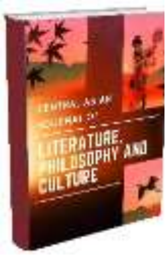

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY AND CULTURE

eISSN: 2660-6828 | Volume: 03 Issue: 03 March 2022
https://cajlpc.centralasianstudies.org

DIACHRONIC CORPUS CREATION PRACTICES IN WORLD SCIENCE: ON THE EXAMPLES OF KOREAN, CHINESE, RUSSIAN AND ARABIC LANGUAGES

Ataboyev Nozimjon Bobojon ugli
Doctor of philosophy (PhD) in philology, associate professor
Dean of the Faculty of Foreign Languages of BukhSU
E-mail: n.b.ataboyev@buxdu.uz
anb929292@gmail.com
https://orcid.org/0000-0002-9756-6849

Received 27thJan2022, Accepted 27thFeb2022, Online 28thMarch2022

Abstract: Corpus linguistics is a special branch of linguistics that studies language through the practice of creating linguistic corpora. The development of the field is closely related to the widespread use of language corpora in finding solutions to linguistic problems. In the article, the description of the progress of this branch of science to create a diachronic corpus in world languages such as Korean, Chinese, Russian and Arabic are clearly demonstrated.

Key words: corpus linguistics, linguistic corpus, language development, diachronic corpus, chronology

INTRODUCTION.

Corpus linguistics has gone through several periods until its current development. It was formed as an independent branch of science at the end of the 70s of the last century, and the research methods, methods and analytical tools of linguistics, which are considered its basis, have been known since the 12th century[1]. In the development of corpora, the emergence of the principles of text size and their sorting spans several periods.

When it comes to corpus linguistics, the term corpus, which is directly studied as an object of the field and a modern approach to working with texts, requires a special explanation.

The dictionary meaning of the word *"corpus"* was initially used in a narrow circle, that is, the composition of the works of a certain writer in the religious and literary genre was arranged in alphabetical order, i.e., in concordance-concordance lines, it was called a corpus. The use of corpus in this sense covers the period before the emergence of electronic corpus. Corpora were formed during this period for religious, literary, and lexicographical research, and this process required long and labor-intensive manual labor. At the same time,

[1] Corpus linguistics // available from: https://en.wikipedia.org/wiki/Corpus_linguistics [Accessed on 03.08.2023]

the speed of finding the searched lexical units increased and the number of users is expanding after the transition of corpora to the online system.

MAIN PART.

The practice of creating corpora focused on linguistic research has been the focus of attention of all scientific schools of the world. For example, in their research, Korean scientists developed corpus analysis methods for classifying types of emotions in the Korean language and developed the Korean Emotion Analysis Corpus[2]. Based on the structure of the created corpus, scientists have not only gained a deeper understanding of the differences between classes of emotions in the classification of emotions, but have also created a new standard data set that allows the evaluation of emotion analysis approaches. This, in turn, shows that corpus bases can enable the researcher to observe and classify based on the reasonable accumulation of data.

The great contribution of Russian linguists to the development of corpus linguistics can be demonstrated by the example of the national corpus of the Russian language (*Национальный корпус русского языка)*[3]. The National Corpus of the Russian Language is a linguistic corpus representing the Russian language that has been partially accessible through an online search interface since April 29, 2004. The Institute of the Russian Language of the Russian Academy of Sciences is constantly researching it. The corpus now contains more than 1 billion word forms, which are automatically lemmatized and POS-/gramme-tags, i.e. covering all possible morphological analyzes for each orthographic form. Lemmas, POS, grammatical elements and their combinations can be searched. In addition, 6 million word forms are contained in subcorpora with manual homonymy. A sub corpus with additional metadata related to morphological homonymy highlighted above is also automatically tagged. The entire corpus contains lexical semantics (LS) searchable tags, which include morphosemantic POS subclasses (noun, reflexive pronoun, etc.), corresponding LS features (subject class, causative, evaluation), derivation (diminutive, adverbial adverbs, adjectives, and etc.).

Arabic linguists are also conducting extensive research on this issue. For example, a new language processing tool, Arabic Corpus Processing Tools ACPTs 4.6 Version, has been developed, which is a stand-alone open/free resource for analyzing large volumes of Arabic and English texts, with a database of more than 50 million words, compatible with more than 8 gigabytes of PCs. keeps The new ACPT has sophisticated state-of-the-art features applicable to the corpus linguistics and corpus linguistic analysis literature, especially statistical packages. When compared with other tools suitable for corpus linguistic analysis, this leads ACPTs to be noted as the most efficient tool for corpus analysis. Focusing on language teaching issues, the most important tasks of appropriate tools that can be used to improve the language teaching/learning process are included in this corpus[4]. As a continuation of such works, it is possible to mention the project of creating the International Corpus of Arabic (ICA), which was created for the first time. It is intended to include 100 million analyzed tokens with an interface that allows users to interact with the corpus data in several ways. ICA is a

[2] Jung Y. and others A corpus-based approach to classifying emotions using Korean linguistic features // Cluster Comput (2017) 20: - pp. 583-595 DOI 10.1007/s10586-017-0777-8

[3] *Национальный корпус русского языка* https://ruscorpora.ru/

[4] Almujaiwel S. and Al-thubaity A. Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching // The Globalization of Second Language Acquisition and Teacher EducationAt: FukuokaVolume: 2. – 2016. 4 p. available at https://www.researchgate.net/publication/309351881_Arabic_Corpus_Processing_Tools_for_Corpus_Linguistics_and_Language_Teaching

representative corpus of Arabic, created in 2006, which is intended to cover the Modern Standard Arabic language used throughout the Arab world. ICA was analyzed by the Bibliotheca Alexandrina Morphological Analysis Enhancer (BAM-AE). BAMAE is based on the Buckwalter Arabic Morphological Analyzer (BAMA)[5].

In our opinion, it is appropriate to focus on the elements that are the basis for the creation of the above-mentioned international corpus of the Arabic language. The website Alexandrina Bibliotheca is the core of the international Arabic corpus. Bibliotheca Alexandrina (BA) is one of the leading international institutions in Egypt, with a place in the dissemination of culture and knowledge, as well as in supporting scientific research. He launched the ambitious project to create the International Arabic Corpus (ICA) as a major effort to create a representative corpus of Arabic spoken throughout the Arab world, as this is the most common way to support research on the language. . After the corpus was created, the analyzed form was the first analyzed Arabic corpus available as a linguistic resource for researchers. It is also the first systematic analysis of cross-national studies in an Arabic-speaking community, which will be a very useful resource for linguists who believe that their theories and descriptions of language should be based on real data rather than on unsubstantiated facts[6].

Chinese linguistics is also one of the leading fields in corpus linguistics research. Because, relying on the experience gained in creating a computer-based translator, the teaching system of the Department of Translation of Guangdong University of Foreign Languages, they summarize the principles and procedures for the development of educational translation and parallel corpus. Further research was conducted by university scientists on the integration of the translation corpus into an online autonomous learning platform for the training of translators. Starting from 2003, by selecting some texts from PACCEL-S, the process of quantitative assessment of pronunciation errors, grammatical errors and pauses was studied using the ParaConc program. According to them, the most frequent mistake in translation work of Chinese students is related to pronunciation, mainly vowels. Among 732 sentences interpreted by 183 examinees, there was an average of one pronunciation error per sentence, and each examinee made 3.64 such errors. The resulting corpus, ParaConc, shows 90 grammatical errors, the most frequent of which are the incorrect use of singular and plural forms and the inappropriate use of speech units[7]. In this way, it can be said on the example of the research of Chinese linguists that it is possible to identify and eliminate errors in pronunciation based on corpus analysis.

It also includes a corpus of spoken Chinese, a collection of transcripts of spoken Chinese produced by non-native speakers and native speakers intended to be publicly available to researchers[8]. This corpus is the first case of creating a hitherto undeveloped colloquial corpus of its target Mandarin Chinese language, and serves as a great resource for the study of many English loanwords found in Chinese.

[5] Nagi M. The International Corpus of Arabic: Compilation, Analysis and Evaluation. // Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). – January 2014. – 8-17 pp. available at https://www.researchgate.net/publication/301404178_The_International_Corpus_of_Arabic_Compilation_Analysis_and_Evaluation

[6] Alexandrina Bibliotheca available at https://www.bibalex.org/ica/en/About.aspx [retrived on 23 september, 2023]

[7] Hu K. and Kim K. H. (eds.), Corpus-based Translation and Interpreting. Studies in Chinese Contexts, Palgrave Studies in Translating and Interpreting, https://doi.org/10.1007/978-3-030-21440-1_3 // written by B. Wang (*) – Guangdong University of Foreign Studies, Guangzhou, China. 2009. – 61-87 pp.

[8] Wang J. Recent Progress in Corpus Linguistics in China // International Journal of Corpus Linguistics // 6(2). DOI: 10.1075/ijcl.6.2.05wan – China, July 2002 – 281-304 pp.

Moreover, Y. Liu, M. Xiaohui Qin, L. Wang and Ch. Chinese scholars such as Huang created CCAE: A Corpus of Chinese-based Asian Englishes[9]. This, in its place, can open a wide way for scientists to study the scope of cross-linguistic interference and the influence of the dominant language on other languages.

In general, the process of interference of languages in the above mentioned five-stage sequence is accelerating today. Because the factors of the current process of communication are fundamentally different from the previous person-person and society-society categories and are being built on the basis of new, modern forms of information exchange. In our opinion, the acceleration of communication in this media field is reducing the scope of the purity and diversity of cultures and languages. Therefore, as long as the study of language development, which is the object of research, is not studied statistically based on the prism of corpora, the practice of preserving the current state of languages and leaving them as a basis for future research will not emerge. And this cannot be improved with any excuse.

According to Swedish scientist M. Kytö, electronic historical corpus and corpus methodology are research methods aimed at studying and evaluating the current state of languages and allowing to consider linguistic changes that may occur in the future. With this, the scientist emphasizes that within the wide range of corpus linguistic methodology, historical corpus linguistics has emerged as a vibrant field that has significantly increased the attractiveness for the study of language history and change. Indeed, the research process of evidence-based historical linguistics would not have been completed without the methodology and new impetus of corpus linguistics. In an era of rapidly changing life and research, increasing competition for academic careers and opportunities for young scientists, there is no doubt that the methodologically easy field has a future. Historical corpora and other electronic resources have also made the study of language history attractive: working on them engages students in an individual and interactive way[10].

CONCLUSION

In our opinion, it is important to create corpora of this type in the case of the English language as well. After all, if the number of manuscripts and published literature decreases at the same time when materials are being digitized, and if they are summarized in corpora in electronic form, any linguist conducting any linguistic research can come to a certain conclusion about the present day based on the history of the language.

REFERENCES

1. Alexandrina Bibliotheca available at https://www.bibalex.org/ica/en/About.aspx [retrived on 23 september, 2023]
2. Almujaiwel S. and Al-thubaity A. Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching // The Globalization of Second Language Acquisition and Teacher EducationAt: FukuokaVolume: 2. – 2016. 4 p. available at https://www.researchgate.net/publication/309351881_Arabic_Corpus_Processing_Tools_for_Corpus_Linguistics_and_Language_Teaching
3. Hu K. and Kim K. H. (eds.), Corpus-based Translation and Interpreting. Studies in Chinese Contexts, Palgrave Studies in Translating and Interpreting, https://doi.org/10.1007/978-3-030-21440-1_3 //

[9] Liu Y. and others CCAE: A Corpus of Chinese-based Asian Englishes available at https://arxiv.org/abs/2310.05381v1 [address date: 30.10.2023]

[10] Kytö M. Corpora and historical linguistics // RBLA, Belo Horizonte, v. 11, n. 2. – Uppsala University, Uppsala: Sweden, 2011 – P. 417.

written by B. Wang (*) – Guangdong University of Foreign Studies, Guangzhou, China. 2009. – 61-87 pp.

4. Jung Y. and others A corpus-based approach to classifying emotions using Korean linguistic features // Cluster Comput (2017) 20: - pp. 583-595 DOI 10.1007/s10586-017-0777-8
5. Kytö M. Corpora and historical linguistics // RBLA, Belo Horizonte, v. 11, n. 2. – Uppsala University, Uppsala: Sweden, 2011 – 417-457 pp.
6. Liu Y. and others CCAE: A Corpus of Chinese-based Asian Englishes available at https://arxiv.org/abs/2310.05381v1 [retrived on 30.10.2023]
7. Nagi M. The International Corpus of Arabic: Compilation, Analysis and Evaluation. // Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). – January 2014. – 8-17 pp. available at https://www.researchgate.net/publication/301404178_The_International_Corpus_of_Arabic_Compilation_Analysis_and_Evaluation
8. Wang J. Recent Progress in Corpus Linguistics in China // International Journal of Corpus Linguistics // 6(2). DOI: 10.1075/ijcl.6.2.05wan – China, July 2002 – 281-304 pp.
9. *Национальный корпус русского языка* https://ruscorpora.ru/ [retrived on October 10, 2023]