Linguistic corpus as a means of adaptation of modern scientific agricultural approaches

Maxim Melnikov*, *Elvira* Mallyamova, and *Marina* Morozova

Department of Foreign Languages, Ulyanovsk State Agrarian University, 432017 Ulyanovsk, Russia

Abstract. The article touches upon the issue of the relevance of creating a specialized agrarian linguistic corpus containing a sufficient amount of scientific publications. The authors substantiate the importance of facilitating prompt access to research results through the global Internet system in agricultural field to increase effectiveness and sustainability of the industry. The study described in the article is aimed to determine the effectiveness of using agrarian linguistic corpus in the process of analysis and adaptation of scientific information in English for academic staff. The problem solution of facilitating the process of obtaining information on scientific discoveries in the agricultural sector is presented in the article. The research team of the Department of Foreign Languages of Ulyanovsk State Agrarian University introduced findings of the study.

1 Introduction

In the modern context of the dynamic development of the agricultural industry on a global scale, the issue of effective and timely adaptation of foreign experience is becoming particularly relevant. Due to objective conditions, agricultural science requires international cooperation to increase its effectiveness and environmental friendliness. International research projects involve knowledge of the English language at the level of adequate communication in the professional field.

One of the major problems of adaptation of modern technologies in the agricultural sector is the access restriction for many Russian researchers to relevant technological solutions.

This restriction is caused by the lack of a sufficient level of language competence of agro-scientists. Access to research results through the global Internet system is beneficial to those who speak English, since most authoritative and authentic sources are mainly in English.

Thus, the problem of facilitating the process of obtaining information on scientific discoveries and technological solutions in the agricultural sector has become so urgent that it needs to be solved in the near future.

Unfortunately, the problem has not been solved so far. One of those solutions could be a specialized agrarian linguistic corpus containing a sufficient amount of scientific publications over the past 5 years.

2 Materials and methods

The basic method for solving the facilitation problem of the adaptation of relevant scientific information, which considers modern approaches and innovations in agricultural science, is to analyze the development of this problem in the specialized literature.

In the modern context of information technologies development, a unique opportunity has arisen for access to authentic information sources in various fields. Currently, the widespread use of corpus linguistics (corpus approach) in many professional fields related to scientific activities at the international level has proven to be efficient [1, 2].

The use of corpus linguistics has been intensified over the past four decades, especially since the mid-1980s. One of the strengths of the linguistic corpus is the empirical nature of data, due to which the hypotheses of researchers are verified by using a huge amount of data, which makes linguistic analysis more objective [3].

The possibility of obtaining the most objective data determines the widespread use of corpora in almost all industries, including, for example, economics, management, translation studies, diachronic studies and discourse analysis.

Corpus linguistics has revolutionized publishing (at least in English), due to which dictionaries and grammar guides published since the 1990s are compiled on the basis of data from linguistic corpora («even people who have never heard of a corpus are using the product of corpus-based investigation» [4].

“Recent critical views on the usefulness of a general academic vocabulary have heightened the relevance of developing discipline specific academic wordlists to

* Corresponding author: mensch777@mail.ru

meet the needs of non-native English writers who must read and publish articles in English" [5].

The most important advantage of linguistic corpus is its digital format, which allows lexicographers extracting all authentic, typical examples of using lexical units from a large volume of text in a few seconds. The second advantage of using the linguistic corpus is the ability to obtain information about the frequency of lexical units and collocations, their quantitative assessment [6]. "The lexical analysis revealed high lexical variation in the corpus and narrow word range.

Academic words provided a lower coverage (6 %) than that usually reported for research articles (10–12 %), and a higher coverage than that reported for newspapers (4 %). The analysis of high-frequency words showed that many of these words, including general and academic words, were closely associated with the discipline of agriculture, and therefore represented the technical vocabulary of the texts" [7].

Frequency determines the basic terminological minimum in the field of highly specialized issues. The possession of the most frequent terminological units determines the appropriate translation of special texts. Another advantage of using linguistic corpus is related to corpus markup and annotations. Many corpora (e.g., British National Corpus, BNC) are encoded with textual markings (e.g. genre, domain etc.).

Corpus annotations, such as marking up part of speech and semantic markup, which allow eliminating the ambiguity of words, also makes it possible to group ambiguous words or homographs more appropriately. Moreover, the constant updating of the corpus allows tracking minor changes in the meaning and use of lexical units, and thus keep the corpus up to date.

Based on the foregoing, it was decide to create the agrarian linguistic corpus for academic staff to implement the scientific research especially in the field of the agricultural industry and international cooperation related to modern approaches and innovations of agricultural science and the adaptation of foreign experience in this field.

The analysis of open sources has led to the conclusion that there are no specialized linguistic corpora that contain academic information in agrarian field available to Russian researchers at the moment. Thus, the research goals and objectives were formulated.

3 Algorithm and design

The algorithm was implemented on the websites listed above. The content of linguistic corpus was filled with articles from various journals included in the list of Scopus, WoS, on topics relevant to the research tasks of the university academic staff.

Particular attention was paid to articles published in 2019: Farmers' Knowledge, Attitude and Practices on Pesticide Safety: A Case Study of Vegetable Farmers in Mount-Bamboutos Agricultural Area (Cameroon Sonchieu Jean1, Fointama Emmanuel, Akono Nantia Edouard, Serri Brownlinda), Effect of irrigation and adjuvant on residual activity of pendimethalin and metazachlor in kohlrabi and soil (Miroslav Jursík, Martin Kočárek, Marie Suchanová, Michaela Kolářová, Jaroslav Šuk), Determination of the physical and mechanical properties of a potato (the Agria variety) in order to mechanise the harvesting and post-harvesting operations (Negar Ahangarnezhad, Gholamhassan Najafi, Ahmad Jahanbakhshi), Effect of land use on soil chemical properties after 190 years of forest to agricultural land conversion (Kateřina Zajícová, Tomáš Chuman) and so on [8–11].

Over 525 342 words were collected in all 120 articles published on different websites (including online journals, magazines related to agrarian science) from 2014 to 2019. We have chosen the above sources as a modern presentation of general and highly specialized vocabulary, as they are the most reliable and relevant.

All the articles in linguistic corpus were scanned online using the Scrappy infrastructure (http://scrapy.org/), and all the data was uploaded to the figshare open source archive, which includes the original site files and information about the site, creating the database and technology. Scanning includes editorial content only.

Advertisements, reader comments, phone numbers, website addresses, and email addresses were ignored. In addition, punctuation was ignored, words separated by hyphens, and apostrophes were deleted (example: 's).

In the course of this research, the agrarian linguistic corpus was created on the basis of materials from selected sources, the glossary of the most common terms and terminological phrases was compiled and the diachronic analysis of concepts used in the context of modern approaches and innovations in agricultural science were carried out.

The obtained data make it possible to carry out appropriate and equivalent translation of scientific articles describing modern approaches of agricultural science. The data also help to follow the frequency of relevant terms and track new ones.

The dynamics of the use of terms reflects the relevance of various approaches in agricultural science at the moment. In addition, the database of selected texts allows quick access to the necessary narrowly specialized information, remaining relevant due to the possibility of its constant replenishment.

4 Results

The department of Foreign Languages of Ulyanovsk State Agrarian University carried out the research work to create an agrarian linguistic corpus facilitating the process of adaptation of relevant scientific information, which analyzes modern approaches and innovations in agricultural science.

The purpose of the study was to identify the effectiveness of using the agrarian linguistic corpus in the process of analysis and adaptation of academic articles in English for academic staff.

In the framework of creating a specialized linguistic corpus, the following basic tasks on which the researchers mostly focused were solved: 1. optimization

of the search for scientific material available on the Internet according to certain criteria determined by the specialists; 2. representation of language units implemented in specific scientific texts; 3. removal of homonymy.

To achieve the goal of the research work, the following objectives were solved:

- search for reliable authentic sources of agricultural scientific literature;
- analysis of scientific data presented by experts in the public domain;
- creation of the agrarian linguistic corpus based on materials from selected sources;
- compiling a glossary of the most common terms and terminological phrases;
- diachronic analysis of concepts used in the context of modern approaches and innovations of agricultural science;
- survey on the use of the agrarian linguistic corpus.

The representatives of the Department of Foreign languages of Ulyanovsk State Agrarian University Associate Professor Mallyamova E.N., Associate Professor Melnikov M.V. and with the participation of Associate Professor of the department of Foreign Languages of Ulyanovsk Institute of Civil Aviation Morozova M.A. analyzed authentic open sources of scientific information in the field of agricultural science.

As a result of the analysis, the most relevant resources were selected such as: Acres USA, Mother Earth News, Capital Press, AgriSeek, Practical Farm Ideas (UK), Progressive Farmer, US Farm Network, AgProfessional, AgWeb, AGCanada, Agricultural publications in Canada, Western Producer (Canada), Modern Agriculture (Canada), Agriculture Today (India), Farmer's Weekly (South Africa), Africa Agribusiness (South Africa), Farm Industry News, Delta Farm Press, Corn and Soybean Digest (Minneapolis, MN), Farmers Guardian, Lancaster Farming, Modern Farmer, Rural Cooperatives.

5 Discussion

In scientific texts, terminology plays an important role, which functions both in the commonly used layer and in the form of rather narrow special designations included in the language of scientific literature. Within the framework of corpus linguistics, it becomes possible to re-describe such an important unit of modern vocabulary as a “term”. Since modern knowledge bases are polythematic, the role of a contextual dictionary is very important because they use the principles and methods of corpus linguistics.

The terms within the framework of the same scientific field can be divided into two classes: 1) broad-based terms which belong to several subject areas and at the same time are included in the common language vocabulary; 2) highly specialized terms belonging to strictly defined subject areas, which some researchers designate as professional jargon.

The specificity of any term, first of all, is that its semantic structure contains broad-based meaning, which has clear semantic boundaries. In this regard, the identification of the meaning of the term through the context is becoming increasingly important.

In the traditional view, academic interaction, both written and oral, is a continuum, including scientific articles written for both experts and non-fiction articles intended for a wide range of readers. One of the main difficulties in adapting scientific texts is the use of special jargon, which is technical terminology or a set of specific idioms used in particular activity or by group of specialists.

Scientific and popular science texts substantiate conclusions and emphasize the uniqueness and novelty of research results.

Academic writing is used in scientific journals, conference proceedings, and scientific books. This style assumes that readers to have prior knowledge of the field, familiarity with the standard structure of a scientific article such as IMRAD (Introduction-Methods-Results-Discussion), and the use of academic vocabulary (e.g., analyze, facilitate), jargon.

General vocabulary is often classified by frequency of lexical units, which is determined by evaluating written and spoken language, often from magazines and online publications.

Websites and online subject matter journals are likely to be the most reliable source of academic information in the framework of the research. The general vocabulary is traditionally divided into high frequency (1000–3000 phrases) and low frequency (above the level of 9000 phrases).

More recently, a med-frequency group (family level of 3.000–9.000 words), created from a common vocabulary, was also presented in academic publications [12, 13].

Many studies aimed to create industry vocabulary lists, for example, agriculture [14], chemistry [15], etc. These lists are intended to facilitate scientific search for researchers with low level of language competence.

In general, studies using vocabulary lists have shown that technical vocabulary (e.g. polymerization) usually accounts for 5 % of academic texts, while the bulk (80 %) consists of high-frequency polythematic words (e.g. soil, animal) and a smaller part (8–10 %) consists of academic vocabulary (for example, derivative, technique) [12].

Hu and Nation found out that a familiarity with 98 % of all vocabulary (circa the first 2,000 word families) in a text is required to precisely comprehend the content [16]. In another study on vocabulary and comprehension in non-native grownup readers, Laufer and Ravenhorst-Kalovski presumed a minimum level of comprehension, requiring knowledge of 95 % of words in a scientific article [17].

We consider the importance of the context as the most critical point in the in the process of analysis and adaptation of scientific information in English.

The industry vocabulary list containing all possible general scientific and specialized terms is not enough for adequate comprehension of scientific texts by non-native specialists. It is necessary to provide academic staff with the access to the linguistic corpus for researchers to

compare analyze and choose the most appropriate meaning in a certain context.

6 Conclusion

The compiled glossary of the most common general scientific and highly specialized lexical units allows carrying out effectively a thematic search for targeted articles throughout the corpus. In addition, diachronic analysis makes it possible to identify current trends in scientific research and technological solutions in the agricultural sector.

The conducted survey on the use of the agrarian linguistic corpus showed it to be efficient in research activity.

Thus, the goal of the study to identify the effectiveness of using the agrarian linguistic corpus in the process of analysis and adaptation of academic articles in English for academic staff was achieved by solving the tasks framed in the study.

As a recommendation for agrarian university academic staff which carries out scientific activities and work according to the training programs for highly qualified specialists, it is recommended to create specialized linguistic corpora according to research goals and tasks.

In order to get general academic information, researchers can use the following linguistic corpora as additional sources on general scientific topics: British National Corpus http://www.natcorp.ox.ac.uk/, American National Corpus (ANC). http:// americannationalcorpus.org/, Corps of the German Language Institute LIMAS-Korpus http://www. korpora.org/Limas/ etc.

References

1. *Lancaster University Department of Linguistics and English Language*, Retrieved from: http://www.research.lancs.ac.uk/portal/en/upmprojects/construction-and-corpusbased-analysis-of-the-british-councillancaster-university-aptis-corpus (3ace9bdf-2dd1-4f72-b9d0-530670fcc2f7).html
2. *The Centre for Corpus Approaches to Social Science is an ESRC-funded research centre*, Retrieved from: https://www.lancaster.ac.uk/users/ moocs/corpus/people/hardie-wk4/index.htm
3. T. McEnery, A. Wilson, *Corpus linguistics*, 2rd ed. (Edinburgh University Press Edinburgh, 2001)
4. S. Hunston, *Corpora in applied linguistics* (Cambridge University Press, Cambridge, 2002)
5. I. Martínez, S. Beck, C. Panza, *Academic vocabulary in agriculture research articles: A corpus-based study*, Engl. for Spec. Purp., **28(3)**, 183–198 (2009)
6. A. Karin, *Corpora and Language Teaching* (John Benjamins, Amsterdam, Philadelphia, 2009)
7. V. Muñoz, *The vocabulary of agriculture semi-popularization articles in English: A corpus-based study*, Engl. for Spec. Purp., **39** (2015)
8. Agricultural Sciences ISSN Print: 2156-8553 ISSN Retrieved from: https://www.scirp.org/journal/as
9. *Plant, Soil and Environment* (PSE) ISSN 1214-1178 (Print) ISSN 1805-9368 (On-line).
10. *Research in Agricultural Engineering* (RAE) ISSN 1212-9151 (Print) ISSN 1805-9376 (On-line).
11. *Soil and Water Research* (SWR) ISSN 1801-5395 (Print) ISSN 1805-9384 (On-line).
12. I.S. Nation, *Learning Vocabulary in Another Language* (Cambridge University Press, New York, 2001)
13. I. Nation, *How Large a Vocabulary is Needed For Reading and Listening?* Can. Mod. Lang. Rev. **63(1)**, 59–82 (University of Toronto Press, Sep. 2006)
14. I.A. Martínez, S.C. Beck, C.B. Panza, *Academic vocabulary in agriculture research articles: A corpus-based study*, Engl. Spec. Purp., **28(3)**, 183–98 (2009)
15. L. Valipouri, H. Nassaji, *A corpus-based study of academic vocabulary in chemistry research articles*, J. Engl. Acad. Purp., **12(4)**, 248–63 (2013)
16. M. Hu, I.S.P. Nation, *Vocabulary density and reading comprehension*, Read a Foreign Lang., 403 (2000)
17. B. Laufer, G.C. Ravenhorst-Kalovski, *Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension*, Read a Foreign Lang., **22(1)** (2010)