Corpus-Based Research On The Language Features Of Corpus Linguistics: In The Example Of ECOCL

Nozimjon Ataboev Bobojon o'g'li

Abstract : The following article deals with the corpus based analyses on the materials regarding the field of corpus linguistics (CL). That is to say, the scientific works – dissertations, articles and manuals on corpus linguistics have been collected in a corpus. The data have been analyzed in order to get more detailed information on CL and the special terms used in it. Moreover, the collected corpus has been named as ECOCL (Experimental Corpus Of Corpus Linguistics) which is devoted to analyze the linguistic features of the scientific materials based on CL. In the article, considering all the discussions and results, the final conclusions have been reached.

Index Terms: Corpus, corpus linguistics, ECOCL, corpus-based analysis, type, token, database, collocational units(CU).

1 INTRODUCTION

Currently, new information technologies are used in almost all spheres of linguistics – both theoretical and applied. Actively improving computer programs and projects for working with text, search and recognition of information influenced on the emerge of the corpus linguistics. One of the types of such projects, allowing both a linguist and any user interested in a language, to access huge volumes of data in the shortest possible time, are language corpus.

2 PROCEDURE

Literature review

Considering CL as a methodology for linguistic researches has become well-known conception among the linguists currently. It has been observed in numerous scholarly works by Charles F. Meyer [4], Paul Baker [14], Douglas Bieber et al. [6], Yuji Kawaguchi et al. [19], Natthapong Chanyoo [13], John Flowerdew and Michaela Mahlberg [9], Lucas Carl Steuber [11], Antoinette Renouf [2], Andrew Hardie and Robbie Love [1], S. Hunston [16], Paul Edward Rayson [15], Miguel Fuster and Begoña Clavel [12]. It need to be noted that it is wrong to accept the notion of CL as a field such as semantics, syntax, sociolinguistics and so on, because it is not a section of linguistics but rather a methodology that does not require interpretation and description. That is to say that CL should be regarded as a research methodology that does not limit itself to only one area of linguistics but has research methods that can be used in any linguistic scientific works.

- ***Nozimjon Ataboev Bobojon o'g'li,*** *Ph.D. student at Uzbek State World Languages University, Tashkent, Uzbekistan anb929292@gmail.com ORCID id: 0000-0002-9756-6849*

It is reasonable to conclude that CL was considered as a research method in the early stages of its development and was considered as part of a network of computer science-based disciplines, but has now become a methodology with its goals and research methods and the wider scientific community. However, in the future, it is worthwhile to consider the high-tech CL as a linguistic research paradigm and / or an individual science that can solve problems in all areas, in accordance to our high-level scientific hypothesis [3][21].

Discussion

The main objectives of the CL as a methodology are: 1) development of the theoretical foundations of the field; 2) to create a school of analytical experience in the development and implementation of different types of linguistics; 3) to work out general requirements for linguistic corpora; 4) setting up corpora for educational and scientific purposes; 5) systematic improvement of the use of corpus texts in various areas of linguistics [20, P.53]. The latter task is of fundamental practical importance. In our view, it would be more appropriate if task 3 is given as to "develop the criteria for classifying linguistic corpora and the principles of structural textual formation." According to Sinclair J. [17, P.171], the corpus is a collection of natural language data and a set of linguistic texts that characterize the natural state of language in the existing language. Scholars such as Walter de Gruyter [18], Douglas Bieber [5], Charles F. Meyer [4], Lily Kucheruk [10], Guy Aston and Lou Burnard [8] cite yet another example of corpus requirements. This is the purpose of the corpus compiling. Any linguistic corpus compiler must set a specific goal before its creation, otherwise the corpus will simply be a language source and will not have a linguistic value. After all, according to G. Aston and L. Burnard [8, P.21], the corpus is not a collection of optional or randomly collected texts. As for us, a corpus is a comprehensive computer-aided automated search engine that combines the principles of processing, labeling, and compilation of sufficient quantities of linguistic texts (spoken and written) that meet the criteria of representation, formulated and categorized according to the pragmatic goals of the corpus compiler. In addition, it is supposed to have a good number of examples and consistent base for results as well as empirical analysis.

3 RESULTS AND ANALYSES

Experimental Corpus Of Corpus Linguistics (ECOCL) was created from the collection of the materials as 73 articles, 19 dissertations and 46 books based on corpus linguistics. The collected data are all in English. The materials have been gathered through the collections of scientific articles from Scopus.com and other internationally recognized scientific journals. The gathered data have been peered reviewed that allows to accept the corpus data as a representative of the linguistic features of the sphere of CL. ECOCL was made by means of software of ANTCONC. All the collected materials had been transformed into the .txt format of the documents before entering them in the corpus. In order to check if the corpus of ECOCL could respond to the principle of representativeness, the lexical layer of the texts entered in the body was automatically sorted by the number of frequencies, which was considered by empirical analysis (see Table 1). As a result, it was noted that the corpus contains 5456566 (five million four hundred fifty-six thousand and five hundred sixty-six) words (tokens), and that amount was based on the repeated use of 100652 types of words. Among these types of words, the highest frequency was recorded by functional words, namely, article, preposition, and conjunctions (see Table 1). Of these twenty top frequently used tokens, the 12th most commonly applied one is the term 'corpus', and the number of frequency is 41,926 times as a word-token. This suggests that the corpus can represent the language aspects of the CL based materials.

TABLE 1.
RESULTS DERIVED FROM CORPUS-BASED ANALYSIS

#Word Types:		*100952*
#Word Tokens:		*5456566*
№	*Frequency*	*Word*
1	*311771*	*The*
2	*213735*	*Of*
3	*141307*	*And*
4	*139836*	*In*
5	*135544*	*X*
6	*118640*	*A*
7	*112821*	*To*
8	*69881*	*Is*
9	*53084*	*That*
10	*49754*	*For*
11	*47341*	*As*
12	*41926*	*Corpus*
13	*37302*	*Are*
14	*35769*	*Be*
15	*31468*	*This*
16	*30931*	*On*
17	*30675*	*With*
18	*28894*	*It*
19	*27342*	*Or*
20	*26044*	*By*

As it has been mentioned above, any corpus should be research-based and goal-oriented. ECOCL has been a data collection for analyzing the language use in the field of CL. That might seem to be an unusual approach because CL has been analyzed through the scope of its own object. However, this shows the relevance of the work. By analyzing the corpus of ECOCL, there were many findings about the use of language in CL. For a corpus, to indicate the collocational units (CU) of a word searched. As the word corpus has been indicated in the search engine with the request of showing the collocational units that come with the target word. The results appeared on the concordance with their frequencies. The frequencies have been arranged into two types: 1) the ones preceding corpus; 2) the ones following corpus (see Table 2).

TABLE 2.
CORPUS-BASED RESULTS ON COLLOCATIONAL UNITS PRECEDING AND FOLLOWING THE TERM 'CORPUS'

#Total No. of Collocate Types:			*16561*	
#Total No. of Collocate Tokens:			*419044*	
Word	*Total frequency*	*Total CUs*	*Preceding CUs*	*Following CUs*
English	*4.85377*	*4142*	*1777*	*2365*
Learner	*5.80161*	*1530*	*1211*	*319*
Language	*3.80161*	*2139*	*950*	*1189*
National	*6.32381*	*922*	*835*	*87*
British	*5.69779*	*1009*	*766*	*243*
Spoken	*4.99988*	*1423*	*751*	*672*
Parallel	*5.89391*	*657*	*582*	*75*
International	*5.58121*	*785*	*643*	*142*
Linguistics	*6.40317*	*5495*	*539*	*4956*
Data	*4.65972*	*1913*	*504*	*1409*
Based	*5.82903*	*3787*	*510*	*3277*
Research	*4.49764*	*1453*	*360*	*1093*
Large	*4.76921*	*607*	*439*	*168*
Annotated	*5.46413*	*348*	*191*	*157*
Tagged	*5.31856*	*310*	*191*	*119*

The information gained from the table 2 allows one to make the following assumptions:

a) The highest and lowest frequencies can indicate the use of the phrases that are available to collocate with corpus.
b) The words such as learner, national, British, spoken, parallel, international, large and tagged mostly precede the term corpus, while language, linguistics, data, based and research mostly prefer to follow it collocating with it.

Moreover, the data in ECOCL enable its user to understand and to acquire more about the language use in CL. ECOCL provides the exact definitions to the terms related to the field. However, in some cases, it does not provide the explanations directly and only supplies with a number of examples to have an empiric analysis to comprehend better. Reclining this positive function of the ECOCL, some of the problematic features of the application of terms in CL can be analyzed. For example, the terms as Parsed corpus, Tag, Tagging, Token, Type can be difficult to understand or differentiate. Moreover, the plural form of the term Corpus is used in the scientific materials as Corpuses or Corpora and both are used equally. However, as a term it is supposed to have one superior form to the other. The collocations as Language corpus, Linguistic corpus and Text corpus are mostly used in the same meaning in CL. Of course, it might be really confusing for the non-native scholars in CL or just linguistics. In order to find a way to solve these problems, the role of ECOCL is valuable. In ECOCL, the grammatical plural form of term corpus has been searched and the following results came out: the frequency of corpora is 13829 and that of corpuses is 53. As it is obvious to note that the grammatical plural form of corpus should be accepted as corpora, however, it cannot reject the existence of corpuses in the language no matter how low degree of existence it has. Before creating the ECOCL, some hesitation about whether the terminological phrase parsed corpus exists or not appeared. After making the corpus and having a search for it, there has been a clear consumption because ECOCL showed 128 results for parsed

corpus in concordance with the sources on the right (see Picture 1). Analyzing the term in a wider context provided by the source materials, it was easy to understand and make a clear definition of it: parsed corpus – a corpus in which the sentences are analyzed into their constituents. In the same way, the terms as tagging, type and token have been analyzed and the following definitions have been made: Token is the smallest unit that each corpus divides to. Typically each word form and punctuation (comma, dot, ...) is a separate token. Tagging is an alternative term for annotation, especially word level annotation such as POS tagging and semantic tagging. Type is a word type in corpus that is a form of the word counted only once no matter how many times it is used repeatedly.

PICTURE 1.
INQUIRY RESULTS FOR parsed corpus IN ECOCL

Now, the next problem about the usage of alternative terminological forms as language corpus, text corpus and linguistic corpus in CL should be concerned. In ECOCL, language corpus was used 170 times, text corpus was applied 100 times and linguistic corpus was in use of 34 times. These results can show that language corpus is the most common one to apply among the others. The exact numbers and manual empirical analyses are always much more objective rather than relying on the intuition.

4 CONCLUSION

The way presented by CL research-methodology is easy and fast as well as reliable and objective. Although the corpus analyses cannot say whether an idea is true or false, there are quantitative results for the researchers that are reliable enough to a make a conclusion regarding the use of a linguistic unit in the language. The steps for reaching these important results are as following: 1) collection of the materials that might represent the whole language or sub-language; 2) creation of a language corpus with the gathered materials in first step; 3) observation of the reliability and representativeness of the corpus; 4) demonstration of the corpus analyses and searches; 5) drawing and making final conclusions about the use of a word or phrase in the language. In this regard, the analysis of the language-use characteristics of corpus linguistics could be presented successfully via creating ECOCL. By this, the problematic issues of using the terms or terminological phrases have been discussed and final assumptions were made according to the exact quantitative results derived from ECOCL.

5 REFERENCES

[1] Andrew Hardie and Robbie Love Corpus Linguistics 2013 376p
[2] Antoinette Renouf Corpus linguistics: past and present University of Central England, Birmingham 13 p
[3] Bobojon o'g'li, N. (2019) ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis. Test engineering and management, (81), 4170-4176.
[4] Charles F. Meyer English Corpus Linguistics: An Introduction University of Massachusetts at Boston. 2004. 168 p
[5] Douglas Biber University Language A corpus-based study of spoken and written registers Northern Arizona University 2006 262 p.
[6] Douglas Biber, Ulla Connor and Thomas A Upton Discourse on the Move Amsterdam 2007 290 p
[7] Guy Aston and Lou Burnard BNC Handbook Exploring the British National Corpus With SARA Edinburgh Textbooks in Empirical Linguistics Edinburgh University Press 1998 255 p.
[8] Guy Aston, Lou Burnard, The BNC Handbook: Exploring the British National Corpus with SARA, Edinburgh University Press, Edinburgh 1997, 256 p.
[9] John Flowerdew and Michaela Mahlberg Lexical Cohesion and Corpus Linguistics Amsterdam. Philadelphia – 2009. 125 p
[10] Liliya Kucheruk Modern English Legal Terminology linguistic and cognitive aspects. University of Oles Honchar – 2015. 384 p.
[11] Lucas Carl Steuber Disordered Thought, Disordered Language: A corpus-based description of the speech of individuals undergoing treatment for schizophrenia. Portland State University – 2011. 73 p
[12] Miguel Fuster and Begoña Clavel Corpus Linguistics and its Applications in Higher Education. University of Valencia – 2010. 51-67 pp.
[13] Natthapong Chanyoo A corpus-based study of connectors and thematic progression in the academic writing of Thai EFL students. University of Pittsburgh – 2013. 146 p
[14] Paul Baker Sociolinguistics and Corpus Linguistics Edinburgh University Press 2010 189 p
[15] Paul Edward Rayson Matrix: A statistical method and software tool for linguistic analysis through corpus comparison Lancaster University 2002 194 p
[16] S. Hunston Corpus LinguisticsUniversity of Birmingham 2006 234-248pp
[17] Sinclair J. Corpus, Concordance, Collocation. Oxford University Press. 1991. 179 pp.
[18] Walter de Gruyter Corpus Linguistics Berlin 2008 776 p.
[19] Yuji Kawaguchi, Toshihiro Takagaki Nobuo Tomimori Yoichiro Tsuruga Corpus-Based Perspectives in Linguistics. Tokyo 2007. 442 p.
[20] Мамонтова Виктория Валерьевна Особенности перевода сложносоставныгх слов с английского языка на русский (на материале корпуса публицистических текстов) Diss. Ставрополь – 2008. 139 с.
[21] Ataboev, N.B. (2019) ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis. Test engineering and management, V(81), 4170-4176.