Pindus Journal Of Culture, Literature, and ELT
ISSN: 2792 – 1883 **Vol 2 No. 3**
https://literature.academicjournal.io

The Theoritical Procedure Of Compiling A Diachronic Corpus

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

Abstract: The article deals with the process of compiling a diachronic corpus and corpus analyses methods. Corpus creation details set out of six closely connected procedural steps have been taken into account in the following paper.
Key words: corpus linguistics, a diachronic corpus, corpus methodology, empirical system

INTRODUCTION.

The independent development of CL was based on computer linguistics, mathematical linguistics and other fields of linguistics. CL was considered as a research method in the early stages of its development and was considered as part or a tool of computer and mathematical sciences; today it has become a methodology with its own goals, objectives and research methods and is widely accepted by the scientific community. In the future, however, it would be appropriate to assess as a high-level scientific hypothesis that the rapidly evolving CL will have all the features of a linguistic research paradigm and/or an independent science that can solve problems in all fields.

The corpus is a set of sufficiently large selection of linguistic texts (oral and written) that can be formed and classified on the basis of strict principles based on the pragmatic purpose of the corpus designer, meet the requirements of representativeness, being a database with sufficient quantity and consistency convenient for reference result and empirical analysis.

The classification of corpora is also a characteristic problem facing Corpus linguistics. Approaches to classification vary, and it can be seen that they are mainly carried out according to the size, design, language, and similar characteristics of the corpus. According to the principle of classifying corpora in terms of the pragmatic purpose of the user, linguistic corpus using existing corpus terms can be applied in one of the areas: a) translation practice and comparative linguistics: b) lexicography: c) literary theory: d) linguodidactics and language teaching: e) etymology and linguistic history research and etc. Indeed, it would be inappropriate to look at the classification of corpora as the final standard form, because when any corpus is created, it brings with it a new purpose, a new design, and a new type of classification.

MAIN PART.

We consider it necessary to divide the corpus linguistics methodology into an empirical system of six methods, aimed at creating corpora based on the principles of the corpus and their application in applied research:

It is reasonable to say that the first type of methods are ***the aim-oriented methods*** before the creation of the corpus. Of course, based on the *aim-orientation* of the corpus, the corpus compiler

ISSN 2792-1883 (online), Published in Vol: 2 No: 3 for the month of March-2022

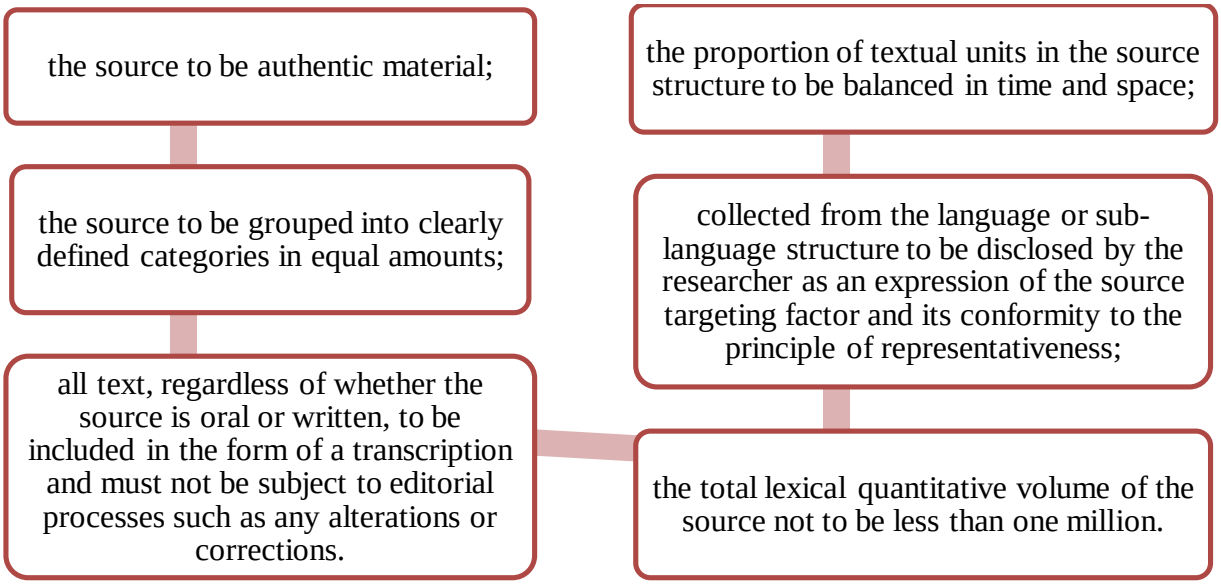
`must clearly define what kind of research the linguistic corpus being created is intended for.

It would be appropriate to call the second stage methods ***data-collection*** methods. 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 corpus



The third stage methods consist of ***metadata addition methods***. The methods included in this group aim to add reference data to the selected texts for inclusion in the corpus. A clear representation of the database is a very simple method, and it allows the information to be searched by category (newspaper material, word of mouth, etc.) and linked as a reference in the example given in the concordance row. The tagging method is important at this stage, and today it is mostly done using computers.

Any syntactic device can also be tagged during the research process. For instance, grammar structure of *I am* can be tagged as PPSS+BEM (PPSS – non-third person nominative personal pronoun ва BEM – am form of Be).

In this case, if VB comes for *the verb*, the third person Z added to it is the present simple tense suffix, while D denotes the verb form in the past tense, G represents the gerund form, and N represents the past participle forms of the verb.

The tag set is formed as the final systematic sum of these individual tags. The labeling system for the Penn Treebank corpus lists a total of 48 tagging codes, including: CC – coordinating conjunction, CD – cardinal number, DT – determiner, FW – foreign word, JJ – adjective, JJS – adjective in superlative degree, JJR – adjective in comparative degree, NN – noun, NNS – plural

ISSN 2792-1883 (online), Published in Vol: 2 No: 3 for the month of March-2022

noun, NNP – proper noun, RB – adverb, RBR – adverb in comparative degree, RBS – adverb in superlative degree and others.

The fourth stage methods are ***automation methods*** aimed at converting the cases into a fully electronic form. The methods at this stage work with all the collected, sorted, linked metadata, and verified texts in the form of an electronic corpus and its design. It is also at this stage that work is carried out to incorporate a wide range of possibilities, such as the formation of concordances in the corpus, the color of the search word and the words it combines, and the ability to graphically express the results obtained on the basis of corpus analysis.

In the fifth stage, it is important to use ***evaluation methods***. The perfection of the created corpus can be determined by identifying whether they can respond to each of the stated principles of the concept of corpus. In this case, each feature of the case must be evaluated.

In the sixth, final stage, ***frequency search-analysis methods*** can be introduced in the compiled corpora. It is possible to include methods that allow to draw conclusions by looking at the frequency of use of words in the corpus. In general, the importance of frequencies is high in all corpus-based analyses, and working with them also requires a clear distinction, because the frequency of words use that are part of the corpus enables us to draw comprehensive conclusions about them. T.S. Gries[1] states the corpus frequencies can be of 3 types:

Raw frequencies – the frequency of use of the word in the corpus in the search, which is given in exact numbers. For example, in COCA, the verb *give* is used 202515 times, which is its raw frequency (RF).

Normalized frequencies – the ratio of the exact number of frequencies indicated by the corpus to the levels of 10 (1000, 1000000, etc.). For example, to find the NF of the verb *give*, we calculate the total amount of COCA (TA), i.e., 570 million. In this case, to find the normalized frequency (NF) of a word relative to a million, we need to divide its RF by 570. It is appropriate to provide the following formula:

$$\mathrm{NF(million/1)} = \frac{\mathrm{RF}}{\mathrm{TA}/1000000}.$$

It can be seen that the NF(million/1) of the verb *give* in COCA is approximately 355,2. This is important when comparing monolingual corpora or interlingual corpora.

Logged frequencies – the degree derived from the primary frequency embedded in a clearly selected grounded logarithm (e = 2.7182818). That is, the logarithm is the inverse function to exponentiation. If we look at the formula $\log_b x=y$, if b = 10, if x = 100, then y = 2. Simply put, the answer to the question of how many levels of 10 will be 100 is 2. In deriving the logarithmic frequency, the above formula b = 2,7182818 (base), x = 202515 (raw frequency), and from this the logarithm shows that the degree is y = 12,218569364782, which is the logged frequency (LF) of the use of the verb *give* in COCA.

It is advisable to compare the corpus data by means of the above mentioned frequencies. In our view, this comparison process can take three different forms*: Internal comparison in a corpus; Interlingual comparison in different corpora of one language; Intercollectual comparison in the corpora of different languages.*

Conclusion. We proposed to divide the Corpus linguistics methodology into six groups of

[1] Gries S.Th. Quantitative designs and statistical techniques/ English Corpus Linguistics edited by Douglas Biber and Randi Reppen. – United Kingdom: Cambridge, 2015. – P. 52.

ISSN 2792-1883 (online), Published in Vol: 2 No: 3 for the month of March-2022

methods, aimed at creating a corpus and applying it in research. It is through the integration of these systematic methods that the creation of an experimental corpus aimed at studying the language features of Corpus linguistics can be justified. Accordingly, the corpus planned to be created is formed in accordance with this process.

As we consider, Corpus linguistics is considered as a methodology of linguistic research, a group of six-step methods that include specific processes based on the consistency of its research methods – aim-oriented methods; data-collection methods; metadata addition methods; automation methods; evaluation methods and frequency search-analysis methods have been improved.

USED LITERATURE:

1. Большакова М.А.Лингвистический контент и программнаяреализация интеллектуального немецко -русского отраслевого словаряДиссер...канд.фил. наук. Москва. – 2013. С 56-57.

2. Джаъфарова Д.Ф. Модели лингвистического анализа текстов таджикского языка (на материале газелей Хафиза) дисс. ... кандидата филологических наук. Душанбе – 2013. С 35

3. Gries S.Th. Quantitative designs and statistical techniques/ English Corpus Linguistics edited by Douglas Biber and Randi Reppen. – United Kingdom: Cambridge, 2015. – P. 52.

4. 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.

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

6. 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 2792-1883 (online), Published in Vol: 2 No: 3 for the month of March-2022