Volume 2, Issue 5, May - 2024 **ISSN (E):** 2938-3811

COCA CORPUS AND ITS COMPONENTS

Fozilova Maftuna Mirzohidovna
Student of Bukhara State University

Abstract

This article provides detailed information about the COCA (The Corpus of Contemporary American English) corpus and its components. The content of the corpus is analyzed from the followig point of views as number of words , type , and genres and others The method used in carrying out the study is descriptive analysis method. The result of the research demonstrates that there are collocational competence of EFL students, history of the CL , genres in the COCA corpus.

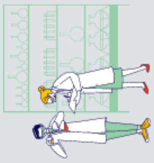
Keywords: COCA corpus, American English , components , linguistic , language , words , genres.

Introduction

Nowodays a **corpus** mentioned as a collection of texts . According to Sinclear , a corpus is selection of real – life language texts that represent a specific state or form of a language . In addition to this illustrative quote , there is today a growing consensus that a corpus is a collection of machine-readable authentic texts sampled to be representative . Thus a corpus is a large principled collection of natural examples of language stored electronically. Thus , the creation of concordances in pre-electronic studies of corpus linguisti c and their treatment as dictionaries or indexes laid the foundation for the emergence of corpora [1: Rakhimov M] . One of the most famous corpus is COCA (The Corpus of Contemporary American English) is a one –billion-word corpus of contemporary American English . IT was created by Mark Davies , retired professor of corpus linguistics at Brigham Young University (BYU).

Main Body.

The Corpus of Contemporary American English (COCA) is one of the largest corpora of American English with over one billion words (November 2021) from various sources collected from magazines , web pages , conversation , and more , thus serving as a comprehensive source for research exploring language patterns across severalregisters or genres . Put differently, this corpus's diverse range of texts enables scholars to comprehend language use in a variety of setting . The corpus is constantly growing: In 2009 it contained more than 385 million words ; In 2010 the corpus grew in size to 400 words ; by March 2019 , the corpus had grown to 560 million words. Thus, for example, from October 1 to October 31, 2019 the number of site users reached 130,000 . Attracting more than hundred



thousand users per month , the most well-development linguistic corpus of this site is COCA . [2:120 pg.] .

The COCA corpora is by far the most widelu-used of these corpora . In early 2020 , dramatically expanded the scope and measurement and aspects of COCA to make it even higher beneficial for researchers , teachers , and learners . The following research about benefited from the elements online corpora related to lexical sophistication and built-in the Corpus of Contemouraty American English (COCA) , www.americancorpus.org , into the writing syllabus of Lebanese EFL undergraduates . COCA was chosen for its significant advantages . It is among the biggest online corpora and is accessible for all clients regardless of their linguistic knowledge . The corpus offers a clear show of phrase frequency vi a its five registers : spoken, news, academic, finction, and magazine .

Results.

The study has reached three fundamental results . First , employing the COCA as a pedagogical corpus tool can enhance the collocational competence of EFL students should a corpus- driven approach be used descriptively in the classroom . Second , the two methodological stages of demonstration and praxis could facilitade the process of topical priority as a significant index of collocational usage and its thematic relevance . Third , more empirically , the naturally occurring collocates of the node ”coronavirus ” have proven significant to the pedagogical situation of teaching the node’s collocational meanings encoded in the syntactic categories of nouns , verbs , adjactives , and adverbs e.g. infection , cause , novel , closely , and respectively . [3:210]

Corpus analyses show that the properties of the collection units provided are important . As an example , in order to analyze the semantic properties of the verbs cause and result in , which have a comparable meaning , phrases that are collocated with them in the COCA were searched . Consequently , the concordance of the corpus used to be used to search for nouns that acquired right here as an object of the verbs cause and result in . In the concordance lines it was found that the verb cause was used 71078 times , and the verb result in was used 84001 times .

J.Sinclair , a scholar who was the was the first in the history of CL to use phrases and individual words using linguistic corpora, analyzed two different phrases. They are : naked eye and true feeling. Examples of the use of phrases in his work are taken from the BNC corpus, and while 148 examples were found for the naked eye , for true feeling this number was 53 . When these collocations were searched in COCA , it was observed that the number of the phrases increased significantly . That is, 654 and 175 results were obtained , respectively . This , in turn, leads to a broader discussion of the phraseological unit sought. [4: 30-36 pg]

The corpus carries greater than one billion words of data , inclusive of 20 million phrases every year from 1990-2019 (with the identical style stability yr with the aid of year). This makes COCA the only corpus of English that is 1) large 2) recent 3) has a vast very of genres . The following indicates the genres in the corpus :

Genre # texts # wordsn # Explanation

Spoken -- 44,803 / 127,396,932 / Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)

Fiction-- 25,992 / 119,505,305 / Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.

Magazines-- 86,292 / 127,352,03 0 / Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.

Newspapers-- 90,243 / 122,958,016 / Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution,

San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.

Academic-- 26,137 / 120,988,361 / More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business

Web (Genl)-- 88,989 / 129,899,42 7 / Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.

Web (Blog)-- 98,748 / 125,496,216 / Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.

TV/Movies --23,975 129,293,467 Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual

"spoken corpora".

- Text : 485,179
- Words : 1,002,889,754 [5 : 1] .

Discussion

The result of this study indicated that students and to English learners become aware of the use of online corpus (COCA) to be beneficial for their English vocabulary development . The results also show that most of the language learners (92.6%) have positive attitudes towards using online corpus in order to increase their vocabulary , helps them to learn collocations and phrases without having difficulty in learning .

As Sinclair said who was the first in the first in the history of CL to use phrases and individual words using linguistic corpora , analyzed two different phrases . They are naked eye and true feeling . For example , in current English the adjective *glad* is found only in number of predicative constructions , glad that ...., glad of ...., glad to ...., etc.., with a rich pattern of collocation in these structures .

The COCA corpus contains more than one billion words of text from eight genres and there has so much data from each of these genres , it provides useful statistics about the frequency of words , phrases , and grammatical constructions across the genres - whether they are very informal.

Conclusion

COCA has a qualitity of elements that units it apart from any different corpus . These encompass its **size** (1.0 billion words) , how upon **to data** it is (texts through Dec 2019) , **genres** (TV/Movie subtitles , spoken ,blogs , webs , finction , magazine , newspaper , academic) , and its **searches** (range of query types , the ease and speed of its searches) , including the ability to limit and to compare across genres and time periods .

All of these features make COCA the ideal corpuss for researchers, teachers and language learners .

References:

1. Amirbek F . 120 page
2. Amir H. Y. Salama , Waheed M. A .Altohami "Enhancing EFL Student's COCA – Induced Collocational Usage of Coronavirus : A Corpus – Driven Approach . " IJACSA – journal Vol.13 , No.2.2022 , 210-page .
3. Rakhimov M. " Korpus Linguistikasi Taraqqiyoti va O'zbek Tilshunosligida Korpus Tahlil Asoslari." In 2023 , 155 -page.
4. Sinclair J. Trust the text : language , corpus , and ddiscourse . London : - Routledge . 2004 . 30-36 page .
5. The COCA corpus (new version released March 2020) https://www.english-corpora.org , 1- page.