MODERN ASPECTS OF ECONOMIC AND SOCIAL SUSTAINABLE DEVELOPMENT
Vol. 12 (2023): Special Issue

INTERNATIONAL E-CONFERENCE-15th September

DATA COLLECTION AND NORMATIVE PRINCIPLES IN CREATING A DIACHRONIC CORPUS

Nozimjon Bobojon ugli Ataboev
Doctor of philosophy (PhD) in philology, associate professor
Dean of the Faculty of Foreign Languages of BukhSU
E-mail: n.b.ataboyev@buxdu.uz

Annotation. The corpus is a set of selected, sufficient linguistic texts (oral or written) that can be formed and classified on the basis of strict principles based on the pragmatic purpose of the corpus compiler, meet the criteria of representativeness, combine principles such as processing, tagging, and perfect computer automation, which is a database with a convenient quantity and consistency for systematic search, reference results, and empirical analysis.
Key words. linguistic data, BNC, ICE, IBM, national corpora.

. INTRODUCTION

According to Tony McEnery and Andrew Wilson: “The concept of corpus was originally used to refer to a collection of texts that contained more than one, and meant nothing more than that”. But in today's developing linguistics, "corpus" is used in a broader sense: one that includes all types of texts and provides examples of them; having a definite size; designed to be readable by computer technology; a collection of texts that provides the opportunity to find information in a standard way for everyone.

MAIN PART

According to the type of linguistic data, corpora are divided into written, spoken and mixed corpora. Most of the first generation corpora were exclusively written. Written texts are much easier to collect. There are three methods of entering written texts into computer:

- re-typing texts;
- using the texts that already exist in electronic form;
- scanning typed texts (but with many mistakes to correct).

Large modern corpora are usually combined, with a preponderance of written texts. Even in the British National Corpus (BNC) only 10% of texts are spoken. International Corpus of English (ICE) stands out with 60% of spoken material.

Meanwhile, a language mainly exists in its oral form; its written form is secondary. That is why spoken or mixed corpora are so important. Most famous specific spoken corpora include the London Lund Corpus (LLC, 1975) and the Lancaster/IBM Spoken English Corpus (1992). The latter contains 52600 words and comes on a CD-ROM with the audio recordings, fully labeled for accents, intonation, pauses, etc. However, it does not contain information about the social status and education of the respondents, which limits its use in sociolinguistics.

ISSN 2319-2836 (online), *With support* APJMMR
https://www.gejournal.net/index.php/APJMMR

Spoken corpora include fewer word uses than written corpora, not only because of the complexity of data collection, but also because fewer words are usually sufficient for prosodic research. So, to study intonation, a corpus of one hundred thousand words is enough.

Usually quite detailed information about the respondents is collected:

- place of recording
- what the respondent is doing
- time
- date
- number of participants
- degree of spontaneity of the conversation
- theme
- sex of participants
- age of participants
- ethnicity of participants
- native language of participants
- profession
- education
- social status
- attitude towards the note-taker
- dialect

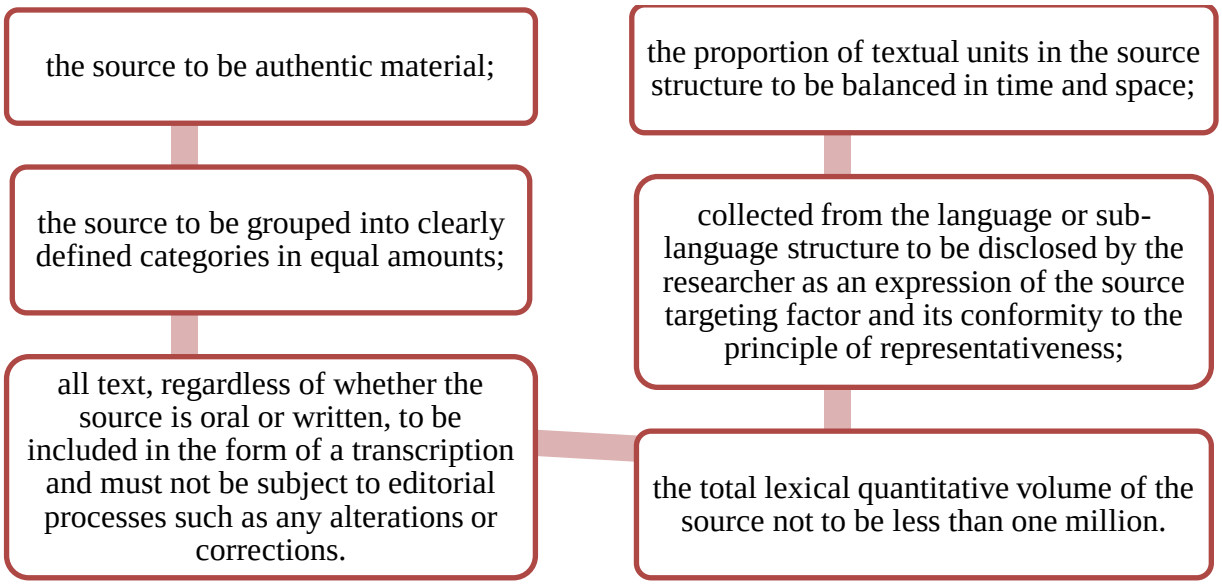
It would be appropriate to consider data-collection methods as following: These methods focus on working with the collection of texts and the sorting them in accordance with the genre, period of time and others. At the same time, requirements have been developed for the text sources that are part of the corpus, and it is recommended that each corpus compiler collects texts based on these six requirements.

Given the fact that insufficient research has been conducted on the criteria for the collection and selection of corpus texts, it is relevant to propose criteria requirements for the textual linguistic information included in the corpus. Accordingly, normative requirements were developed for the textual content of the linguistic base within the scope of the corpus, and they constituted six (see diagram 1).

Diagram 1

Normative requirements for texts of a diachronic corpus

ISSN 2319-2836 (online), *With support* APJMMR
https://www.gejournal.net/index.php/APJMMR



The third stage methods consist

CONCLUSION

In my opinion, the biggest and most important positive opportunity offered by corpus linguistics is that it collects and conveys to the next generation the texts that reflect the features of representativeness selected from the point of view of time, space and field of natural language. The creation of national corpora, which can reflect such features as the principles of existence and development of the language - its status as a state language (space), its users, i.e. the presence of language owners (relatively, time and field), makes it possible to compare the present and future of the language. Various foreign concepts introduced as a result of globalization have a negative impact on the language content. If a national corpus of a language is created, the language content of the time period covered is clearly preserved. An example of this is the British National Corpus (BNC). The BNC, which contains a linguistic database of 100 million words, represents the language between 1993 and 1997, and this database is kept unchanged even after several centuries.

REFERENCES

1. Gries S. Introduction to S. Gries and A. Stefanowitsch (eds.).Corpora In Cognitive Linguistics: Corpus–Based Approaches To Syntax And Lexis, – New York: Mouton de Gruyter. 2006. – 1-18 pp.

2. Мамонтова В.В. Особенности перевода сложносоставныгх слов с английского языка на русский (на материале корпуса публицистических текстов)дис. канд. фил.наук. Ставрополь: 2008. – С 54.

3. Bobojon o'g'li, N. A. Corpus-Based Research On The Language Features Of Corpus Linguistics: In The Example Of ECOCL. *Language*, *3*(2139), 950.

4. Ataboev, N. B. (2019). Problematic issues of corpus analysis and its shortcomings. ISJ Theoretical & Applied Science, 10(78), 170-173.

5. Ataboev, N. B. (2019). ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis. *Test engineering and management*, *81*, 4170-4176.

ISSN 2319-2836 (online), *With support* APJMMR
https://www.gejournal.net/index.php/APJMMR