THE INSTRUMENTAL OBJECTIVE ROLE OF CORPUS IN RESEARCHES OF VARIOUS FIELDS

Vasliddinova Kamola
Phd student of Uzbekistan State World Languages University.

Annotation: In this article the instrumental objective role of corpus in the scientific researches is given and shows how the corpus helps to improve efficiency in researches and using of the corpus in the language learning process in education system and its fairy effective aspects in other fields are highlighted.

Keywords: Corpus, computational linguistics, corpus theory, corpus and discource, corpus-based methods, research, assessments, chat bot, server.

Аннотация: В данной статье представлена инструментальная объективная роль корпуса в научных исследованиях и показано, как корпус помогает повысить эффективность исследований и использования корпуса в процессе изучения языка в системе образования, а также выделены его эффективные аспекты в других областях.

Ключевые слова: Корпус, компьютерная лингвистика, корпусная теория, корпус и дискурс, корпусные методы, исследования, оценки, чат-бот, сервер.

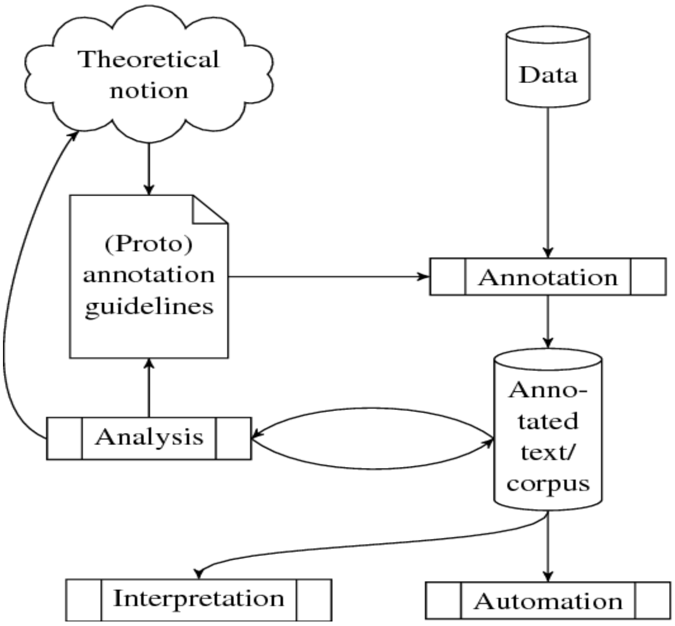
Linguists have relied on various methodologies for many years to unravel the mysteries of human communication from deciphering ancient scripts to analyzing contemporary discourse. Among these methodologies, corpus linguistics stands out as a powerful tool that has revolutionized the study of language in the digital age.

Corpus linguistics involves the systematic analysis of large collections of text, known as corpora, to investigate patterns of language use. These corpora can encompass a wide range of texts, including written works, transcripts of spoken language, social media posts, and more. By examining these vast repositories of linguistic data, researchers can uncover insights into vocabulary usage, grammatical structures, semantic relationships, and even cultural phenomena.

The interest to a corpus in the field of linguistics has been growing rapidly over the ten-year periods. Corpus in the simple term is defined as the compilation of texts that has been gathered for a specific reason [W.Cheng., 2011: 2].

The meaning of the term "corpus" varies slightly throughout academic areas. It usually refers to a grouping of texts; in literary studies; in theology, any collection of data (narrative texts or single sentences) elicited for linguistic research [Sebba, Fligelstone.,1994: 769].

It is clearly that early corpus-based research was carried out in English and then became the main focus of research by linguists around the world. Although corpus-based methods emerged in the 1960s, they have emerged as a new field in linguistics since the 1980s and corpora have developed new theories and directions in various linguistic fields.



1-rasm. Framework of corpus-linguistic techniques in the process of theorizing.

The first researches carried out in the field of corpus were in the areas of educational processes, in the field of linguistics, cross-cultural studies and corpus & discourse. On learning processes Barker (2010), Bieber (2012), Boulton & Tyne (2015), Chen (2014), Conrad (2009), Fuster Márquez & Clavel, Arroitia (2010), K Hyland (2012), K Hyland & Wong (2013), Lee & Webster (2012), Ma (2012), Walsh (2010) carried out early basic scientific research. In the field of linguistics, the scientific works of Gries (2009), Kaszubski (2003), Moisl (2015), Cheng (2011), Lukin (2017) are highly recognized. Wiegand & Mahlberg (2019) investigated the importance of corpora in cross-cultural research. Baker (2006) and Partington & Marchi (2015) elaborated on the differences and similarities between corpus and discourse [A.Mohammed Saleh Al-Hamzi, A.Gougui, Y.S.Amalia, T.Suhardijanto., 2020: 176-181].

If we think over that corpus helps to evaluate scientific researches, can any researches use to carry out their searches with the help of corpus? Before answering this question, we should analyze how to do it and with which method a researcher go to a destination. Elena Tognini-Bonelli (2001) has made a distinction between "corpus-based" research and "corpus-driven" research [J. Jitpaiboon1, A.Rungswang., 2022: 151].

In corpus- based research a researcher looks for in corpus with a specific intended approach. That is to say, by grammar or statistical rules a researcher tries to analyze how is the rule working? how it should be in fact? how many language users follow the rules or how many of them use wrong in context or speech.

In contrast to the method described above, "corpus-driven" approach serves to see what types of language patterns surface from the corpus. Different language variation examples come out during the analysis like lexical bundles other language varieties and collocations. Corpus - driven method especially evaluates the efficiency of language teaching in classrooms. Because they serve as the authentic materials in teaching.

A number of advantages of using corpora in language teaching and learning have been identified by many corpus linguists and thereby its implementation into the language classroom has been highly recommended. The advocates of using corpora have argued that corpora can provide a powerful tool with which learners can explore and discover patterns of authentic language, providing such information as collocations, colligation, and semantic prosody that are hardly obtainable [Lee, Shinwoong., 2011: 159-178].

According to famous corpus researchers Aijmer, Kaltenbock, Mehlmauer-Larcher, it has been also contended that corpus-based language teaching has potentials to motivate learners and promote learner autonomy that are highly valued in pedagogy. I absolutely agree the mentioned ideas that at present we may take real life materials to enhance classroom environment with the help of corpus. In the following we may see some agreements by a few scolars who are doing on corpus.

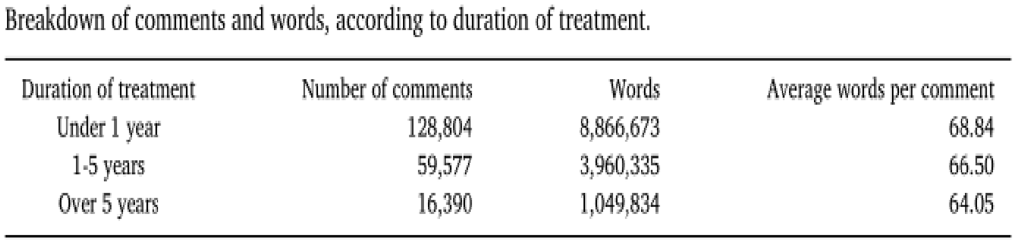
Due to those potentials of corpora in language teaching and learning, a number of researchers (Aston, 1997; Braun, 2007; Conrad, 2004; Hunston, 2002; Tribble, 2001) offered them as an inventive teaching tool and a useful resource, and language specialists have begun to view their use as fairly trendy.

Many projects also review relevant factors related to the use of corpora in higher education. In a certain study titled "Corpus Linguistics and its Applications in Higher Education" by Fuster Márquez & Clavel Arroitia (2010), they are set out to depict implied essentials of corpus linguistics and its progress in relevance to theoretical linguistics and its implementations in modern teaching pursuits.

Although the first researches have been done in linguistics and methodology, later corpus has become one of the main objects to investigate in different fields such as politics, economics and especially medicine in recent years.

Firstly, we should mention that the latest researches in medicine especially statistical ones are being done on corpus. For example, based on 214,340 written patient comments (14,403,694 words) about National Health Service (NHS) cancer care in England from 2015 to 2018, the analysis was conducted with the help of corpus-based method. Patients are categorized based on the length of their therapy, and their qualitative assessments of their experiences are contrasted based on the keywords that best describe their language, the themes of their positive and negative comments, and the feedback ratings they provided [G. Brookes, P. Baker., 2021:1].

Another copus-based quantitative analysis is related to expressions of epistemicity. That is to say, a category covering the expression of commitment to the information transmitted and comprising epistemic modality and evidentiality, in a corpus of 400 newspaper articles from The Guardian concerning the COVID-19 pandemic.



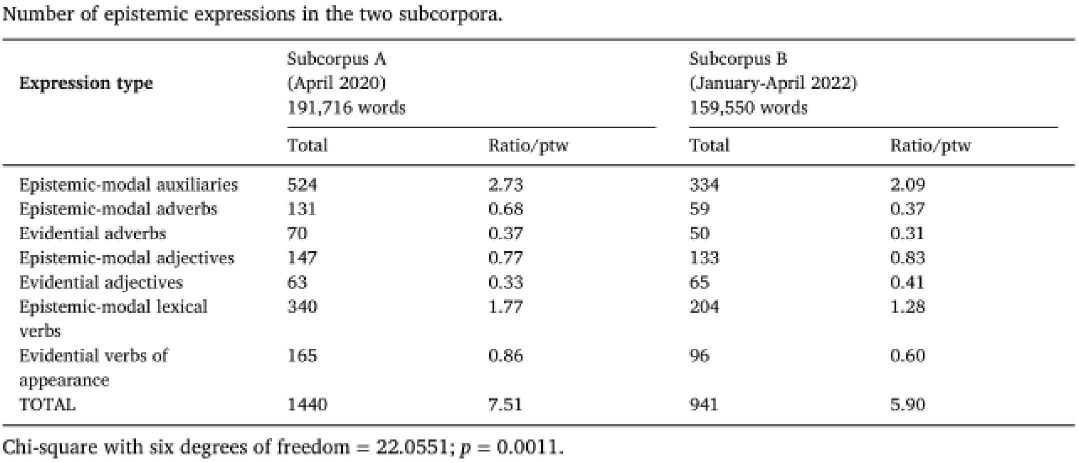
Breakdown of comments and words, according to duration of treatment.

Duration of treatment	Number of comments	Words	Average words per comment
Under 1 year	128,804	8,866,673	68.84
1-5 years	59,577	3,960,335	66.50
Over 5 years	16,390	1,049,834	64.05

2-rasm. Breakdown of comments and words according to duration of treatment in the Covid-19 pandemic [G. Brookes, P. Baker., 2021:1].

200 articles were written in April 2020; the other 200 were written between January and April 2022, after massive vaccination and an extraordinary increase in medical knowledge. The analysis distinguishes between a number of subtypes of epistemic expressions and three kinds of

authorial voice. The results show that the April 2020 articles contain more epistemic expressions, of both weak commitment (might, perhaps, apparently …) and strong commitment (know, clearly, surely), which suggests a greater need to distinguish the known from the unknown in this period, due to the pervasive state of uncertainty. The analysis has social implications, since it gives readers an opportunity to appreciate the careful assessments of epistemicity found in the corpus and therefore to consider the convenience of obtaining information from quality media. These social implications, together with the methodology of the analysis, contribute to the potential of the paper for pedagogical applications [C.Jones, D.Oakey, K.O'Halloran., 2023:1].



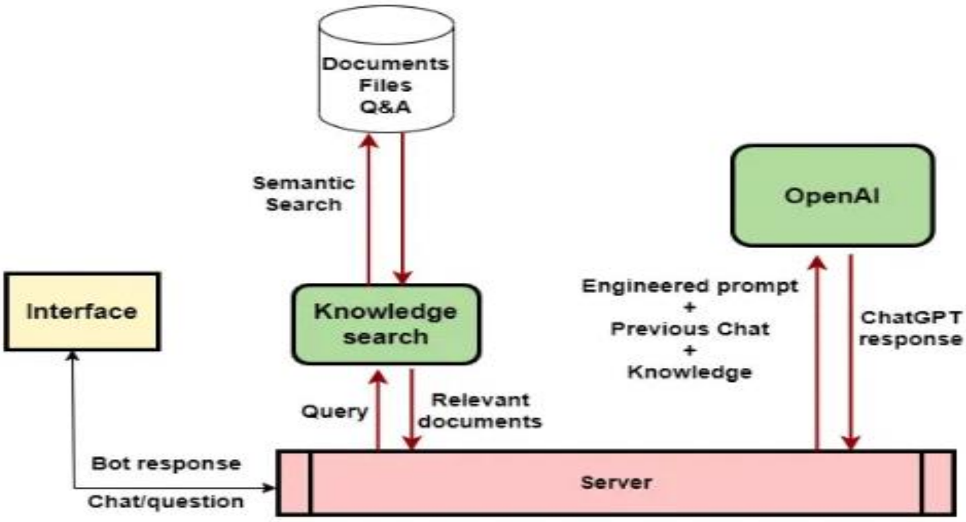
Number of epistemic expressions in the two subcorpora.

Expression type	Subcorpus A (April 2020) 191,716 words		Subcorpus B (January-April 2022) 159,550 words	
	Total	Ratio/ptw	Total	Ratio/ptw
Epistemic-modal auxiliaries	524	2.73	334	2.09
Epistemic-modal adverbs	131	0.68	59	0.37
Evidential adverbs	70	0.37	50	0.31
Epistemic-modal adjectives	147	0.77	133	0.83
Evidential adjectives	63	0.33	65	0.41
Epistemic-modal lexical verbs	340	1.77	204	1.28
Evidential verbs of appearance	165	0.86	96	0.60
TOTAL	1440	7.51	941	5.90

Chi-square with six degrees of freedom = 22.0551; $p = 0.0011$.

3-rasm. Number of epistemic expressions in the two subcorpora. [C.Jones, D.Oakey, K.O'Halloran., 2023:1].

We know that these days chat bots have become popular in this digital era. A chatbot is a machine conversation system which interacts with human users via natural conversational language. Software to machine-learn conversational patterns from a transcribed dialogue corpus has been used to generate a range of chatbots speaking various languages and sublanguages including varieties of English, as well as French, Arabic and Afrikaans.



4-rasm. A functional model of chat bot.

One of the most powerful language models ChatGPT was created by OpenAI in 2022 and this model has the ability to understand and interpret human language, making it a valuable tool for a wide range of applications. The model is trained on a large corpus of text data, which enables it to generate human-like responses to user queries.

In conclusion we can see that corpus has been becoming a central object to conduct researches not only in linguistics, but also in various fileds like economy, medicine, politics and others. Corpus focuses and characterizes authenticity, for that reason it is so fairy tool or method to carry out, investigate or analyze.

References:

1. A.Mohammed Saleh Al-Hamzi, A.Gougui, Y.S.Amalia, T.Suhardijanto. Corpus Linguistics and Corpus-Based Research and its Implication in Applied Linguistics: A Systematic Review Parole: Journal of Linguistics and Education, 10 (2), 2020. pp. 176-181.
2. B.A.Shawar, E.S.Atweel. Using corpora in machine-learning chatbot systems. International journal of corpus linguistics, vol.10, 4, 2025. pp. 489-516. https://doi.org/10.1075/ijcl.10.4.06sha
3. C.Jones, D.Oakey, K.O’Halloran. I will say the picture of the background is not related to the words: using corpus linguistics and focus groups to reveal how speakers of English as an additional language perceive the effectiveness of the phraseology and imagery in UK public health tweets during COVID-19. Applied Corpus Linguistics 3, 2023. P.1
4. G. Brookes, P. Baker. Patient feedback and duration of treatment: A corpus-based analysis of written comments on cancer care in England. Applied Corpus Linguistics 1, 2021. P.1. https://doi.org/10.1016/j.acorp.2021.100010
5. J. Jitpaiboon1, A.Rungswang. Trends of Corpus Linguistics Used in English for Specific Purposes Research: A Case of Asian ESP Journal.2022.20(2), P.151
6. Lee, Shinwoong. Challenges of Using Corpora in Language Teaching and Learning. Linguistic Research.2011. 28(1), P.159-178
7. Sebba, Fligelstone conference ,1994. P.769
8. W.Cheng. Exploring corpus linguistics, London: 2011. – P.2