Branch Lexical Catalog as an Independent Genre of Educational Lexicography

Olga Ilyina, Marina Zolotareva

Bauman Moscow State Technical University

ABSTRACT

This article is written in the aspect of teaching a foreign language for special purposes when the teacher himself is often quite far from the subject he usually teaches. The article considers the genre of the branch of the lexical catalogue. Educational lexicography traditionally works with different types of restrictive lexical lists – frequency dictionaries, lexical minima, thematic dictionaries, etc. The branch lexical catalogue differs from other genres in its content and purpose. It is carried out to make the lexical material of a certain subject area transparent, far from natural everyday communication. The main task of this genre is to catalogue all the vocabulary that is included in the corpus of texts of a certain subject-communicative area. The development of the industry lexical catalogue is carried out using automated data processing and subsequent manual revision. As a result, teachers receive a complete list of words that are in demand in a specific professional field. This list of words can be ordered alphabetically or by frequency. In the second case, it can be used to output a frequency dictionary of specialized vocabulary. The process of creating a branch lexical catalogue is given for the speciality "Rocket Science".

Keywords: Corpus-oriented lexicography, Lexical approach in foreign language teaching, Language in specific purposes (LSP)

1. Introduction

The material of this publication concerns two thematic areas – teaching a foreign language for special purposes (LSP) and the development of restrictive lexical lists for foreign language learners. The intersection of the two regions have arisen by chance: learning a foreign language for special purposes in the same way as learning a foreign language for general knowledge, needs resolving the issue of selection of teaching content, and the amount of foreign language vocabulary is an integral part of the concept of "learning content".

Most research in the field of modern pedagogical lexicography as an independent branch of linguodidactics aims to develop various types of restrictive lexical lists for the study of a foreign language of general proficiency and not for special purposes. In the practice of teaching Russian as a foreign language (RFL), a similar situation has developed: a stable position has been formed on the allocation of the lexical base for different levels of RFL proficiency, fixed in lexical minima. But few branch RFL lexical minima catalogue the vocabulary of certain subject areas. They are all designed to solve a particular lingua didactic problem of teaching international students at a particular Russian university.

Modern means of automatic information processing make it possible to simplify and speed up developing lexical minima in the Language of the speciality; the main thing is to understand how to organize this work.
It is known that restrictive lexical lists for learning a foreign language of general proficiency are created on the basis of data from frequency dictionaries and analysis of national language corpora. Thus, modern pedagogical lexicography uses such genres as dictionaries and various types of restrictive lexical lists – lexical minima, terminological minima, frequency lists of words, etc.

What can serve as a source base for creating lexical minima of LSP if there are no frequency dictionaries for LSP? This article attempts to describe such a genre of pedagogical lexicography as a branch lexical catalogue. The creation of a branch lexical catalogue is the first stage in developing a branch lexical minimum. The following is material on developing a branch lexical catalogue in the speciality "Rocket Science" is presented.

2. Literature Review

Let's look at what genres are most in-demand in educational lexicography.

Systematic research in this area began in the twentieth century. However, earlier attempts were made to determine the lexical base for learning a foreign language. The predecessors of modern lexical minima can be considered a dictionary compiled in 1643 by the famous Czech teacher Jan Amos Komenský in order to study the Czech language. It was called "The World of sensible things in pictures" and was compiled on the principle of a thematic reference book (van Amos Komenský, 1967). It included such parts as "House", "Man", "City", "State", "Marriage union", "Feast", "Elements", "Journey", "Animate and above all birds", "Four-legged and above all domestic animals". Each thematic part contained a list of words related to the stated topic, definitions of words, and illustrations to complement the interpretation.

Komenský's book creates a general outline of future editions of educational lexicography: the most significant vocabulary from the point of view of solving communicative tasks is selected. The vocabulary material is presented in thematic groups and accompanied by the necessary illustrations that perform a semanticizing function. Subsequently, these provisions are transformed into the principles of organizing lexical work in a foreign language class and, in particular, into the principles of developing lexical minima.

In contrast to academic lexicography, educational lexicography seeks to minimize the volume of vocabulary material, separating the lexical core, the elemental composition of which should be sufficient for successful human communicative activity. Therefore, the main genres in this area are various restrictive word lists – dictionaries, lexical and terminological minima, frequency dictionaries, thematic word lists, etc.

It can be assumed that the first restrictive word lists began to appear in the twentieth century. In the United States, dictionaries containing the frequency vocabulary of the English Language were published, among them "Teachers word book" (Thorndike, 1921) and "Dictionary of basic English" (Ogden, 1930). The authors of both publications relied on the principle of frequency and relevance for everyday communication when selecting words.

In our time, the compilation of frequency dictionaries is also an urgent task. Although the creation of frequency dictionaries is now a linguistic and not a lingua didactic task, they are actively used to teach a foreign language. Modern automated information processing systems make it possible to work with large bodies of texts, allow for more accurate calculations, and complex computer algorithms help avoid many problems when counting, for example, perceiving word forms as a single word. Modern frequency dictionaries promise a high density of coverage of any text. So, R. Johns and E. Tschirner, in 2006, based on a multimillion-dollar corpus, compiled a frequency list of German words. It included 4,000 units and promised a non-specialized text coverage density of 80-90% (Johns, Tschirner, 2006).

In addition to frequency dictionaries, there is another equally common genre of educational lexicography – lexical minimum. Much has been done in this area by Russian scientists because Russian educational lexicography has focused on developing lexical minima since the middle of the XX century since the beginning of development. At various times, lexical minimums were created for national schools of the Soviet Union and international students of Russian universities, where much was done by N.M. Shanskij (Shanskij, 1976), V.V. Morkovkin (Morkovkin, 1985). Currently, lexical minima have been developed for five levels of general proficiency in Russian as a foreign language (Andryushina, 2011). There are also industry-specific lexical minimums created within the framework of the LSP, for example, the Lexical Minimum for training in mathematics, physics, biology and Chemistry (Artemyeva, 2004), Lexical minimum of mathematical terms (Zharov, Klimova, Kuznecova, 2003), Dictionary of chemical terms (Garshin, Morkovkin, 2017), Lexical minimum in the speciality "Robotics" (Ilyina, Zolotareva, 2019), etc. It should be noted that the creators of all lexical minima believe that the frequency principle is the main one for including words in the minimum. Still, they also resort to subjective assessments of the demand for words in communicative situations.

Comparing different methods of creating restrictive word lists, Muravyov and Ol'shevskaya conclude that the greatest density of coverage of non-specialized texts is realized by those minima that rely on statistical methods and the frequency principle (Murav’ev, Ol’shevskaya, 2019).
3. Methodology

Teaching a foreign language for special purposes differs from the model of teaching a general foreign language (Škoricova, Orlov, Romanova, Pinevich, 2019; Pinevich, Petrova, Orlov, Romanova, 2019). These differences are noticeable both at the initial stage of the educational process-goal setting and subsequent steps. They are also manifested in the content of training, in the organization of the lesson, in the ways of mastering the language material by students.

The external circumstances of the pedagogical process also affect the process of mastering a foreign language for special purposes. These include the density of training sessions per day, per week, and per semester, students' immersion in the academic environment, immersion in the language discourse of the speciality, and interaction with colleagues who choose a foreign language for professional communication.

Russian as a foreign language for special purposes is mostly associated with universities where international students receive Russian professional education. It should be noted that we are talking about a large group of foreigners. Statistics show that in 2019, about 300 thousand international students were studying at Russian universities. Further growth in the number of foreigners studying in various professional higher education programs is expected (Gohberg, Ozerova, Sautina, SHugal’, 2020). Although Russian universities offer different training programs for international students, there are nevertheless common features that are characteristic in general for the model of linguistic support for international students in Russian universities.

When international students move to the main stage of study at a Russian university, classes in the Russian Language do not stop, but their number decreases. International students study mainly Russian as a tool for gaining knowledge in their chosen professional field, so the main course for teaching Russian at a Russian university is LSP (Ijina, 2018).

Since Russian teaching to foreigners in Russian universities focuses on the LSP aspect, teachers are forced to conduct research in selecting and minimizing vocabulary for study independently.

As shown in the literature review, to identify the nuclear vocabulary, which becomes mandatory for foreigners to study at certain foreign language proficiency levels, methodologists rely on dictionaries of frequency vocabulary. The principle of frequency is recognized as the main one in the compilation of lexical minima by all scientists. Mathematical methods have become mandatory in modern corpus studies (Sidnyayev, Butenko, Garazha, 2019; Sidnyayev, Butenko, Bolotova, 2020). Thus, the stage preceding the development of the lexical minimum is cataloguing the vocabulary of a certain part of the Language. To do this, the dictionary base of the Language is first formed, and then the subsequent calculation of the uses of the detected words takes place. The formation of the dictionary base is the most significant part of the project and the most controversial at its implementation stage. Even now, when scientists have automated programs for the derivation of linguistic statistics and when it is possible to study the multi-million word-use cases of national languages, the question remains of the correctness of the selection of text material is subsequently split into individual words, included in it.

Suppose we are talking about lexical minima of general proficiency. In that case, cataloguing and calculating vocabulary is a linguistic task, not a methodological one, since the developers of lexical minima rely on the frequency dictionaries of the national Language and other lexicographic publications. The methodology for developing frequency dictionaries based on big data processing results is based on mathematical principles.

Thus, we will fix the stages of the complex path of developing lexical minima:

- processing large amounts of language data to identify the lexical core-frequency vocabulary;
- definition of the principles of the development of lexical minima for the purpose of teaching a foreign language
- Minimizing the national Language vocabulary for lingua didactic purposes, recorded in restrictive lexical lists, for example, lexical minima.

When it comes to teaching a language for special purposes, teachers working in Russian universities need to independently engage in lexicographic work and make lexical minima for the speciality language. Moving along this route involves creating a corpus of texts and cataloguing its vocabulary. To understand what needs to be reduced and to what minimum, you must first accurately describe the boundaries of the compressible material. To do this, it is advisable to make a preliminary branch lexical catalogue.

The development of the branch lexical catalogue for the speciality “Rocket Science” is presented in the next section.

4. Results

The branch lexical catalogue is an alphabetically ordered complete list of words that are in demand in a specific area of human communicative and speech activity. Most often, this area is the sphere of specific professional communication. It can also be assumed that lexical catalogues may reflect the vocabulary space of communication of certain social groups in a certain historical period. The branch lexical catalogue genre could also include
lexicographic publications that describe the vocabulary of the texts of the masters of the artistic word, for example, “Pushkin’s Dictionary of the Language” (Vinogradov, 2000).

The main purpose of the lexical catalogue is to list all the words used within the same communicative sphere. Passportization of vocabulary allows you to calculate the number of words and describe the morphological landscape and popular word-formation models. The lexical catalogue genre involves creating a list of words without writing dictionary entries and posting an accompanying linguistic commentary.

In a general sense, the lexical catalogue should be considered an independent lexicographic genre because it performs a unique function and has a transparent and uniform structure.

Let’s consider the process of creating a branch lexical catalogue for the speciality "Rocket Science".

4.1. Compilation of the Corpus of Texts

Work on any branch lexical catalogue begins with the definition of the principles of creating a database of language material.

The project described here is about teaching international students communication in Russian within the framework of engineering discourse, so the source base should be arrays of texts (oral and written), which together reflect the thematic field of this communicative sphere. In the described case, we consider the professional, communicative sphere of "Rocket Science" and the educational and professional one. The educational process in any declared speciality at the university involves the student’s classroom and independent work. In the first case, the communicative actions will be performed mainly orally, in the second – in writing. It is physically difficult to collect and process the corpus of oral texts created by teachers and students due to educational interaction, so the source base of the industry lexical catalogue in the speciality "Rocket Science" was written texts.

The principles of formation of the corpus

- ✓ Together, the collected body of texts reflects the subject field of the speciality "Rocket Science".
- ✓ Textbooks and manuals are used as publications that contain the necessary amount of knowledge in this speciality.
- ✓ Textbooks and manuals should be written by teachers of specialized departments responsible for implementing training in this speciality or recommended by them for use in the educational process.
- ✓ Books should be written by different authors to avoid the consequences of the manifestation of idioystyle.

The corpus of texts is compiled on the basis of 34 sources in the speciality "Rocket Science", the total volume of which is 3306 pages. The total volume of the final branch lexical catalogue is 9490 words.

4.2. Preparation of the Dictionary Material of the Catalog

To obtain a complete list of words presented in the corpus of texts in the speciality "Rocket Science", we used the Wordstat automatic vocabulary processing program. This program was created by Russian developers; it allows you to catalogue the lexical composition of large text arrays by alphabetical or frequency characteristics.

The program copes with the technical part of the work, automatically splitting the processed texts into words. To complete the creation of the catalogue, you need to process it manually. Processing involves several directions. The first direction is eliminating technical errors, such as combining word forms that the program mistakes for independent words, getting rid of fragments of words included in the list due to the machine’s misunderstanding of word transfer algorithms.

The second direction is linguistic. Since the genre of the described product involves cataloging only words, in the process of manual processing, all elements that are not a word are removed, for example, numbers or combinations with numbers – the 12th, 8th, etc.

The third direction is the lingua didactic. Since the lexical catalogue is understood as the basis for creating a lexical minimum for learning the Russian Language, all words written in foreign languages are removed from it. Proper names written in foreign and Russian languages and adjectives formed from them, such as Princeton, Moscow, are also excluded.

Russian as a foreign language is an integral part of the course of Russian as a foreign language at the Russian Technical University and is thought of as a logical continuation of the course of the Russian Language of general use. Russian basic grammar study precedes the study of the Language of the speciality; therefore, by the time of transition to the LSP course, non-complementary morphological classes of words of the Russian Language are fully known to foreign speakers, so lexical units related to such morphological classes can be taken out of the created catalogue.

All numerals are taken out of the created catalogue – quantitative and ordinal: two, fifteen, fiftieth. But the numeral one is preserved since it can be homonymous to the indefinite pronoun one (meaning "some, some") and is represented in this meaning in the source text material. The numeral both is also preserved since it is not perceived as a typical numeral by foreign speakers.
Closed morphological classes are not included in the catalogue:
particles, for example: whether, would, let, even, not, nor, only
non-derived compositional conjunctions, for example a, but, and, or
non-derivative prepositions, for example: on, for, of, under, before, at, in, for, o, from, except, over
personal and possessive pronouns, for example you, he, my, our, their, his, hers.

The following words are not included in the catalogue: American, German, German, Irish, Canadian, European, African, etc. The inclusion of words with similar semantics in the catalogue is impractical because they also represent a closed class of words. Their number is limited by the number of ethnic groups living in the world. However, unlike grammatically closed classes, this lexical group is very voluminous. In the studied texts, its elements are random and reflect the author’s subject preferences. The inclusion of such thematically related lexical groups in lexical minima is possible if it is systematic and not accidental. This means that for the formation of the lexical minimum in the speciality “Rocket Science”, this group of words will not matter.

4.3. Errors

When developing a lexical catalogue, some errors are unavoidable. Some of them are related to technical errors in the program’s recognition of words in the source text. According to subjective estimates, we have to admit that part of the vocabulary of the corpus of texts is lost. However, according to subjective assessments, this part is less than 1% of the total volume of the corpus.

Another part of the errors is due to the grammatical features of the Russian Language when it is impossible to determine the original form of a word without context. Therefore its partial status since the same word can belong to different parts of speech. Let’s look at some examples.

Simple (Russ. простой) can be both an adjective – simple (простой, adj.) example, and a verbal noun – downtime (простой, verbal noun) of technique

we measure (Russ. измеряем) can be a personal form of the verb to measure and a short participle from the full-measured (Russ. измеряем).

And, finally, without context, it is impossible to determine whether the words unacceptable, unstable, etc., are an adverb or a short participle or a short adjective. The solution to these grammatical problems is performed by manually checking the contexts in which these words are used in the corpus.

5. Conclusion

We consider the branch lexical catalogue as an independent genre of educational lexicography. It is a complete list of words contained in the corpus of texts under study and is a kind of dictionary passport of this corpus. This work is carried out to improve the effectiveness of teaching a foreign language for special purposes. LSP teachers need industry-specific lexical catalogues and industry-specific lexical minimums since the existing lexical minimums are usually designed to teach a foreign language of general use. The speciality language is associated with a narrow linguistic area and is of interest only to a limited contingent.

Knowing the methodology for creating industry-specific lexical minimums, LSP teachers can create them independently, using modern means of automated information processing. The development of a lexical catalogue precedes the stage of vocabulary minimization, as it allows you to cover the boundaries of the lexical space of the studied discourse, to establish its quantitative composition, to determine the frequency vocabulary, to see the word-formation and morphological features of the lexicon.

References


Morkovkin V.V. (1985) Leksicheskie minimumy sovremennogo russkogo yazyka Moskva, Russkij yazyk