Impact Factor: 9.2 ISSN-L: 2544-980X

CORPUS ANALYSIS IN THE EXPRESSION OF STATUS VERBS

Karimov Nodirbek Nosirjon o‘g’li[1]

A corpus is a language resource consisting of a large and systematized set of texts. In corpus linguistics, they are used to perform statistical analyses, to test views, linguistic phenomena or theoretical rules within a specific language or a specific section of the language. A corpus can consist of textual data in one language or several languages. A corpus usually means a textual corpus, but nowadays corpora are no longer just texts. Therefore, instead of the word corpus, we use the concept of text corpus. Corpora are annotated to make language research more efficient. For example, one type of corpus annotation is word tagging (POS-tagging). This means tagging based on the category of the word and the categories of this category. That is, the word "kutdim" carries the following information: verb, singular, tense, person-number. The same information is attached to the word through tags. Another form of annotation is lemmatization, which is to indicate the base form of a word. For example, the base of the words "kutdim", "kutgandim", "kutganimga" is the same - the verb "to wait". This is called lemmatization. The concepts of root and base should not be confused here. For example, the word "bostirma" is formed in the form "bostir+ma", but we cannot consider the word “bostir” as a lemma in its rooting, “bostirma” is a single word. If we need to root the words "bostirmada", "bostirmaga", "bostirmaning", then it will be correct to take the word suppression. In simple terms, a lemma is a part of a word that omits form-forming suffixes.

Simply put, the corpus helps to conduct research in any field of linguistics to be qualitative and effective. A corpus is a set of texts submitted to a search program in order to determine the characteristics of language units, a set of written or spoken texts in natural language stored in electronic form, a set of texts that are placed in a computerized search system on the basis of software, and which work in an online or offline system. Linguistic corpora are an undeniable tool for language research and problem solving. It differs from a regular electronic library. The purpose of the electronic library is to cover the artistic and journalistic works reflecting the socio-political, spiritual and economic life of the people. Texts of the electronic library are not processed from the point of view of

[1] PhD student of Namangan State University e-mail: topcoder1600@gmail.com

language, so they are inconvenient for research. Because the electronic library will not be created for the purpose of preparing a base of scientific research material, but will aim to collect the national spiritual heritage. Unlike the electronic library, the language corpus means the collection of necessary, useful and interesting texts for language learning and research. The first factor that distinguishes a corpus from an electronic library is the character of the text and its enrichment with additional information. If users need a word, a standard text editor will find it. But working with a software system that understands the meaning, content and structure of the language phenomenon in the text is very preferable and convenient. Searching for a language unit, if needed, such software, a corpus, can be of great help to the researcher or user. While it took months and years for a researcher to find examples for his work and transfer them to a file (in the era before the development of computer technology), today, with the help of world language corpora, he has the opportunity to find and work on hundreds of examples in a matter of minutes.

A special search system consists of several sets of programs designed to obtain information from the corpus. It provides statistical information and search results in a user-friendly format. It is advisable to expand the coverage of the corpus and use the material of written and oral speech in order to clearly imagine what process is taking place in the language. With the help of such a corpus, it is possible to make a clear conclusion about the change that has occurred and is expected in the language as a result of development. Corpus linguistics involves the collection and analysis of collections of spoken and written texts as a source of evidence describing the nature, structure, and use of languages. This work leads to a quantitative measure for describing languages, usually including information about the probability of occurrence of linguistic objects or processes in specific contexts. Corpora vary greatly in size and design, but most are available electronically with computer programs specifically designed to support analysis. Current corpus linguistics, which studies all levels of language, including phonology, lexicology, grammar, and discourse, has grown out of a long tradition of using texts as an empirical basis for linguistic description.

A corpora is often annotated to represent grammatical classes and functions. Software to analyze grammatical structures or identify compounds using concordance has revolutionized text analysis. Corpus linguistics has shed light on the systematic change of languages in different historical, regional, and sociolinguistic contexts, genres, and registers, focusing on the company that tends to preserve particular words and the ways in which language users typically express themselves. Probabilistic descriptions of languages can complement other methodologies used by linguists and have implications for work in a range of fields beyond linguistic description. These include Natural Language Processing, Language Learning and Cognitive Linguistics.

The term corpus linguistics generally refers to corpus-based linguistic research (Biber et al., 1998; Tognini-Bonelli, 2001 et al.). Archetypal corpus work existed long before the modern digital

age, as exemplified by early efforts to index words and harmonize the Christian Bible in the thirteenth century. However, the emergence of corpus linguistics as an academic discipline is closely related to the rapid development of computer technology, as well as the availability of digital tools to record, store, and examine corpora associated with the availability of digital content from the second half of the twentieth century. Although the use of computational and statistical tools plays a crucial role in corpus linguistics today, it should be noted that two essential elements in corpus linguistics are the study of linguistic issues and the use of corpora. Thus, the term "armchair linguistics" popularized by Charles Fillmore in 1992 (Fillmore, 1992) applies to corpus linguistics to the extent that both linguistic theories and argumentation are required to provide a foundation for corpus-based processing and analysis will be done.

The first modern corpus is the present-day Brown University Standard Corpus of American English ("Brown Corpus"; Francis & Kučera, 1979; Kučera & Francis, 1967), compiled at Brown University in the 1960s. With approximately one million words of well-balanced, diversely sourced text, The Brown Case is a breakthrough in corpus size and corpus design. But in the following years, the standard of hull size increased rapidly. Thanks to the improved storage and processing power of modern computers, it is now not uncommon to encounter a corpus containing hundreds or thousands of millions of words.

What came with the increase in scale was the expansion of the genre. In addition to "balanced" corpora such as brown corpora (see definition of balanced below), corpora are also available for highly specialized domains and non-traditional data types (e.g., multimodal language data, collected text) developed.

A corpus is a collection of systematically collected texts or transcribed discourses to represent a specific function and description of a language that can serve as a basis for linguistic analysis.

While conducting research on the corpus analysis of case verbs, we will first bring to your attention the analysis in the pragmatic direction, referring to the largest syntactic unit.

TEXT

“Hasan bugun bozorga sehrlandi. U yo‘lda ketar ekan, sharqirab oqqan suvning qo‘shig‘idan sehrlangandek ariq bo‘yida to‘xtadi. Qalbida nechuk kemtiklik, bir yeri xuvullardi. Ko‘p o‘tmay hayolida o‘sha Mastura gavdalandi. Uning kulgilari, qo‘shiqlari qulog‘i ostida jonlandi. Go‘zal yuzlari, ko‘zlari nur taratardi. Egnidagi yarashgan libosi lov-lov tovlanardi. Birpas hushi yo‘qoldi. O‘zini qo‘lga olgan Hasan yana yo‘lida odimladi. O‘tmishni eslarkan, chuqur xo‘rsinib qo‘ydi. ” (Dilafruz No‘monova)

This text contains 58 words and phrases, 15 of which are verbs. We will see their analysis below.

STATE VERBS	GRAMMAR	QUANTITY

	INDICATORS	
Sehrlandi, sehrlangandek	Sehr+lan+di, sehr+lan+gan+dek	
Sharqirab	Sharq+ira+b	
Xuvullardi	Xuv+ulla+r+di	
Gavdalandi	Gavda+lan+di	
Jonlandi	Jon+lan+di	
Nur taratardi	Nur tara+t+ar+di	
Yarashgan	Yarash+gan	
Tovlanardi	Tovla+n+ar+di	
Yo‘qoldi	Yo‘q+ol+di	
Qo‘lga olgan	Qo‘l+ga ol+gan	
Odimladi	Odim+la+di	
Eslarkan	Es+la+r+kan	
Xo‘rsinib qo‘ydi	Xo‘rsin+ib qo‘y+di	

Hence, corpus linguistics is one of the resources describing the structure and use of languages as well as processing in computer science for various applications such as natural language or language teaching and learning in language education (Kennedy 1998). Because corpus linguistics is large texts and observing the frequency of certain linguistic units (such as words or grammar) reveals categories in the corpus.