Therapeutic Chemical Classification System. 195 instructions to medical products have been analyzed.

Linguistic research corpus contains three sub-corpora: Ukrainian, Polish and German. Frequency dictionaries contain 3000 words that are specific to medical products.

The results of the study confirm the need for further research on medical terminology. The processed material allows not only to reveal the peculiarities of Ukrainian, Polish and German terminology, but also to understand better the differences between Ukrainian, Polish and German cultures. It is necessary to continue improving the standardization of medical terminology and compiling dictionaries of various types on the basis of instructions to medical products.

The current state of intercultural communication and the world globalization requires such studies on the material of other languages.

References

1. Petrukh L., Holovko I.: Ukrainian-Latin-English Medical Encyclopedic Dictionary, VSV "Medicine", Kyiv (2012).
2. Badziński A.: Medyczny słownik kolokacji polsko-angielski-angielsko-polski, Warszawa (2011).
3. Nechai S.: Russian-Ukrainian Medical Dictionary with foreign names, Fund "Third Millennium", Kyiv (2000).
4. Baranov A.: Automation of linguistic research: the body of texts as a lingual-caustic problem. In: Russian studies today, vol. 1-2., pp. 179-191. (1998).
5. Baker M.: Corpora in translation studies: An overview and some suggestions for future research, pp. 223-243. Target (1995).
6. Hardie A, Baker M, McEnery T.: A Glossary of Corpus Linguistics, Edinburgh Univ Press, (2006).
7. Buk S., Levus Ye., Yavorskyi Ye.: Algorithm for reflecting changes in the lexical saturation of the text. In: Scientific journal of Lviv Polytechnic National University, vol. 771, pp. 349-353. Lviv (2013).
8. Karpilovska Ye: Introduction to Applied Linguistics: Computer Linguistics, LLC "Southeast", Donetsk (2006).
9. Maksymiv O.: Corpus of the Persian language as a source of material for the frequency dictionary, Eastern world,vol. 4, pp. 109-114. (2008).
10. Shyrokov V., Buhakov O., Hriaznukhina T., Kostyshyn O., Kryhin M.: Corpus linguistics, Dovira, Kyiv (2005).
11. Demska O.: Text Corpus: the idea of another form, VPTS NaUKMA, Kyiv (2011).
12. Świdziński M.: Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy. In: LingVaria, vol. 1 pp. 23-32. (2006).

13. Sinclair J.: Corpus, Concordance, Collocation, Oxford University Press, Oxford (1991).
14. Regulatiry and directive documents of the Ministry of Healthcare of Ukraine, www.mozdocs.kiev.ua/liki.php, last accessed 2018/03/21.
15. Rejestr Produktów Leczniczych, http://pub.rejestrymedyczne.csioz.gov.pl.375, last accessed 2018/03/21.
16. European Medicines Agency, www.ema.europa.eu/drugs, last accessed 2018/03/21.
17. Kwic, http://www.chs.nihon- u.ac.jp/eng_dpt/tukamoto/kwic_e.html, last accessed 2018/04/02.
18. Buk S.: Statistical characteristics of the novel “Basis of the Society” by Ivan Franko (based on the frequency dictionary of the novel). In: Scientific journal of Lviv Polytechnic National University, vol. 676, pp. 90-93. Lviv (2010).
19. Buk S.: Quantitative parametrization of Ivan Franko’s texts: project and its realization. IIn: Scientific journal of Lviv Polytechnic National University, vol. 58, pp. 290-307. Lviv (2013).
20. Perkhach R.-Yu.: Terminology in patient information leaflets: cognitive linguistic and cultural linguistic aspects (based on Ukrainian, Polish, German): dys. kand. filol. n.: 10.02.15, Odesa (2017) .
21. Perkhach R.-Yu, Shyika Y.: The methology of frequency dictionaries to the instructions to medical products. (based on trilingual corpus) In: Materials of 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT 2018), vol. 2, pp. 26-30. Lviv (2018).