BMSTU Corpus of Scientific and Technical Texts: Conceptual Framework

Iuliia Ivanovna Butenko¹,
Vladislav Vladlenovich Garazha²

¹Department of Roman-German Languages, Bauman Moscow State Technical University, Russia
²Department of English for Industrial Engineering, Bauman Moscow State Technical University, Russia

Corresponding Author: Iuliia Ivanovna Butenko
E-mail: iuliibutenko2015@yandex.ru


Received Date: December 12, 2020
Accepted Date: January 22, 2021
Online Date: March 5, 2021

Publisher: Kare Publishing
© 2021 Applied Linguistics Research Journal
E-ISSN: 2651-2629

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

ABSTRACT

The paper describes the conceptual framework for developing the BMSTU corpus of scientific and technical texts. Scientific works on corpus linguistics are reviewed. The theoretical foundation is based on the collected data. The main methods include both direct and indirect analyses. The direct analysis implies interaction with a corpus itself. Publications devoted to corpora are considered as a part of the indirect analysis. The main characteristics of scientific and technical texts taken into account are highlighted. They include compositional characteristics of the analyzed texts and structures of terms. The compositional structure of scientific and technical texts is supposed to be solved by adding additional structural mark up that would show the place of each text element in the document. It is offered to create a database of terminological word combinations to automate the marking in the corpus. The lexical entry framework, which describes the semantic characteristics of each term, is given in detail. The corpus of technical texts being developed is considered to become a tool for Bauman Moscow State University researchers, teachers and students.

Keywords: corpus linguistics, scientific and technical texts, markup, lexical entry, compositional structure of a text.

Introduction

Corpus linguistics has come a long way since its birth in the 1960s. Its terminological base was developed in the subsequent period. However, views on how this field should be classified differ up to this day. In his work, Corpus linguistics: the specifics of modern meta descriptions of language (Barkovich, 2016), A.A. Barkovich introduces corpus linguistics as a metalinguistic tool and a methodology for the design, use, and study of corpora for the metalinguistic interpretation of the speech patterns with computer tools; technical progress contributed to the design of corpus linguistics as a special field of science. It became possible to transform various data blocks into corpora and create databases with the help of computer technology. According to I.A. Pakholkova, the mission of corpus linguistics is to develop general principles for the construction and use of linguistic data through the use of computer technology. She noted that the growing amount of data necessitated the structuring of the language material, particularly in the form of corpora. The texts corpus is a corpus of data, the units of which are texts or parts of them (Pakholkova, 2012).

The nature of the text corpora is discussed in detail in the book by V.P. Zakharov and S.Yu. Bogdanova Corpus linguistics. The authors review the rationale behind the building of the text corpora. First, a sufficiently large and balanced corpus provides a representative picture of linguistic phenomena.
Second, the data in the corpus is compiled in its natural contextual form, which increases the objectivity of the studies. Third, the compiled and processed data set can be used multiple times by different researchers and different purposes (Zakharov, & Bogdanova, 2020). Besides the data array can become a basis for studying the properties of the language. It is possible to carry out statistical studies to build linguistic models only with a sufficient sample size found in the database (Sidnyaev, Butenko, & Garazha, 2019).

Corpora can be classified in accordance with various criteria. Still, if the goal is to build a specific corpus, it is important to determine how comprehensive it is expected to be. As stated in G.V. Kolpakova, Linguistics and lexicography, the small special cases are processed manually, and categorization significantly impacts the results of analyses. However, in large corpora, where categorization is even more important, automatic processing is used. This necessitates the marking of corpus data by parts of speech (Kolpakova, 2011). The variability of corpora is reviewed in N.V. Kozlova's article Linguistic corpora: definition of basic concepts and typology where corpora classification is described and explained in detail (Kozlova, 2013).

As always, it is important to consider the experience of predecessors. Considering this, special attention is paid to works dedicated to the current national corpora. Despite the fact that they have a different focus, some issues are common in all cases. This includes the task of resolving ambiguity. Finding a solution to this issue would allow us to introduce semantic filters into the corpus. The problem of semantic markup and filters was considered in researchers dealing with the Russian National Corpus (Kustova, G. I., Lyashevskaya, O. N., Rakhлина, E. V., & Shemanova, O. Yu., 2006).

The study of corpora of scientific and technical texts requires consideration of works with respective specifics. A.C. Fuentes analyzed the vocabulary aspect of academic and technical corpora, which makes his work especially important concerning the objectives of our research (Fuentes, 2001). Along with this, it is worth citing an article that also attracted attention for its subject orientation. In the article by Vjolica Belegu-Caka, the use of corpora is revealed from the standpoint of using English for special purposes teaching (Vjolica, 2001).

The Corpus manager approach was reviewed in work by D.Sh. Suleimanov and D.R. Mukhamedshin (Suleimanov & Mukhamedshin, 2018). Similar aspects were studied in the review of corpus capabilities by Z.B. Dolgikh. The latter provides his definition of the corpus manager based on the paper as mentioned above by V.P. Zakharaeva and S.Yu. Bogdanova (Zakharaeva & Bogdanova, 2020). The corpus manager means a specialized search system that includes software for searching data in a corpora or corporea to obtain statistical information and provide the user with results in a convenient form (Dolgikh, 2018).

Scientific and technical texts deserve special attention. The lexical and grammatical form of the language in such texts is generalized and has abstract nature. The approach to annotating and marking up the corpus can be adjusted according to such cases. Users will need to work with scientific and technical texts to conduct research in different linguistics fields but use the corpora as the source for data and knowledge bases when developing natural language processing software or its components. As an example, we can see the expert's activity in the field of software certification. One of the stages of certification is normative profile forming—which is searching for specific requirements for software being analyzed. The collection of normative documents and standards to be studied is relatively large and may cause many difficulties, especially for experts (Butenko, 2019).

Currently, the corpus linguistics is represented by national corpora for the Russian, English, Hungarian, Slovenian and other languages, as well as various specialized corpora, such as A.S. Gribojedov's corpus, the corpus of Latin texts Perseus etc. A researcher can use one or several corpora to source data. At the same time, organizations such as private companies and universities continually conduct research and select tools in accordance with their needs. Unfortunately, there is no corpus of Russian scientific and technical texts available for users. The constant growth of scientific and technical texts requires developing tools for creating a corpus of scientific and technical texts. The building of such corpus would contribute to the standardization of terminology, terminological dictionaries development, databases for automatic processing technical documentation.

Consequently, it becomes necessary to create marked-up collections of scientific and technical texts asssures for further research of terminological units. This poses other challenges, such as determining the type of markup and considering ways to implement it. Thus, this paper aims to develop a concept of annotated BMSTU corpus of scientific and technical texts.

Methodology

The methodology includes a review of scientific literature in the field of corpus linguistics in Russia and internationally. Among them are the articles that show the design of already existing corpora. While collecting research materials main focus was on modern sources. Corpora were selected by potentially having suitable functions and means for the corpus of scientific and technical texts. Such corpora do not have to be similar to the one we design in all aspects. For example, some British National Corpus principles may be suitable, despite the fact that it is
a corpus of English texts. The corpus features concerning the markup of Russian texts are demonstrated in the works devoted to the Russian National Corpus.

The analysis of publicly available corpora and the synthesis of the theoretical basis for the development of a new corpus were carried out. The study used direct and indirect analysis. The result of the direct analysis is the material obtained from working directly with the analyzed corpora. This approach was not applied to all corpora, as some of them do not have free access. In some cases, the database could be operated via the corpus web page. The objects of indirect analysis included publications on corpus linguistics. Articles on the topic of a single corpus reveal the theoretical foundations behind its development. These works are reliable sources since the article is devoted to the corpora developed, among others, by the cited authors.

At the same time, we aim to take into account the specific characteristics of scientific and technical texts that may influence the design and functions of the BMSTU corpus of scientific and technical texts. Such specific characteristics of scientific and technical texts include the compositional structure of texts and wide use of terminology that requires developing additional software tools for their annotation and representation in the BMSTU corpora of scientific and technical texts.

Results

As a result of researching the corpora and related works, it was possible to identify several tools and principles that would help develop the BMSTU corpus of scientific and technical texts. In the study, priority was given to the principles stated in work by M.V. Kopotev and A. Mustajoki (Kopotev, &Mustajoki, 2003). Their publication is dedicated to the creation of the Helsinki Annotated Corpus; thus, the principles presented there have been slightly adapted (see Table 1). Not all approaches were found to be suitable for the corpus under development. In the Turkish national corpus, all texts are divided into ‘imaginative’ and ‘informative’ (Akşan, 2012). This division is not helpful for a corpus of technical texts since all texts would be informative. Informative texts - texts from scientific literature, while creative texts refer to fiction.

<table>
<thead>
<tr>
<th>Table 1. Annotated corpus design principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wide usability for various audiences</td>
</tr>
<tr>
<td>Grammatical coverage priority</td>
</tr>
<tr>
<td>Focus on multilevel grammatical information</td>
</tr>
<tr>
<td>The choice of linguistic legacy</td>
</tr>
<tr>
<td>Multiple interpretations of one language unit</td>
</tr>
</tbody>
</table>

The development of the BMSTU technical texts corpus reveals a number of specific characteristics of such texts, which are to be considered during the corpus compilation. These features include the compositional structure of texts, the availability of many terminologies, and the need to develop software tools to represent them.

The texts to be added and annotated in the corpora are technical educational and instructional literature, technical manuals, scientific articles published in specialized journals, encyclopedias, the regulatory base, including various standards (Il'vina, & Zolotareva, 2019).

Some types of texts require additional structural annotation. They include legal texts, regulatory and technical documentation, description of standards in different subject fields. As an example, the compositional structure of standards is reviewed below. A text of a standard is a written rule, objectified in the form of a written document, consisting of a number of statements, united by various types of lexical, grammatical and logical connections, having a certain moral nature, pragmatic attitude and processed in a respective literary manner. An externally designed compositional structure distinguishes texts of standards as a class of documents. One of the means of formalizing this structure is rubrication, i.e. the division of the text into constituent graphically distributed parts and the use of generalized abstract lexical units in the presentation of the requirements.

The main part of the standard text is usually divided into sections, subsections, paragraphs, sub-paragraphs. It is mandatory to have headings of sections and subsections. Headings (subheadings) are an important means of rubrication and, therefore, an element of structural formalization. The structural elements of the standard are divided
into the following elements: title page, preface, content, introduction, title, scope, normative references, terms and definitions of concepts, designations and abbreviations, requirements to the object of standardization, appendices and bibliographic data. The title page is the first page; the preface is on the second page. The content includes serial numbers and headings of sections, applications with the designation of their headings. An introduction is provided when it is necessary to justify the reasons for developing the standard. The structural element scope is given to indicate the scope of the standard's applicability and, if necessary, specify the object of standardization. The structural element of the references contains a list of standards, which are referred to in the text of the standard. The terms and designations contain definitions of clarifying and defining the meaning of the terms used in the standard. Appendices may contain materials that complement the provisions of the standards. Appendices, for example, can be graphical materials, large-format tables, calculations, description of devices, algorithms, and programs. Appendices can be obligatory and informative. Informative appendices may be advisory and reference. The standard also specifies the bibliographic data. The structural elements, with the exception of such elements as Title Page, Preface Name, and Requirements for the object of standardization, are used if necessary, depending on the characteristics of the object of standardization. The titles of structural elements in texts of standards are short and characterize the object of standardization accurately. The titles of standards, as a rule, do not contain abbreviations, Roman numerals, mathematical signs, and Greek letters. The example above shows that while searching for information in the collection of a regulatory framework, the structural characteristics of the texts should be taken into account.

Another distinctive feature of the technical texts corpus is the terminological annotation. Lexical units are annotated according to the presence or absence of correlation with a certain subject area (Lesnikov, 2019). Moreover, it is necessary to specify the models of terms, which would allow searching for proper terms in the technical texts corpus. As for terminological annotation, it is necessary to consider multi-component terms, which the system should perceive as a single terminological phrase. For example, the phrase hydraulic amplifier should be understood collectively as a single terminological unit and not as separate lexemes hydraulic and amplifier. The terminology of the different subject fields contains many other features reflected in the annotation, so it is important to consider some paradigmatic features of aerospace terminology (Danilenko, 1977).

The automation of terminological markup is a necessary element of the corpus. It is a complex theoretical and practical task that requires both particular linguistics and programming methods (Sidnyav, Butenko, & Bolotova, 2020). The greatest difficulty in automatic markup is multi-component terms expressing a single concept but represented by several lexical units in their composition. The above circumstances indicate the need to create a super-corpus structural models database of terminological word combinations that will allow using the developed base for further terminological markup of technical texts corpus in different subject areas.

Thus, to simplify the approach to the markup of the corpus, it became necessary to select a well-established array of terms that could demonstrate the specific structural formations inherent in terminologies and term systems. For the array under study, we chose the term system of the Welding types subject area. On the one hand, it has been researched for a long time, and on the other hand, it was compiled mainly in the 20th century, which simplifies the selection and analysis of terms. Due to the fact that the terminological system of this subject area is sufficiently studied and standardized, a super-corpus database is expected to be created on its basis to describe the structural models of terminological phrases. Considering the formal structure of the terminological system elements of Welding types it should be noted that the most productive model is the combination of the nuclear element with a noun or adjective in the function of prepositional definition. This model is most clearly seen in two-component word combinations. Still, the analysis of more complex formations shows that the model left definition, attached to the core of the term, is also present in them, demonstrating generic characteristics. The further complication of the terminological word combination most often occurs with the complication of the postpositional definition, which carries the generic features. This approach will make it possible not only to research the particular characteristics of terminological phrase models but also to form all possible productive models of terminological phrases that will coincide with phrases from other subject areas.

Annotation of terms in the corpus requires the development of a terminological database containing information on each term and its already known or new meanings (Bozhenkova, Bozhenkova, & Mirzaeva, 2015). It is supposed to be an initial stage for further research on Russian terminology processing and standardization. It is supposed that each term in the annotated text will be linked to a certain dictionary entry. The structure of the dictionary entry can be represented in the form of eight:

T = \langle A, B, C, D, E, F, G, H \rangle

The elements of the structure:

A = \langle Ai, A2 \rangle – denotation of a concept, where Ai is the unique name of the concept or the name of the dictionary entry; A2 is the sign of the concept, which corresponds to the identifier of the frame.

B = \{B1, B2, B3, \ldots \} – is the set of concept definitions. There can be several definitions for one concept. Usually, a word is defined as a verbally expressed intensional enough to define an extensional to distinguish substantive and
formal definitions. The set of substantive definitions will be stored in the thesaurus in a natural language. Considering that a formal definition for a concept describes through other concepts and nominally through its own properties and relations to other concepts, the formal definition must be represented in both kinds of the thesaurus.

\[ C = <C_1, C_2, C_3, C_4> \] – the conceptual object set of types being described, in which four types are distinguished according to Dahlberg:
- \( C_1 \): essence: tangible and intangible objects, ways of looking at them;
- \( C_2 \): properties: quantitative, qualitative, relational (relations);
- \( C_3 \): actions: operations, processes, states;
- \( C_4 \): values (dimensions): time, position, space.

\[ D = <D_1, D_2> \] – a pair of sets of properties of a concept, where \( D_1 \) – set of qualitative properties; \( D_2 \) – set of quantitative properties.

\[ E = \{E_1, E_2, E_3, \ldots\} \] – set of notions, describing methods/functions, peculiar to a given notion, and reflecting pragmatics, connected with a given notion.

\[ F = \{F_1, F_2, F_3, \ldots\} \] – set of concept synonyms, or, in other words, set of concepts having quantitative relations (identity relation) with a given concept.

\[ G = \{G_1, G_2, G_3, \ldots\} \] – the set of correlates, or, in other words, the set of concepts having the relation of opposition to a given concept.

\[ H = <H_1, H_2> \] – a pair of sets of concepts having qualitative relations with the given concept, where:
- \( H_1 \) – the set of concepts that constitute a generalization relation with the data \( <H_1, H_2> \);
- \( H_2 \) – the set of species concepts;
- \( H_3 \) – the set of concepts constituting the aggregation relation with the data \( <H_1, H_2> \);
- \( H_4 \) – the concept that is “whole” concerning the describable, \( H_5 \) – the set of concepts that are a part of the describable.

Such detailed representation of dictionary entry will allow researchers to study paradigmatic features of terms and see their place in the term system (Miller, & Biber, 2015). Further representation of terms in the dictionary has several advantages: any dictionary entry can be updated at any time, and the updates will be available for users online.

**Discussion**

Several challenges were encountered in the course of the research. The development of the optimal design for the corpus of scientific and technical texts is a challenging task, as there are few examples and analogues to make a comparison for such a corpus. The development of conceptual foundations for the corpus of scientific and technical texts gives an insight of what features should corpus have at the end.

There are many more challenges to be solved to prepare the BMSTU corpus for users' scientific and technical texts. The stage of concept design is completed, and now we are working on the software for the corpus markup system and corpus concordance. These stages require working with programmers, and a number of technical problems arise. For example, there is a need for techniques to mark up terminological word combinations as part of its software realization.

At the same time, we have to make the corpora applicable for many practical tasks like teaching academic writing, conducting research in linguistics, serving as a database for natural language processing tools, developing corpus linguistics, computational linguistics, and lexicography, etc.

Another goal of the BMSTU corpus of scientific and technical texts is to fulfill the educational function. From this point of view, it would be appropriate to include the useability feature. It would allow a user to refer to the list containing words sorted by usage in different fields. This is relevant when studying Russian for general academic purposes. In addition to this, sorting by source and by field of science has proven to be a useful sorting method (Kozlova, 2013).

**Conclusions**

This paper has clearly shown that the available scientific works devoted to the development of corpora enable us to synthesize a theoretical basis for the development of a new corpus of scientific and technical texts. The standard tools and methods proved to be inapplicable for designing the corpus of scientific and technical texts. The specific characteristics of scientific and technical texts influencing the corpus being developed were analyzed, and possible solutions described. The compositional structure of scientific and technical texts is supposed to be solved by adding additional structural markup that would show the place of each text element in the document. The terminological word combinations are another problem that can be solved by adding a super-corpus database suitable for all subject areas. The dictionary entry presented reflects all paradigmatic features, thus providing sufficient information for Russian terminology processing and further standardization. The BMSTU corpus of scientific and technical texts is supposed to be a unique resource as it can be used for a large variety of human and technological tasks. At the same time, it reflects up-to-date circumstances on text resources, amount of data processed and quick update. Creating a marked-up scientific and technical texts collection is a multilateral issue. However, by resolving issues individually, it
is possible to find answers in other not similar natural language processing tasks. The BMSTU corpus of scientific and technical texts would become a valuable practical tool for teachers and students under the condition that the reviewed achievements of the existing corpora are implemented there.

References


