November-December 2019
ISSN: 0193-4120 Page No. 4170 - 4176

ICT in Linguistic Studies: Application of Electronic Language Corpus and Corpus-based Analysis

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

Article Info
Volume 81
Page Number: 4170 - 4176
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *20 December 2019*

Abstract:
The following article deals with the application of the Information Communication Tools in the sphere of linguistics. To be more precise, it is devoted to the newly emerged language corpora and advantages of analyses derived from the use of it. The possible fields of linguistic studies in which the corpus can be applied are mentioned in the work, as well as the essential practical analysis results are considered. Concordance-based searches on modern e-corpora as BNC and COCA are available in the work that indicates the theoretical value of the research.

Keywords: *ICT, corpus, corpus analysis, BNC, COCA, concordance, linguistics.*

I. INTRODUCTION

At the time being, the world is undertaking many positive changes that are the results of the use of ICT, i.e., computers in all spheres of the world. It is noteworthy that the field of linguistics is not an exception as well. To be more precise, the emergence of the following sciences as computational linguistics, applied linguistics, computer linguistics and corpus linguistics can be a perfect proof of the inter-active relationship between computer technologies and linguistic studies. Being one of the above-mentioned, corpus linguistics fundamentally emerged in the late of 19th century. In 1897, German linguist J. Kading used a large corpus consisting of about 11 million words to analyse distribution of the letters and their sequences in German language. The impressively sized corpus corresponding with the size of a modern corpus was revolutionary at that time [8]. As is clear, the process of the corpus compiling was issued without application of ICT tools as computers at the very beginning. Later on, in the early 1960s the first electronic corpus, i.e. Brown Corpus has been created. That means that it was the first English corpus stored with the help of computers. The notion 'Corpus' refers to a collection of the natural language materials in both spoken and written format with the help of computers.

II. OBJECTIVES/PURPOSE OF THE STUDY

The main purpose of the article is to consider the language corpora as representation of a language, based on the size and number of modern linguistic corpuses, and to reveal their advantages in

field of linguistics through these corpus analyses. This purpose sets the following research objectives:

- presenting the opportunities of a language corpus;

- analytically analyzing the corpus analysis results;

- making conclusions on the advantages of using ICT, a corpus in Linguistics.

III. METHODOLOGY

The paper draws on quantitative results based on empirical analyzes, using a corpus-based approach to generate scientific conclusions.

IV. VARIOUS APPLICATIONS OF THE CORPUS ANALYSES

Corpus Linguistics is an approach that aims at investigating language and all its properties by analyzing large collections of text samples. As for S.Th. Gries, corpus linguistics is an autonomous methodological paradigm within linguistics [13; P. 4]. That means corpus analyses can be implied in many areas of linguistic researches. They are as following [11]:

- *translation studies;*
- *lexicography*;
- *teaching;*
- *sociolinguistics*;
- *discourse analysis*;
- *morphology*;
- *phonology*;
- *syntax*;
- *comparative typology*.

Now let's consider some of the above mentioned ones. They are as following:

A) Corpus analysis in translation:

In the practice of translation, an interpreter always needs a perfect knowledge of words to give a proper as well as valuable equivalent of the text into the target language. This knowledge can be based on the dictionaries, mostly bilingual ones. However, the dictionaries might not supply the users with all the required data regarding the application of words. Concerning the nature of bilingual dictionaries, Pinchuck stated that the translator should bear in mind that [12]:

(a) a dictionary, and therefore also a bilingual dictionary, is always out of date:

That means the source which is always up to date is needed, that is called a corpus of a language. The corpus based analyses can provide the researchers with the up-to-date example given in the exact numbers;

(b) many of the recorded expressions are no longer in common use:

But, corpora are the collections of the naturally occurring texts which are generally recognized among the directly chosen individuals in their native lands. Moreover, most of computer-aided corpora as COCA are added newly collected text materials annually, of course, it is a valid date of the present state of a language;

(c) expressions referred to as colloquial or non-standard may have risen into more formal use:

Whereas, the corpus gives the data and statistics and it doesn't judge itself, it allows the users to do so. In other words, in all types of annotated corpora it is common to give the source of the text and that can easily indicate the genre of the material, i.e. formal of informal, publicistic or scientific etc.;

(d) most commonly, new expressions have come into use but are not yet recorded.

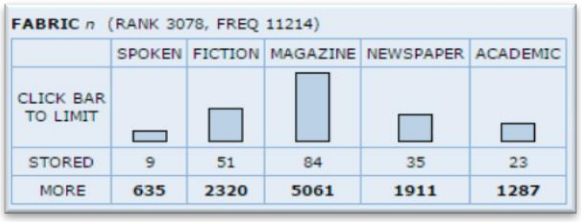
However, corpora are available to be updated daily.

For example, COCA (Corpus Of Contemporary American English) can give the results in the graphs. As it is depicted in the figure 1, the lexeme 'fabric' has been searched and the results of the frequency have been expressed according to the sections of the source materials in a bar graph. In fact, the translator can have an option to use the examples of the word use in accordance with the area of the interpretation, i.e. if the text to be translated belongs to the sphere of journalism, the interpreter needs to opt to click on

November-December 2019
ISSN: 0193-4120 Page No. 4170 - 4176

the sections of NEWSPAPER or MAGAZINE due to the needs.

Figure 1. Frequency of word usage according to sections



(The picture 1 was taken from COCA [5])

B) Corpus analysis in lexicography:

The job of creating a dictionary is requiring the compiler a great deal of difficulties as the modern life is thirsty about the dictionaries that are both easier and faster as well as valuable to supply reliable and up-to-date examples for the use of words. In order to overcome this problem, the lexicographers are applying the corpus-based analyses. In fact, the development of Corpus Linguistics has given birth to *Corpus-based Lexicography* and a new corpus-based generation of dictionaries. A large and well-constructed corpus gives quantitative information about frequency, distribution, and typicality of linguistic features – such as words, collocations, spellings, pronunciations, and grammatical constructions.

As the English scholar Meyer asserts Corpus has given a lot of contributions in language study since its emergence, but its impact on lexicography started in 1989 [Meyer 1989]. Due to advance of computer software, automated data lexicographers can now save their time and the tremendous amount of work needed for compiling a dictionary. Typically, a dictionary includes information on the part of speech, usage, meaning, pronunciation, etymology of a word. Before the advent of corpora, all this information had to be gathered manually, which can be found from the corpus results easily. To look back at the history of dictionary compiling one can conclude that the lexicographers needed to do the hard labor of collecting slips of paper containing text that they intend to include in the dictionary. For this reason, it took roughly 50 years to complete Oxford English Dictionary, which was later known as New English Dictionary. Currently, a half century is not needed to accomplish the above-mentioned process if the computer software, e.g. corpora are used.

The advantages of using corpora in dictionary compiling include the followings:

– With corpora, dictionary makers can now use a large sample of authentic spoken and written text as a source to illustrate how each word in their list is used in real life [10]. The citation used in dictionary comes from real-life discourse which provides accurate, well-defined lexical meanings in the definition of a word in dictionary;

– One huge improvement in dictionary making is the rich information available for words that have many invariant meanings such as ‘take’, ‘go’, and ‘time’, which tend to be overlooked in the previous dictionary practice [8];

– Another huge advantage of using corpora in lexicography is that information on word frequency can also be obtained. This way, lexicographers can assign whether a word is among the first 500 most common words, the next 500 and so on [10]

Moreover, this corpus represents the present-day English language, which makes it the source of real-life examples.

That’s why, one can easily make conclusion that the use of corpora in dictionary-making practices gives a lexicographer a lot of opportunities, among the most important ones are [6]:

1) to produce and revise dictionaries very quickly, thus providing up-to-date information about the language;

2) to give more complete and precise definitions since a larger number of natural examples are examined;

3) to keep on top of new words entering the language, or existing words changing their meanings;

4) to describe usages of words or phrases typical for particular varieties and genres;

5) to organize examples extracted from corpora in order to explain words and collocations;

6) to treat phrases and collocations more systematically due to the existence of mutual information tools which establish relationship between co-occurring words;

7) to register cultural connotations and underlying ideologies which a language has.

Here, we have taken several analyses on the expression of the lexeme 'fabric' in a dictionary we have chosen was *the Cambridge Advanced Learner's Dictionary, Cambridge University Press 2008* [3]. According to this dictionary we could find the following data concerning the word '*fabric*':

fabric /ˈfæb.rɪk/ *noun* **CLOTH**

1. [C or U] (a type of) cloth or woven material

dress fabric

seats upholstered in hard-wearing fabric cotton fabrics

fabric /ˈfæb.rɪk/ *noun* **STRUCTURE**

2. **the fabric of *sth***

the structure or parts especially of a social unit or a building

the fabric of society

Unhappiness was woven into the natural fabric of people's lives.

We must invest in the fabric of our hospitals and start rebuilding them.

Linen is a coarse-grained fabric.

The fibres are woven into fabric.

This fabric is similar to wool, only cheaper.

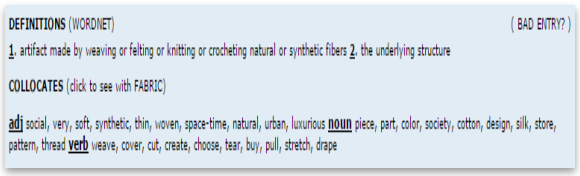
The designs are printed onto the fabric by hand.

She designs her own fabrics using woodblocks and stencils to create patterns on the material.

However, in our opinion, the dictionary should provide the user with more information about the word forms, parts of speech in one search and, of course, more examples ought to be provided. This dictionary explains the meaning of the word very well, however, lack of examples cannot give a general view of understanding it. It means that the users are not provided with enough examples as if only theory but not practice. No synonyms or antonyms are mentioned that means the learners cannot acquire sufficiently.

It would be better to compare the use of the dictionary with the corpus-based results. Practically, if one needs the definition of the word 'fabric', it is possible to search for it from the corpus 'COCA' (see picture 1)

Picture 1. Definition of the word



(The picture 1 was taken from Corpus of Google English and COCA [5])

Moreover, it provides the users with the information about the word formation of the word 'fabric' in one table as it is shown in the table 1. Here, the signs have their specific meanings, that is, N-noun, V-verb, ADJ-adjective. A glance at the given table supplies the striking reveals that N-noun form of the word is the most common among the others, indicating the total amount of 11,214 examples. Another very important feature of this table is that it allowed us to find examples according to the genre of usage. One can make an assumption that the word 'fabric' is mostly used in magazines comparing to the rest of the genres.

November-December 2019
ISSN: 0193-4120 Page No. 4170 - 4176

Table 1. Word formation structure

CLICK ON A WORD FOR CONTEXT AND SYNONYMS (CAN ALSO LIMIT TO GENRE)

	RANK #	PoS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	3078	N	FABRIC	11214	635	2320	5061	1911	1287
2	10618	V	FABRICATE	1511	199	126	403	238	545
3	11783	N	FABRICATION	1296	125	114	262	152	643
4	28331	J	FABRICATED	187	15	22	47	32	71
5	32366	N	FABRICATOR	146	10	11	44	34	47
6	51714	J	FABRIC-COVERED	32	1	8	15	7	1

(The table 1 was taken from COCA [5])

Moreover, if the information about the etymology of the word is needed, the search results can be given in the following way. As it is depicted in figure 2, the word use frequency has been indicated through ups and downs per years in the horizontal axis. This refers to the development of the word use in a language.

Figure 2. Chronological line of word usage in a graph.

(Figure 2 was taken from Corpus of GOOGLE BOOKS: British English [7])

C) *Corpus analysis in teaching foreign languages*:

In the EFL classes, the teacher always needs both reliable and authentic teaching materials to make the classes better in quality. In this regard, the corpus analyses can provide the educators with the data they need, i.e. the annotation about the parts of speech as well as synonymic or antonymic lines of the word semantically. Corpus German linguist, Ute Römer divides the use of corpora in language learning and teaching into two types: [14]

- indirect applications: hands-on for researchers and materials writers;

Here, the author includes the specialized corpora and the research results which can help the language learning process.

- direct applications: hands-on for teachers and learners (data driven learning, DDL).

Whereas, the direct way of use indicates the important features of the general corpora from which every learner can do searches so as to get the necessary information like grammar constructions and others. Moreover, the teachers are able to visit the corpora in order to get some authentic materials for teaching purposes.

Here are some of the examples which show the eficiency of the corpus anaysis in teaching languages. As for us, the practical application of corpus analysis in the classroom can be devided in three categories in accordance with the purpose and the audience of the users. In our opinion, they can be classified as following:

- using a linguistic corpus as teacher:
  - for textual materials;
  - for collocational examples;
  - for the task evolution on grammar, phonetics, and vocabulary.
- using a linguistic corpus as a language learner:
  - for self correction;
  - for getting more examples to acquire knowledge about the word;
  - for downloading an authentic material.
- using a linguistic corpus as a linguist or a student majoring in linguistics, translation and philology:
  - for compiling corpora with specific aims to investigate a language of a part of the language;
  - for making scientific assumptions about the target language;
  - for the authentic evidences concerning the language use.

Actually, both language teachers and learners can use the corpus for making sure about the application of words in terms of grammar and spelling. For example, the teacher wants to teach the

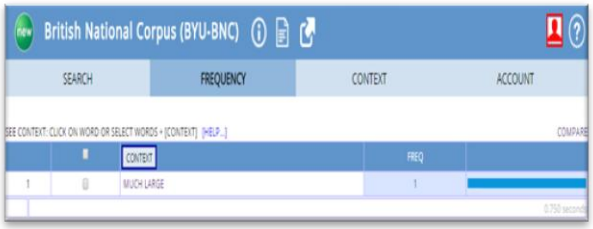
November-December 2019
ISSN: 0193-4120 Page No. 4170 - 4176

degrees of adjective in English and the usage of adverbs of degree with the comparative adjective forms. For this, the teacher needs to prove that 'much larger' is common collocation what the English speakers use, but 'much large' is not. Here, corpus should be applied. In BNC (British National Corpus), the two collocations have been searched and the results are in Picture 2 and picture 3. As it is clear, the number of results for 'much larger' is 299, while the quantity of results for 'much large' is only 1. That means, the provement of the grammar rules can be derived from corpus-based analyses.

Picture 2. The frequency results for 'Much larger'

(The picture 2 was taken from BNC [2])

Picture 3. The frequency results for 'Much large'



(The picture 3 was taken from BNC [2])

V. CONCLUSION

To sum up, corpus analysis can be used in different spheres of linguistics. The use of corpus analysis in the practice of translation provides the users with many advantages as the old-fashioned dictionaries do not meet the requirements demanded by the translator at the time being. The next issue undertaken in the article was the application of the corpus results in compiling dictionaries. Comparing the history and the present day lexicography, in both periods the lexicographers have a great need for the reliable source of the target language which is authentic. In this sense, we consider the language corpora to be very useful tool for researchers supply them with the infinite number of the examples regarding the target word. Moreover, the comparison of the traditional dictionaries existing now and the corpus-based dictionaries gives an overview that compiling dictionaries aiding corpus analysis is more superior to the latter. A language corpus contains not only a bunch of examples but also the etymology of the words, synonyms and the collocational units can be revealed from the corpora. The analysis gained from a corpus of any language can be utilized in the sphere of teaching foreign languages as well. As far as we are concerned, the application of corpus results are available to be applied in learning process, besides, according to the users' status in the classroom and their purpose of using the corpus results can be in different use. Taking into account the mentioned tasks and their efficiencies, we think that the use of corpus analysis by the teachers in the foreign language classrooms would be more successful. In our point of view, the application of corpus analysis in teaching is a modern trend and, that's why, the interests by the learners would be higher because of being computer aided. We are strongly convinced that every lesson designed with the use of corpora will be much superior to any other traditional language teaching classrooms. It is because of the unlimited source of data concerning the target language provided by the corpus.

REFERENCES

[1] Amy B M Tsui *What teachers have always wanted to know - and how corpora can help*, The University of Hong Kong, 2009. 177 pp.

[2] British National Corpus http://corpus.byu.edu/bnc/ [Accessed on September 12, 2019]

[3] Cambridge Advanced Learner's Dictionary © Cambridge University Press 2008 Produced for Cambridge University Press by Armada.

[4] Casey Mari Keck *Corpus linguistics and language teaching research: bridging the gap* Language Teaching Research 2004; Northern Arizona University. 83 pp.
[5] Corpus of Contemporary American English http://corpus.byu.edu/coca/ [Accessed on September 12, 2019]
[6] Geoffrey N. Leech, Corpora: The Linguistics Encyclopedia, ed. by Kirsten Malmkjaer. – Routledge, 1995. 256 pp.
[7] Google books British English: http://googlebooks.byu.edu/help/intro_e.asp?w=&h= [Accessed on September 15, 2019]
[8] Hans Lindquist. Corpus Linguistics and the Description of English. – Edinburg: Edinburgh University Press, 2009. 219 pp.
[9] Lindquist, H. Corpus Linguistics and the Description of English. Edinburgh: Edinburgh University Press. 2009. 224 pp.
[10] Meyer, C.F. English Corpus Linguistics.Cambridge: Cambridge University Press. 2002. 168 pp.
[11] Nikola DobrićMisic, Lopicic (Eds..) Language, literature and identity, Serbia. 2009, pp. 359-363.
[12] Pinchuck, I. Scientific and Technical Translation. London: Deutsch. 1977. 264 pp.
[13] Stefan Th. Gries and Anatol Stefanowitsch (eds.). Corpora in Cognitive Linguistics. Corpus Based Approaches to Syntax and Lexis. Trends in Linguistics: Studies and Monographs 172. New York: Mouton de Gruyter, 2006, 352 pp.
[14] Ute Römer, I. *Origin and history of corpus linguistics _ corpus linguistics vis-a`-vis other disciplines// 7. Corpora and language teaching*: Hannover (Germany) – 2008. P. 112-122