ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 2, pp. 423-431, February 2024
DOI: https://doi.org/10.17507/tpls.1402.13

State-of-the-Art Review of the Corpus Linguistics Field From the Beginning Until the Development of ChatGPT

Yaser M. Altameemi*
Department of English, University of Ha'il, Ha'il, Kingdom of Saudi Arabia

Abstract—**The present paper highlights the recent state of and development in the corpus linguistics (CL) field. Although several reviews have been conducted on CL, these reviews have focused on specific areas, such as education, or did not provide an overall clear overview of the future implications of the field (Baker et al., 2008; Biber & Reppen, 2020; Biber et al., 1998; G. N. Leech, 1991; Mcenery et al., 2019; McEnery & Hardie, 2012). The author begins this paper with providing an overview that can guide new researchers in this field as well as postgraduates who require a general historical and thematic map of CL. The general overview discusses the publications of scholars who have participated in this field as well as the central tools that have been applied in CL. For specific details regarding the development of the field, the author analysed 217 articles from the 3 highest-impact factor journals according to the Web of Science over the last four years (2019–2022). The findings reveal a rapid development of the field in terms of practical and methodological perspectives, specifically regarding the investigations of language uses in different contexts. Thus, this paper indicates a significantly strong correlation between CL and technological development, such as natural language processing (NLP), and how this approach could fill the research gap of utilising CL in other areas of linguistics.**

Index Terms—**corpus linguistics, ChatGPT, systematic review, natural langue processing, corpus linguistics journals**

I. Introduction

Corpus linguistics (CL) is a field of linguistics that involves the analysis of large collections of texts which are known as corpora. Over the last few decades, there has been an increased use of CL in linguistic research to facilitate the collection and analysis of large amounts of texts. A key advantage in CL is that it studies broader representative samples and linguistic instances than the traditional method—that is, manual analysis of texts). CL enables analysts to notice findings of language that are not apparent in the analysis of small-scale discourse. Further, CL assists linguists in investigating changes in language patterns over different phases of time or regions. These perspectives could enable researchers to analyse the use of language in various contexts using authentic samples of language use. However, CL has a few limitations. Two main challenges that CL faces will be introduced. The first one is defining the appropriate corpora for analysis. The second is that CL requires specialised software and researchers should be aware of utilising and processing the data using a CL software tool. These key advantages and disadvantages are the central aspects of discussion in this paper.

The current paper aims to review the field of CL by exploring its methodological developments and current applications. I first present an overview of the field, including scholars in the field, areas of focus (e.g., English grammar and vocabulary), and salient tools. Then, the methodology adopted in this article will be provided. Thereafter, I discuss the analysed articles and the development in the field over the last four years (2019–2022). To conclude, I discuss a few challenges and limitations that CL faces and recommend future directions for research in this field.

II. Background of CL

Numerous studies have provided practical, theoretical, and methodological viewpoints of investigations as well as discussed tools that have been used in CL (Ali et al., 2011; Baker et al., 2008; Biber & Reppen, 2020; Biber et al., 1998; G. N. Leech, 1991; Mcenery et al., 2019; McEnery & Hardie, 2012; Nartey & Mwinlaaru, 2019; Nurdiyani & Nadra, 2021). However, CL has become messy and connected to other fields such as computer science without robust connection between the basis of the various fields. In this section, I highlight the major developments of CL by presenting the linguistic areas and related figures in CL. Then, I briefly discuss the salient tools that were employed by several studies and the features of these tools.

A. Overview of Linguistics Areas and Related Figures in CL

* Corresponding Author.

Various areas of linguistics have extensively employed CL, such as discourse analysis, computational linguistics, language teaching and learning, pragmatics, and sociolinguistics. These linguistic areas have developed the use of CL with the following considerations. First, the creation of large corpora like the British National Corpus (BNC), which contains millions of words. This type of corpora has enabled the study of different uses of language in unprecedented linguistic features. Second, advances in CL tools (e.g. collocation analysis tools) have enhanced the efficiency of processes and analyses of a large volume of linguistic data. Third, interdisciplinary collaboration between CL and other fields, such as natural language processing (NLP), has led to meaningful insights into language use. Fourth, methodological innovations through the development of new methods of analysing linguistic data, such as corpus-based discourse analysis, keyword analysis, and multi-dimensional analysis. Finally, CL has been applied to practical problems, such as common and frequent issues that relate to language learning and machine translation. Thus, CL has enabled researchers to explore new ways of investigating language uses.

Table 1 below presents the prominent scholars in the field of CL. The table presents the name of scholar, the area in which he/she focused on, and his/her area of interest. The last column includes the well-known published works of the scholars.

TABLE 1
PROMINENT SCHOLARS IN THE FIELD OF CL[1]

Name of scholar	Area of focus	Salient works
John Sinclair	Considered as the founder of the field	(Sinclair, 1991; McH. & Sinclair, 1975)
Michael Halliday	Uses CL to analyse the use of language in context (i.e. his work on systemic functional linguistics)	(Halliday, 1985a, 1985b; Halliday & Hasan, 1976; M. Halliday & Hasan, 1989)
Douglas Biber	Contributes to analysing differences in the use of language across various registers and genres	(Biber, 1993, 1991; Biber & Reppen, 2020; Biber et al., 1998)
Geoffrey Leech	Corpus-based approach to English grammar and vocabulary	(Leech, 2009; Leech, 1991; Leech & Paul, 2014)
Susan Conrad	Uses CL in language teaching	(Biber et al., 1998; Conrad, 2005)
Tony McEnery	Analyses large dataset by considering a wide range of topics (e.g. discourse analysis and computational methods).	(Baker et al., 2008; McEnery, 2019; McEnery et al., 2006; Mcenery et al., 2019; McEnery & Hardie, 2012)
Mark Davies	Creates corpora such as the Corpus of Contemporary American English (COCA) and the Corpus of Historical American English (COHA)	(Davies, 2010, 2012)
Antoinette Renouf	Different topics in CL (e.g. discourse analysis, genre studies, and multimodal communication).	(Baker & Renouf, 2005; Renouf & Sinclair, 1991)
Paul Baker	Gender and sexuality in language use	(Baker, 2006a, 2006b; Baker et al., 2008; McEnery et al., 2019)
Stefan Th. Gries	Topics and methods in CL, such as collocation analysis and construction grammar	(Gries, 2003, 2016, 2013)

Table 1 above indicates that the scholars have considered practical methods and applications that can facilitate the efficiency of analysing linguistic patterns as well as theoretical bases.

Thus, it is indicated that CL has been utilised to fix issues faced by linguists who employ traditional methods in various areas as mentioned in the second column. One of the salient fields in which CL has been applied in descriptive linguistics is in the analysis of a large corpus of texts. Analysing a large corpus of texts enables linguists to observe and identify patterns in language use and structure. Second, CL has been applied in language teaching, particularly materials that are used to develop a corpus for teaching English as a second/foreign language (ESL/EFL). Learners are aided in identifying useful and common patterns of language to familiarize them with the most frequent linguistic patterns in a specific genre. Another area is lexicography. Lexicographers have applied CL to analyse a large number of texts to identify uses of words, their meanings, and changes in the use of words based on language. Discourse analysis has also employed CL to investigate meaning in a representative sample of texts with different contexts. The last remarkable field is computational linguistics that examines how computer software deals with languages. By applying CL, computational linguists aim to enhance the efficiency of the programmes to recognize linguistic patterns and language use through the development of algorithms. Taking these general areas into account, CL varies from descriptive studies to computational modelling, and this has enhanced the approaches utilised in linguistics, specifically those related to

[1] Please note that the related texts are referenced as in the text citation, and the complete references are included in the references section.

language use.

B. *Review of the Salient Corpus Tools*

An important aspect that CL should consider is focusing on which tool facilitates answering the research question(s). In this section, I provide a general overview of CL software tools. Table 2 presents the general features of salient CL software tools as well as the common advantages and disadvantages of these tools with their extra resources (if more information is required).

TABLE 2
OVERVIEW OF SALIENT CL SOFTWARE TOOLS WITH THEIR FEATURES[2]

CTool	Advantages	Disadvantages	Related references
AntConc	- User-friendly interface that has clear functions for users. - It is a free tool. - Open source - Multiple functions (e.g. analysis of word frequency, collocation, and cluster analysis). - Fast and efficient tool processing.	- Only processes plain text files. - No specific technical support is provided. - Not compatible with other CL programmes.	- Smith (2021) - Anthony and Young-Scholten (2014)
Sketch Engine	- Valuable resource, as it has over four billion words - It is user-friendly. - Queries can be customised by using parameters such as frequency, lemma, and parts of speech. - It has multiple languages. - It has advanced tools, such as concordance analysis, keywords, and collocation analysis.	- It is a paid service. - Slow loading in the process. - Types of queries are limited.	1. Kilgarriff et al. (2004) 2. McEnery and Wilson (2017)
WordSmith	- Easy to navigate and use the software. - Powerful tool that quickly operates processes. - Multiple tools (e.g., collocation and keyword lists). - It supports various languages.	- Paid and expensive tool. - It has almost the same functions that are available in free tools. - As it is a paid tool, there are a few issues reported by users regarding customer support.	1. Golinkoff and Kathy (2006) 2. Reppen (2001) 3. Scott (2001)
Voyant	- User-friendly and enables text visualisation. - It allows sharing the work and projects with others in real-time. - It is a web-based tool and no need for installation.	- It has limited features in comparison to other tools (e.g., LancsBox). - Limited data size.	1. Graham and Milligan (2015) 2. Terras (2017)
LancsBox	- Multiple tools (e.g. collocation and keyword lists). - It visualizes the findings that enable finding the relationship between collocates through collocation network. - It supports advanced queries according to the purpose of the analysis, such as Delta P. - It supports multiple languages.	- Limited size of corpus up to fifty million words. - As it is a free source, it has limited technical support.	- Brezina et al. (2015) - Gablasova et al. (2017)

The above table presents the general differences between various well-known CL tools. All the tools have been developed over time and have various versions. Numerous corpus tools require subscription and are expensive, and this has been an obstacle for many researchers. For example, as a feature, Sketch Engine attempts to provide analysts with the data that already existed in the tool to enable analysts access the required data for research with multi-languages. Nowadays, it is evident that the tools do not focus much on providing data, as many types of data are available using different sources, such as downloading data from social media. In contrast, the features of the tools are focused upon more than actually building the data these days. For example, certain tools focus on simplicity, such as AntConc which provides a user-friendly interface; his tool can be used by both beginners and professionals. Other tools such as Voyant and LancsBox do not only work on the efficiency of processing analytical tools, such as collocation, but they also provide a nice visualization for the findings to facilitate the mapping of the relationship between collocates/concepts.

III. STEPS TO ANALYSE THE LATEST DEVELOPMENTS IN CL JOURNALS

This paper provides a detailed review of the recent development in the field by examining recently published articles in the field of CL. Various journals specialize in CL.

First, we review articles in the highest impact factor journals that are indexed in Web of Science and Scopus. These two indexes classify the journals according to most popular articles by considering the citation and the impact factor. The two indexes have been selected in this research, as numerous studies use them as central sources for reviewing

[2] Please note that the related texts are referenced as in the text citation, and the complete references are included in the references section.

high-impact articles in a specific field (e.g. Ali et al., 2011; Alkhalil et al., 2021; Ma & Mei, 2021; Nurdiyani & Nadra, 2021). For this study, the researcher selected the following high-impact journals: Corpus Linguistics and Linguistics Theory, International Journal of Corpus Linguistics, and Journal of Corpus Pragmatics. These journals are highly indexed in major databases, such as Web of Science and Scopus. All these journals are peer-reviewed academic journals.

Corpus Linguistics and Linguistics Theory has an impact factor of 2.3. This journal publishes articles that discuss theoretical issues, methodological issues, and applications. The International Journal of Corpus Linguistics encompasses various domains such as pragmatics, sociolinguistics, syntax, and discourse analysis. The impact factor of this journal is 1.5. The Journal of Corpus Pragmatics is more specific than the first two journals with regard to the application of CL. It focuses on researching issues in pragmatics, such as discourse markers, speech acts, implicatures, and politeness strategies. It has an impact factor of 1.3. The contributions of articles in this journal are centralized around the importance of considering the various contexts in which linguistic patterns are utilised.

The researcher applied the following steps to analyse the articles in the three journals:

1. All the articles that were published in the period 2019–2022 were analysed.
2. All the types of information were entered manually into an excel sheet. Two-hundred-and-seventeen articles were analysed from the three journals.
3. Articles were ranked chronologically according to the volumes and issues.
4. The researcher considered the following central points for each article: name of the journal, publication date, volume, issue, author(s), abstract, keywords.
5. The articles have been categorized according to the applied methodology, tools of the analysis, area of research, research question(s), and main findings.

IV. Reviewing the Articles

The analysed articles were generally classified into two main areas. First, numerous studies investigate the language used in different contexts, such as in political discourse, academic writing, education, or media communication. This type of studies often investigates patterns of grammatical structures, vocabulary use, and discourse markers within specific contexts. Other studies focus on methodological issues, including building a corpus, processing the data, and analysis of the data (e.g., Egbert et al., 2020; Larsson et al., 2020). This type of studies focuses on building corpora, annotating corpora, and establishing/developing software tools. In this section, I classify the articles into 10 additional categories based on a conductive approach. Table 3 presents the themes and areas of the overall classification across the three journals.

Table 3
Themes of Articles

Row Labels	Article count
CL and Discourse Studies	59
CL and Pure linguistics	49
CL and Education	28
CL and Methodological Perspectives	24
CL and Language Use in Context	21
CL and Methodological perspectives and Discourse Studies	15
CL, Pragmatics, and Discourse Studies	10
CL and Translation	5
CL, Cognitive Linguistics, and Discourse Studies	4
CL and Phonological Studies	2
Grand Total	217

In the table above, it is evident that the areas which were most commonly represented by the articles include CL and Discourse Studies and CL and Pure Linguistics. The analysis of the articles suggests that a central overarching area utilised by the journals is CL and discourse. This fundamental coexistence between CL and Discourse Studies can be linked to the number of scholars in CL who specialize in discourse studies. Moreover, this reason might be the same for the second most frequent area, which is CL and pure linguistics (e.g., Hundt et al., 2021; Sanders et al., 2021; Schneider, 2022). CL and Discourse Studies have a positive correlation as they share the idea of investigating the various uses of language in different occasions as well as analysing representative data that enable analysts to better discuss a social phenomenon.

Further, the articles in the CL and Pure Linguistics theme discuss the use of language, but with a focus on a linguistic structure such as syntax. The third most commonly discussed area is CL and Education. In this area, the articles apply CL to investigate issues that are specifically related to language learning. The articles in this area investigate linguistic errors of learners and consider the possible causes of these errors. Other articles in this area investigate the application of CL in the educational process and how this tool might be useful for learning a foreign language. For example, CL has been applied to help learners to master the target language through the phrases and collocations in the corpora.

The fourth most common theme is CL and Methodological Perspectives in which the articles contribute to fill in gaps that are related to the steps and methods of applying CL. A very close area to this one is CL and Methodological Perspectives and Discourse Studies (i.e., the sixth most common theme). However, the articles in this theme aim to achieve both practical and methodological contributions. In the fifth most common theme, we have language use in context, with 21 articles. This theme has a connection to the seventh theme of Pragmatics and Discourse Studies (10 articles). The articles in the fifth most common theme include topics of language use in context that relate more to discourse studies than being specific to pragmatics.

The last three domains are Translation, Cognitive Studies (Mehl, 2021), and CL and Phonological Studies (Brand & Ernestus, 2021). Translation has developed from the perspective of machine translation, but few articles have been published in the three leading corpus journals. A possible reason for this is that the field of translation has its journals that focus on translation. With regard to cognitive and phonological studies, it might be difficult to interpret these findings in relation to the various contexts. Therefore, this issue is debatable in the field with regard to the extent to which numbers are used as facts and can lead to interpreting abstract phenomena (e.g., emotions and motivations). The overall distribution of the themes is presented in Figure 1 below to present the overall picture.

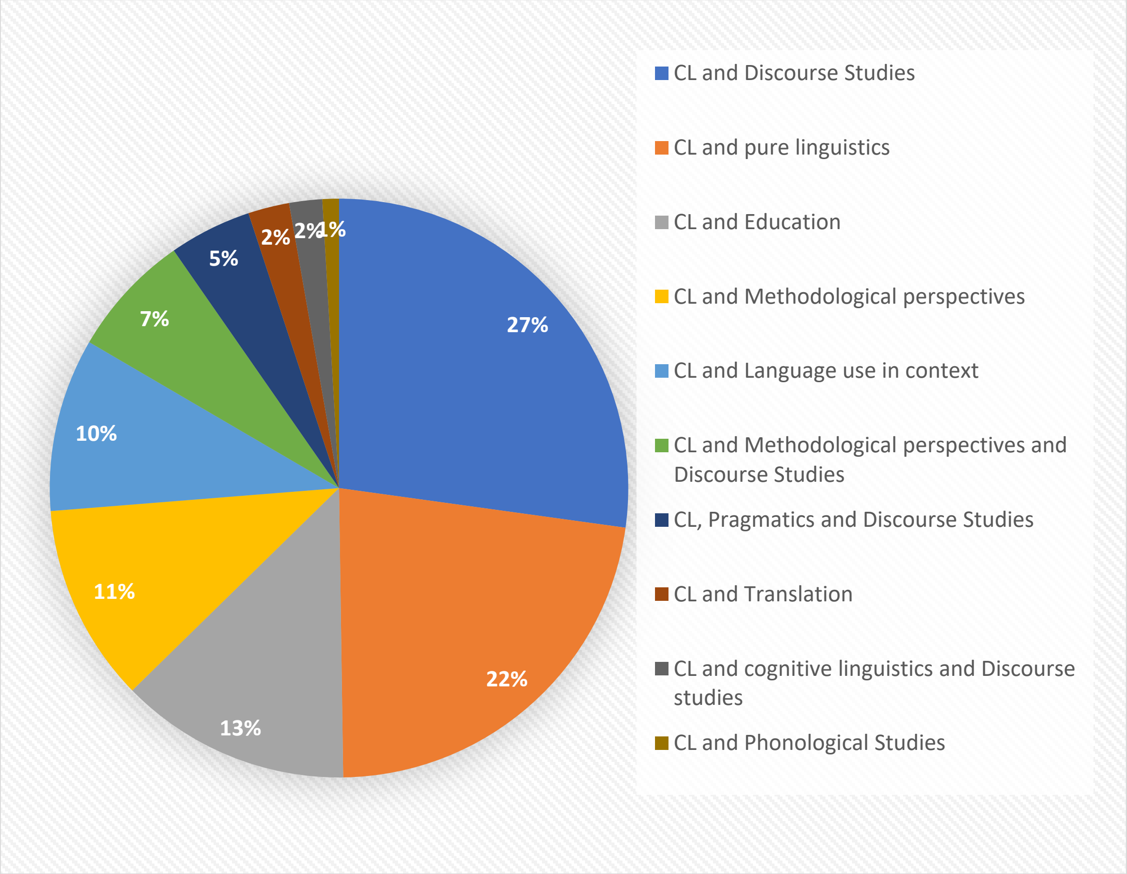

Figure 1. Distribution of Articles Based on the Areas of CL

Table 4 is a more detailed table than the previous one, as it presents the distribution of areas according to the selected journals.

TABLE 4
THE DISTRIBUTION OF THEMES ACCORDING TO THE SELECTED JOURNALS

Row Labels	**Count of Areas and their themes**
CL and Discourse Studies	**59**
Corpus Linguistics and Linguistic Theory	9
Corpus Pragmatics	27
International Journal of Corpus Linguistics	23
CL and Pure Linguistics	**49**
Corpus Linguistics and Linguistic Theory	30
Corpus Pragmatics	5
International Journal of Corpus Linguistics	14
CL and Education	**28**
Corpus Linguistics and Linguistic Theory	8
Corpus Pragmatics	12
International Journal of Corpus Linguistics	8
CL and Methodological perspectives	**24**
Corpus Linguistics and Linguistic Theory	8
International Journal of Corpus Linguistics	16
CL and Language Use in Context	**21**
Corpus Linguistics and Linguistic Theory	10
Corpus Pragmatics	4
International Journal of Corpus Linguistics	7
CL and Methodological Perspectives and Discourse Studies	**15**
Corpus Linguistics and Linguistic Theory	10
Corpus Pragmatics	1
International Journal of Corpus Linguistics	4
CL, Pragmatics and Discourse Studies	**10**
Corpus Pragmatics	10
CL and Translation	**5**
Corpus Linguistics and Linguistic Theory	1
Corpus Pragmatics	2
International Journal of Corpus Linguistics	2
CL and Cognitive Linguistics and Discourse Studies	**4**
Corpus Linguistics and Linguistic Theory	1
Corpus Pragmatics	3
CL and Phonological Studies	**2**
Corpus Linguistics and Linguistic Theory	2
Grand Total	**217**

The above table presents the distribution of the three investigated journals in terms of the areas that emerged in the analysis of the articles. In general, most of the articles in the International Journal of Corpus Linguistics investigate both practical aspects of language use and methodological aspects in the field. In contrast, the articles in Corpus Linguistics and Linguistic Theory focus mainly on investigating theoretical linguistics and how the applications of CL enhance the efficiency of analysing linguistic patterns. Lastly, the articles in Corpus Pragmatics investigate the uses of language in different contexts (Hanks & Egbert, 2022), although he focus of Corpus Pragmatics is on how to integrate pragmatics using computer-assisted methods (Groom, 2019; Yaylali, 2020).

Based on the analysis of the articles, it is evident that applying CL is centralized on decreasing subjectivity and bias, representativeness of the data that enable analysts to make judgments, contextualization, and levels of contexts. For example, numerous studies in the cognitive area apply manual analyses for the sake of having a fuller understanding of a socio-linguistic phenomenon. Other studies may require less context in the analysis, such as the smaller grammatical elements of linguistic patterns like spelling errors and grammar mistakes, as well as the learning of a second/foreign language. These are considered to focus mainly on the writing skills of students and the use of collocation. Overall, both practical findings and methodological perspectives are important areas in CL. While certain studies may prioritize one over the other, many researchers strive to strike a balance between exploring new insights into language use and advancing the research methods in the field.

V. THE FUTURE OF CORPUS LINGUISTICS WITH THE EXISTENCE OF CHATGPT

Since CL is dependent on the application of technological tools, it is important to discuss how the rapid development of technology influences the field of CL. Chat Generative Pre-trained Transformer (ChatGPT) is a search engine that utilises NLP. ChatGPT was selected for this research because it shares the notion with CL of focusing on applying the practical use of language in various contexts. Further, ChatGPT is more advanced than other CL tools by utilizing features such as summarizing large texts and thematic analysis of corpora (Altameemi & Altamimi, 2023). These features can be applied by any user by asking the tool in the chat box without much effort. However, these features still need development in ChatGPT, specifically with regard to the accuracy of the results. In this section, I highlight the importance of applying such ChatGPT in linguistics. Then, I discuss the potential methods of being up to date with NLP.

First, I examine the significance of the collaboration between CL and NLP, specifically with the rapid development of technology and the services it can provide for linguists. Although there might be risks of applying ChatGPT without knowing how it processes language, many linguists need to change the manner in which they think of applying technology. In other words, instead of being cautious in applying ChatGPT, linguists should examine the importance of merging CL and ChatGPT. Even linguistic academic programmes should consider the importance of applying technology in the study plan of their degrees. Moreover, it is not only linguists who must take this point into account, but scholars in other fields who should consider these aspects and think about the effective utilisation of technology in studying fields of human knowledge.

The second aspect that needs to be mentioned is how linguists benefit from NLP such as ChatGPT. This is a critical question that may expand the field of research. However, I the following possible practices that may assist linguists in benefitting from this technology. One, linguists may investigate the linguistic structure of ChatGPT and they may find issues in the language use; simultaneously, they may collaborate with developers to work on these issues. They may look at the reasons that allowed ChatGPT to produce such linguistic structures and how these structures could be further developed. Further, ChatGPT might be developed to participate in providing data and larger corpora than the specialized corpora that were manually built. This aspect would also decrease the bias of the selected corpus, as ChatGPT may help in the automatic building and annotating of a corpus. Moreover, ChatGPT might be developed from a theoretical linguistic structure to analyse, classify, and discuss the salient linguistic findings of a corpus/corpora. For example, ChatGPT now is able to edit and find errors in a written text, and in the future, it might be able to identify the reasons for errors and provide steps to overcome issues in writing.

Overall, CL has proven to be a significant area of studying various language uses. As knowledge has been distributed in different means and huge data has become available to researchers, it is likely that CL will develop and have more sophisticated uses in linguistic research. Simultaneously, the changes in the field of CL should be considered with the rapid development of ChatGPT.

VI. Conclusion

In conclusion, the field of CL has revolutionised the methods utilised in investigating language and the insightful results of linguistic phenomena. By analysing large collections of data, corpus linguists are able to uncover linguistic findings that were previously inaccessible through traditional methods. A major strength of applying CL is providing empirical evidence for linguistic theories through real-life examples. The findings of this approach have challenged a few traditional assumptions and spotlighted unexplored areas of language use. However, CL also faces certain challenges, such as building and maintaining large corpora, representativeness, and bias. Looking ahead, CL is likely to continue developing as technology advances further. With the evolvement of more sophisticated linguistic tools, researchers can delve even deeper into new research questions to investigate unexplored linguistic phenomena. Additionally, the integration of CL with ChatGPT holds great potential for the understanding of language in the digital age. In conclusion, the development of CL will continue to enhance our understanding of language uses in real-life contexts as well as the manner in which we understand human communication.

Acknowledgements

The author wishes to thank the University of Ha’il.

References

[1] Ali, R., Khan, M. A., Ahmad, I., & Ahmad, Z. (2011). *A state-of-the-art review of corpus linguistics journals*, *13*(1), 1-22.

[2] Alkhalil, A., Abdallah, M. A. E., Alogali, A., & Aljaloud, A. (2021). Applying big data analytics in higher education: A systematic mapping study. *International Journal of Information and Communication Technology Education*, *17*(3), 29–51. https://doi.org/10.4018/IJICTE.20210701.oa3

[3] Altameemi, Y., & Altamimi, M. (2023). Thematic Analysis: A Corpus-Based Method for Understanding Themes/Topics of a Corpus through a Classification Process Using Long Short-Term Memory (LSTM). *Applied Sciences*, *13*(5), 3308, 1-12.

[4] Anthony, L., & Young-Scholten, M. (2014). The benefits and limitations of using corpus linguistic techniques in authorship attribution studies. *Journal of English Linguistics*, *42*(2), 119–141.

[5] Baker, P. (2006a). *Glossary of corpus linguistics*. Edinburgh University Press.

[6] Baker, P. (2006b). *Using corpora in discourse analysis*. Bloomsbury Academic.

[7] Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, *4*(2), 197–226.

[8] Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306. https://doi.org/10.1177/0957926508088962

[9] Baker, P., & Egbert, J. (2018). Introduction: The emergence of corpus pragmatics. *Journal of Corpus Pragmatics*, *2*(1), 1-6.

[10] Biber, D. (1993a). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

[11] Biber, D., & Reppen, R. (2020). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

[12] Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

[13] Biber, D., Susan, C., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

[14] Brand, S., & Ernestus, M. (2021). Reduction of word-final obstruent-liquid-schwa clusters in Parisian French. *Corpus Linguistics and Linguistic Theory*, *17*(1), 249-285.

[15] Brezina, V., & Flowerdew, L. (Eds.) (2019). Learner corpus research: New perspectives and applications. *International Journal of Corpus Linguistics*. Bloomsbury Publishing.

[16] Brezina, V., Mcenery, T., & Wattam, S. (2015). Collocations in context a new perspective on collocation networks*. *International Journal of Corpus Linguistics*, *202*(2015), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

[17] Brezina, V., & Meyerhoff, M. (2019). Corpora and discourse: Integrating pragmatics into linguistic analysis with computer-assisted methods. *Journal of Pragmatics*, *145*, 1–7.

[18] Conrad, S. (2005). Corpus linguistics and L2 teaching. In *Handbook of research in second language teaching and learning* (pp. 393–409). Routledge.

[19] Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *4*(25), 447–464.

[20] Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, *2*(7), 121–157.

[21] Egbert, J., Tove, L., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Pres.

[22] Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*(June), 155–179. https://doi.org/10.1111/lang.12225

[23] Golinkoff, R. Michnick., & Kathy, H.-P. (2006). Baby wordsmith: From associationist to social sophisticate. *Current Directions in Psychological Science*, *1*(15), 30-33.

[24] Graham, S., & Milligan, I. (2015). Getting started with text mining using voyant tools. In *the historian's macroscope: Big digital history* (pp. 293–308). Imperial College Press.

[25] Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. A&C Black.

[26] Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Taylor & Francis.

[27] Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

[28] Groom, C., & Charles, M. (2018). Corpus pragmatics: A decade of research. *Journal of Pragmatics*, *130*, 1–9.

[29] Groom, N. (2019). Construction grammar and the corpus-based analysis of discourses the case of the way-in-which construction. *International Journal of Corpus Linguistics*, *24*(3), 291-323.

[30] Halliday, M. A. K. (1985a). *An introduction to functional grammar* (Second, 19). Arnold.

[31] Halliday, M. A. K. (1985b). *An introduction to functional grammar* (Second, 19). Arnold.

[32] Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

[33] Halliday, M., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective* (Second). Oxford University Press.

[34] Hanks, E., & Egbert, J. (2022). The interplay of laughter and communicative purpose in conversational discourse: A corpus-based study of British English. *Corpus Pragmatics*, *6*(4), 261-290.

[35] Larsson, T., Plonsky, L., & Hancock, G. R. (2020). On the benefits of structural equation modelling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, *3*(17), 683–714.

[36] Hundt, M., R öthlisberger, M., & Seoane, E. (2021). Predicting voice alternation across academic Englishes. *Corpus Linguistics and Linguistic Theory*, *17*(1), 189-222.

[37] Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Asr, F. T., Zufferey, S., & Evers-Vermeul, J. (2021). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, *17*(1), 1-71.

[38] Schneider, K. P. (2022). Referring to Speech Acts in Communication: Exploring Meta-Illocutionary Expressions in ICE-Ireland. *Corpus Pragmatics*, *6*(2), 155-174.

[39] Leech, G. (2009). *Change in contemporary English: A grammatical study*. Cambridge University Press.

[40] Leech, G. N. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 8–29). Longman.

[41] Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

[42] Ma, Q., & Mei, F. (2021). Review of corpus tools for vocabulary teaching and learning. *Journal of China Computer-Assisted Language Learning*, *1*(1), 177–190. https://doi.org/10.1515/jccall-2021-2008

[43] McEnery, T. (2019). *Corpus linguistics*. Edinburgh University Press.

[44] McEnery, T., Brezina, V., & Baker, H. (2019). Usage fluctuation analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, *24*(4), 413–444. https://doi.org/10.1075/IJCL.18096.MCE

[45] Mcenery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyse language use. *Annual Review of Applied Linguistics*, *39*, 74–92. https://doi.org/10.1017/S0267190519000096

[46] McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory, and practice*. Cambridge University Press.

[47] McEnery, T., Richard, X., & Yukio, T. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

[48] Nartey, M., & Mwinlaaru, I. N. (2019). Towards a decade of synergising corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora, 14*(2), 203–235. https://doi.org/10.3366/cor.2019.0169

[49] Nurdiyani, N., & Nadra, N. (2021). Review of corpus linguistics for education: A guide for research. *Corpus Pragmatics, 5*(4), 543–547. https://doi.org/10.1007/s41701-021-00111-6

[50] Reppen, R. (2001). Review of MonoConc Pro and WordSmith Tools. *Language Learning & Technology, 5*(3), 32-36.

[51] Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. *Small Corpus Studies and ELT*, 47–67.

[52] Smith, E. L. (2021). AntConc (Version 3.5. 8)/WordSmith Tools (Version 8). *Early Modern Digital Review*, *4*(1), 200-214.

[53] Mehl, S. (2021). What we talk about when we talk about corpus frequency: The example of polysemous verbs with light and concrete senses. *Corpus Linguistics and Linguistic Theory*, *17*(1), 223-247.

[54] Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

[55] Sinclair, J., & McH., C. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.

[56] Terras, M. (2017). Text mining with Voyant tools: A review. *Journal of Digital Humanities*, *6*(1), 1–10.

[57] Yaylali, A. (2020). Brezina, V., & Flowerdew, L. (Eds.). (2019). *Learner Corpus Research: New Perspectives and Applications*. Bloomsbury Publishing.

Yaser Mohammed Altameemi is Assistant Professor of Applied Linguistics at the department of English. He is interested in Critical Discourse Analysis, Discourse Studies, and Corpus linguistics. He is currently interested in multidisciplinary research to fill gaps and practical issues such as automating the analyses of language uses in society. He gained his PhD from Cardiff University in Language and Communication. After his completion, he rejoined the University of Ha'il as an Assistant Professor in the English Department.

He has been nominated for various positions such as the Vice-Dean of Quality and Development for Performance Measurement and Institutional Excellence- Director of E-Learning Unit at the college of Arts- Executive Member in the Strategic Office of the University. Also, he has been the manager of different projects such as the manager of international accreditation for the academic programs of the university- and the manager of developing and establishing academic programs.

Dr. Altameemi, Yaser was a member of the Advisory Committee of Applied Linguistics program at Prince Nourah University.