

Exploring corpus linguistics via computational tool analysis: key finding review

Wan Nur Aida Sakinah Wan Jusoh¹, Norfaizah Abdul Jobar²,
Md Zahril Nizam Md Yusoff², Hanifah Mahat³

¹Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris, Tanjong Malim, Malaysia

²Department of Language and Malay Literature, Faculty of Languages and Communication,
Universiti Pendidikan Sultan Idris, Tanjong Malim, Malaysia

³Department of Geography and Environment, Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Tanjong Malim, Malaysia

Article Info

Article history:

Received Nov 2, 2023

Revised Jan 22, 2024

Accepted Feb 16, 2024

Keywords:

Computational linguistic

Corpus analysis

Corpus linguistic

Text corpora

Tool support

ABSTRACT

Corpus linguistics investigates language using extensive text databases. Tools assist researchers in analyzing, extracting, and interpreting linguistic information efficiently. Furthermore, if researchers only use traditional tools in corpus linguistic analysis, they will lack the comprehensiveness and efficiency required to effectively navigate and derive valuable insights from language data. This paper employed the preferred reporting items for systematic reviews and meta-analyses (PRISMA) approach to find the primary data based on a few keywords in corpus linguistic, corpus analysis, computational linguistic, text corpora and tool support. Based on this method, we used advanced searching techniques on Scopus and Web of Science (WoS) and discovered (N=28) data pertinent to the study. Expert scholars decide on a theme based on the problem, which is (i) types of corpus tools and their uses; (ii) their contributions and their capabilities, and (iii) limitations of corpus tools. All the tools were used in interdisciplinary studies. In summary, this systematic review uncovers pivotal key findings at the intersection of computational tools and corpus analysis, enriching linguistic knowledge. It highlights the interdisciplinary potential of corpus-based analysis in advancing linguistic tools and, their applications, as well as language analysis.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Wan Nur Aida Sakinah Wan Jusoh

Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris

35900 Tanjong Malim, Malaysia

Email: csakinah00@gmail.com

1. INTRODUCTION

In this dynamic landscape, synergy between human expertise and computational power has become a hallmark of contemporary language studies. Researchers increasingly rely on advanced computational linguistics, machine learning algorithms, and corpus-based online tools to efficiently shift through massive corpora and to identify patterns, frequencies, trends, and variations in language use. A corpus is a collection of texts compiled for a certain purpose [1], [2], while corpus analysis is a text analysis method that allows large-scale comparisons between textual objects. This enables the identification of patterns in the grammatical use of recurring phrases and statistically likely or unlikely phrases within a collection of documents. This analysis is useful for testing intuitions regarding texts, triangulating results from other digital methods, and facilitating the retrieval and interpretation of authentic examples of language phenomena. This involves an exploratory, inductive approach to studying the meanings and functions of

linguistic forms in the context of real communication, whereas computational tools are an interdisciplinary field that offers a computational viewpoint on natural language. It also automates linguistic operations that were formerly performed manually, including text analysis [3].

Currently, many corpora require some exploration, and this analysis offers a major contribution to the work of linguists, researchers, and educators [4] in comprehending the natural use of languages in certain communities. After several decades, it was noticed that several linguists and researchers employed corpus linguistics as a key methodology [5]. Regardless of the preceding definition, the concept of corpus linguistics as a theory or methodology has become a “polemic” among linguists [6]. One of the key reasons for the rapid growth of corpus linguistics is its ability to provide a vast amount of authentic language data for analysis [7] and to analyze data with multipurpose functions in one tool. However, the use of corpus tools involves several issues and problems because they do not fully meet the specific needs of researchers in terms of corpus size or diversity [8]. Several tools have limited access and restricted applied domains, which can hinder their usability and applicability in different research contexts [9], lack accuracy in both automatic and manual annotation, and are not user-friendly. This may affect the reliability and validity of corpus data. Additionally, the tool has not kept up with evolving research needs or integration between corpus linguistics methodologies and cutting-edge computational approaches that affect the final result, and they struggle to adapt to the demands of diverse language types and modalities [10]. Another problem is the lack of evaluation skills and knowledge of corpus tools, which can make it challenging for beginners to understand and utilize them effectively [11]. Furthermore, the development of corpus analysis platforms faces challenges in managing and analyzing large amounts of data while ensuring open accessibility and satisfying scientific demands [9].

For this reason, an analysis of the latest studies on the use of computational tools must be conducted to obtain the latest information on all these problems. Such problems will be examined based on i) types, ii) uses, iii) contributions of computational tools, and iv) limitations of computational tools that the researcher found from the systematic literature review. This review provides a structured overview and understanding of a specific phenomenon or topic [12]. The purpose of the systematic literature review is to create a better theoretically grounded understanding of the topic, identify trends, describe managerial implications, and identify future research directions [13].

Additionally, this study contributes to bridging the gap between traditional linguistic analyses and new computational approaches, fostering collaborative efforts to improve linguistic analyses, data accessibility, and other considerations. Ultimately, this study aimed to enhance the quality and impact of language research in diverse fields. In previous systematic literature reviews (SLRs) on calculation tools, the majority have focused on only one subtool. For example, a machine learning tool [14] but not a variety of corpora or computational tools. On the other hand, one study [15] focused on software support, but this research leans more towards discourse-based textual analysis, as well as a few sub-sections on the criteria for tools and their guidelines. Antidze *et al.* [16] focused on software for the composition of natural languages. Therefore, the motivation of this study was to more deeply investigate the computational tools used in corpus linguistics to identify the functions, advantages, and limitations of the tools used by past researchers based on statements from the articles reviewed.

2. METHOD

This section describes the selection of search terms and literature required for the next stage of the mapping phase. It is essential to analyze how researchers have conducted previous studies and their contributions to the field. This literature review summarizes previous research, but it is not a comprehensive analysis of all studies in this area. This section presents the preliminary findings from the SLR conducted to analyze the described domain, starting from the identification and screening phases and including data extraction and analysis, as shown in Figure 1, before the end of this section.

2.1. Identification

The process of selecting papers appropriate for this report involved three key phases as part of the systematic review process. Once the pertinent keywords have been determined, search strings are formulated for the Scopus and Web of Science (WoS) databases, as outlined in Table 1. During the initial stage of the systematic review procedure, we successfully obtained a total of 4,273 papers from multiple databases. For a manuscript to be published, all the work must meet the standards of publication quality.

2.2. Screening

Duplicate papers were eliminated during the initial screening. All papers underwent two screening stages, with the first stage resulting in the rejection of 4,125 papers. Consequently, 148 publications were eliminated based on specific criteria. Overall, the rigorous screening process ensured the elimination of duplicate papers and the selection of relevant publications to meet the objectives of the study.

Table 1. The search strings from Scopus and WoS

Name of database	The search string
Scopus	TITLE-ABS-KEY ("corpus linguistic" OR "corpus analysis" OR "computational linguistic" OR "text corpora" OR "text analysis" AND "tools support") AND PUBYEAR > 2020 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "SOCI")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (OA, "all"))
Web of Science	"corpus linguistic" OR "corpus analysis" OR "computational linguistic" OR "text corpora" OR "text analysis" AND "tools support" (All Fields) and 2023 or 2022 or 2021 (Publication Years) and Article (Document Types) and English (Languages) and Computer Science or Social Sciences Other Topics (Research Areas) and All Open Access (Open Access)

2.3. Eligibility

In the third phase of the research methodology, referred to as eligibility assessment, we meticulously examined the article titles and pertinent textual content to ascertain compliance with the inclusion criteria and alignment with the study objectives. A total of 144 participants were included in this study. This rigorous scrutiny led to the removal of 116 papers because their titles and abstracts did not demonstrate significant relevance to our study's objective and were based on empirical data as shown in Table 2. The selected articles met the necessary criteria and provided valuable insights and information that aligned with our research aims.

Table 2. The criterion for choosing is searching

Criterion	Inclusion	Exclusion
Language	English	Non-English
Timeline	2021-2023	< 2021
Literature type	Journal (article)	Conference, book, review
Subject area	Computer science, social sciences	Besides computer sciences / social sciences / others
Document type	Article	Besides article
Open access document	Open access	Besides open access

2.4. Data extraction and analysis

The authors then investigated the most important studies on corpus linguistic tool classification conducted thus far. The methods used in each study and research results were examined. Subsequently, the author collaborated with several researchers to derive thematic categories grounded in the empirical data within the context of this study. Figure 1 shows how the authors carefully reviewed 28 publications on claims or information relevant to the main topics of the study using a PRISMA diagram [17], [18]. Throughout the data analysis process, a log was maintained to document pertinent explanations, perspectives, enigmas, or other concepts deemed significant in comprehending the data. Finally, we conducted a comparative analysis of the findings to ascertain any potential issues pertaining to the construction of the theme. It is important to note that if the ideas did not match, writers would discuss them. The established themes were changed to ensure they were the same.

3. RESULTS AND DISCUSSION

The findings of this study include several main themes, namely, the types and uses of the corpus tools used in the 24 articles reviewed. These articles used word analysis tools, natural language processing (NLP) tools, and combinations of several NLP and technological tools. The selection of this corpus tool was based on the need to solve the study problem, along with the type of field and data used. This result was divided according to the types of corpus tools and their uses, contributions, capabilities, and limitations.

3.1. Types of computational tool and uses

Computational tools in corpus linguistics can be categorized into several types, based on their functions and applications. First, basic corpus tools such as AntConc [19] and keyword in context (KWIC) [20] allow users to search for word phrases and view every instance in its immediate context. They generated lists of words and phrases, frequency lists, and other visualization methods [21], [22]. The purpose of this software is to thoroughly analyze textual data using known corpus statistical standards. Examples include concordance tools, collocation graphs, network tools, description tools, wordlists, keyword tools, n-gram tools, text tools, GraphColl, and Wizard [23]. However, these tools cannot support language processing analysis; they are limited to the surface features of the text, such as concordance, word and phrase search, and word frequency analysis, and are still excellent for the detailed manual analysis of texts. Thus, for language processing analysis (NLP) is the alternative. It is more diverse and is used in a wider range of applications, from automated text summarization and sentiment analysis to language translation and speech recognition.

NLP tools often use complex algorithms and Microsoft Excel to process unstructured text data, allowing them to summarize large texts, extract key points and information, and improve over time through machine learning and artificial intelligence (AI) [24], [25], visualization data analysis [26], database management systems (DBMS) [27], [28] (e.g., Python’s NLTK or spaCy) [29], integrated development environments (IDEs), web scraping and data extraction tools, speech processing tools, and statistical analysis software to analyze and understand the text. These tools can discern patterns and meanings in a manner like that of human comprehension [30]. Among the most notable NLP tools are Wordify [8], Mlphon [31], morphological analyzer [32], Runyakitara tool [33], LexTutor [34], [35], Coh-Metric [36], [37] linguistic queries and word count (LIWC) [38], UAM corpus tool [39], [40], SketchEngine (SkE) [41], Wmatrix [37] MultiAzterTest [42], sublanguage corpus analysis toolkit (SubCAT) [43], EnvText [44], InLang [45], Berri corpus manager [46], UCREL semantic analysis system (USAS) [47], PyMongo (Mongo DB, Python technology, Flask) [46], LancsBox 5.1.2 [48], LancsBox 4.5 [49], LancsBox [50], NooJ platform [51] and Bi-LSTM [52].

These NLP tools encompass core areas, such as syntax, semantics, pragmatics, and discourse, providing a range of capabilities suitable for complex and dynamic applications. The combination of various theories and language models with NLP models has led to new discoveries in the field. These advancements have been observed in areas such as natural language tasks, language modelling, syntactic parsing, machine translation, sentiment analysis, and question answering [53]. The NLP tool provides a broader and more sophisticated range of capabilities suitable for a wider array of complex and dynamic applications in today’s data-driven world [54]. The types of corpus tools were reviewed on a one-on-one basis and their uses were explored in 28 articles as shown in Table 3. All tools were tagged by the article and found to be a corpus tool and a combination of tools from other supporting tools.

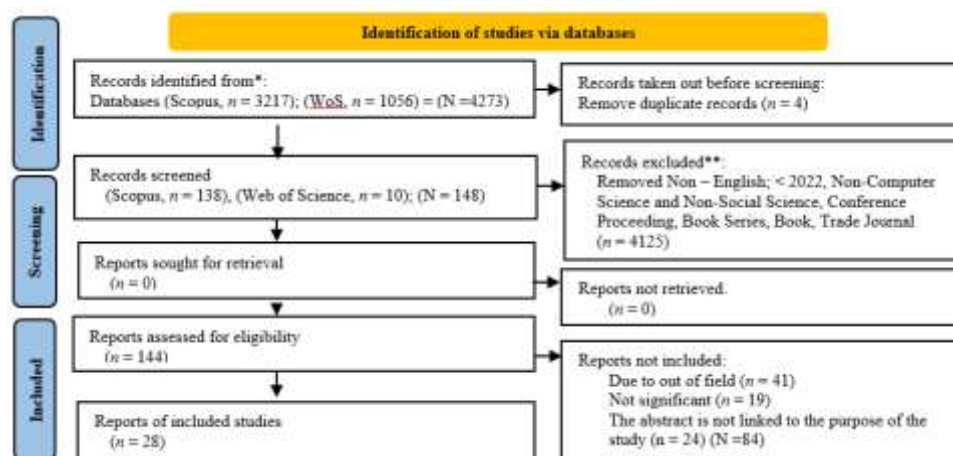


Figure 1. Flow chart of the recommended search study from the PRISMA diagram

Table 3. List of computational tools based on research

No.	Types of tools	No.	Types of tools
1	Wordify [8]	16	SubCAT [43]
2	TermoStat Web 3.0 and AntConc [19]	17	EnvText [44]
3	KWIC [20]	18	InLang and fMRI [45]
4	Mlphon [31]	19	Berri Corpus Manager using Mongo DB, Mongo DB, Python technology, Flask (PyMongo) [46]
5	Morphological analyzer [32]	20	USAS, Wmatrix, Sketch engine [47]
6	Runyakitara tools [33]	21, 22, 23	LancsBox 5.1.2 and iWeb corpus [48] LancsBox 4.5 [49] [50]
7		24	NooJ and POS tagger [51]
8	LexTutor [34]	25	Bi-LSTM [52]
9	LexTutor and SPSS [35],	26	Python 3.0, spaCy, Latent dirichlet allocation (LDA), pyLDAvis, Gensim [55]
10	Coh-Metric and Gramulator [36]	27	Optical character recognition (OCR), Granularity, ABBYY Finereader 14, CtexTools [56]
11	Coh-Metrix 3.0 and Wmatrix [37]	28	TF-IDF algorithm [57]
12	LIWC, Google perspective API tools [38]	29	pyLDAvis-LDA for topic Modeling™ & Python library spaCy [58]
13	UAM corpus tool [39] [40]	30	EcoLexicon and data driven learning (DDL) [59]
14	SketchEngine (SkE) [41]	31	TF-IDF algorithms and Python [60]
15	MultiAzterTest [42]		

Thirteen main functions were determined for the computational tools. All these uses comprise the main functions and key points. All tools, including concordance, NLP, and technology tools, were used by the researcher based on the articles found. Some studies have developed new corpus tools based on the current need to solve problems in a particular language such as Wordify [8], Mlphon [31], Runyakitara tool [33], PyMongo (Mongo DB, Python technology, Flask) [46], InLang [45], MultiAzterTest [42], PyLDAvis [58], and NooJ [51]. The types and functions of these computational tools are summarized in a mind map as shown in Figure 2.

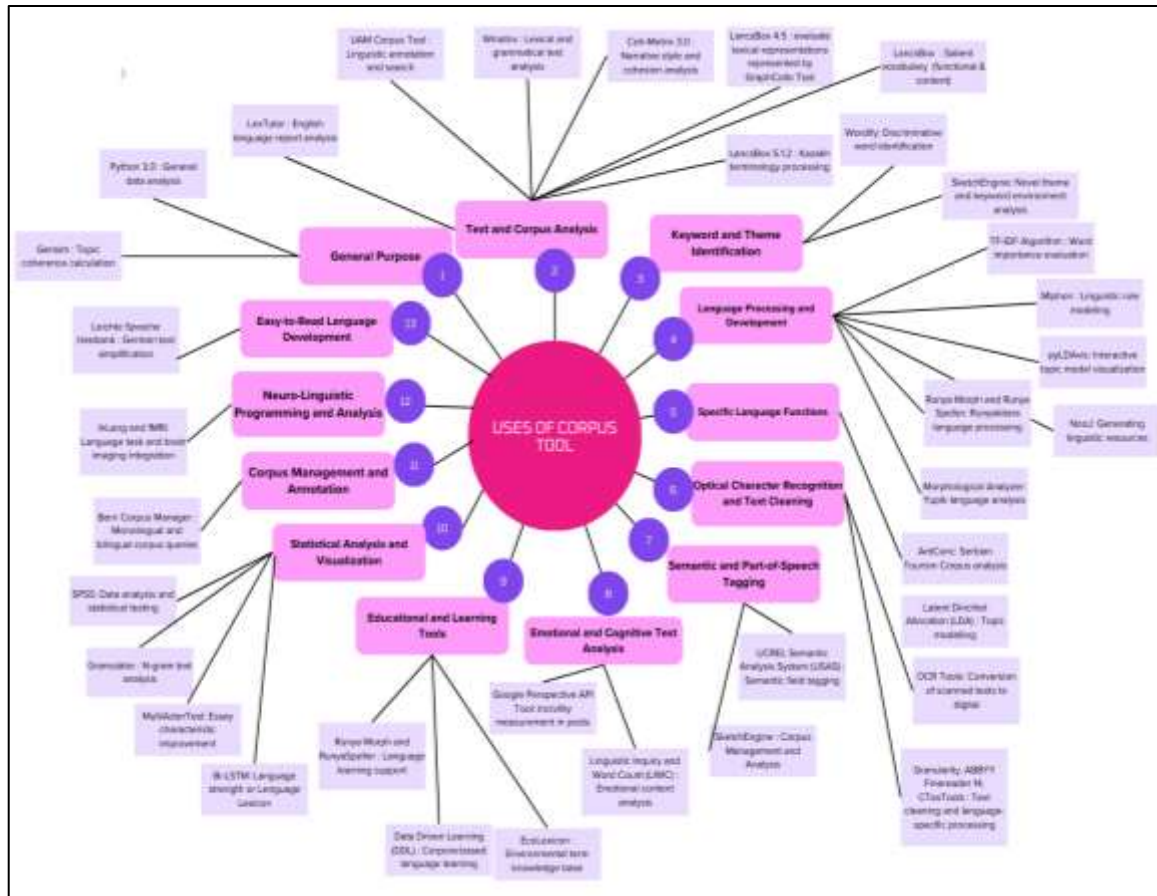


Figure 2. Overview of the use of computational tools based on selected articles

3.2. Contribution and capabilities of computational tools

This section discusses the role of each tool used in the study. Of the studies examined, 26 articulated the impact of the tool on research and its utility in data analysis. There are also articles that do not state the advantages or features of the corpus tools they use or justify their use compared to other tools. They simply state the type of corpus tool and apply it to obtain data. Corpus tools are often not explicitly described in this study. One reason is the lack of evaluation and knowledge of the tools, particularly for beginners [61]. In addition, researchers may overlook describing the contribution of corpus tools because they focus more on integrating the tools to obtain more important data than on describing the contribution of the tool [62]. Therefore, researchers should strive to strike a balance by discussing substantive contributions and providing sufficient methodological details, including explanations of corpus tool selection, functionality, and effectiveness in generating precise data that align with the research goals.

Of the 28 data articles in this previous study, 31 corpus tools were discussed. This shows that there are various contributions from the application of the corpus tools that they have chosen. However, the value of each tool is contingent when the specific aims involve the nature of the research questions and the characteristics of the data influence to ensure that knowledge is advanced in their respective fields [63].

The application of this corpus tool meets the needs of the study and has helped researchers analyze the language and improve their skills. These explanations are detailed in Figures 3 and 4, respectively. There are various benefits of all the corpus tool that have been expressed and benefited researchers, and this tool has supported their study data from various aspects e.g. in terms of linguistic competence and features [42], language functioning and patterns [33], digital corpora [56], lexical richness [35], flexible database, multidimensional measures of language quality [37], and multilingual analysis [50].

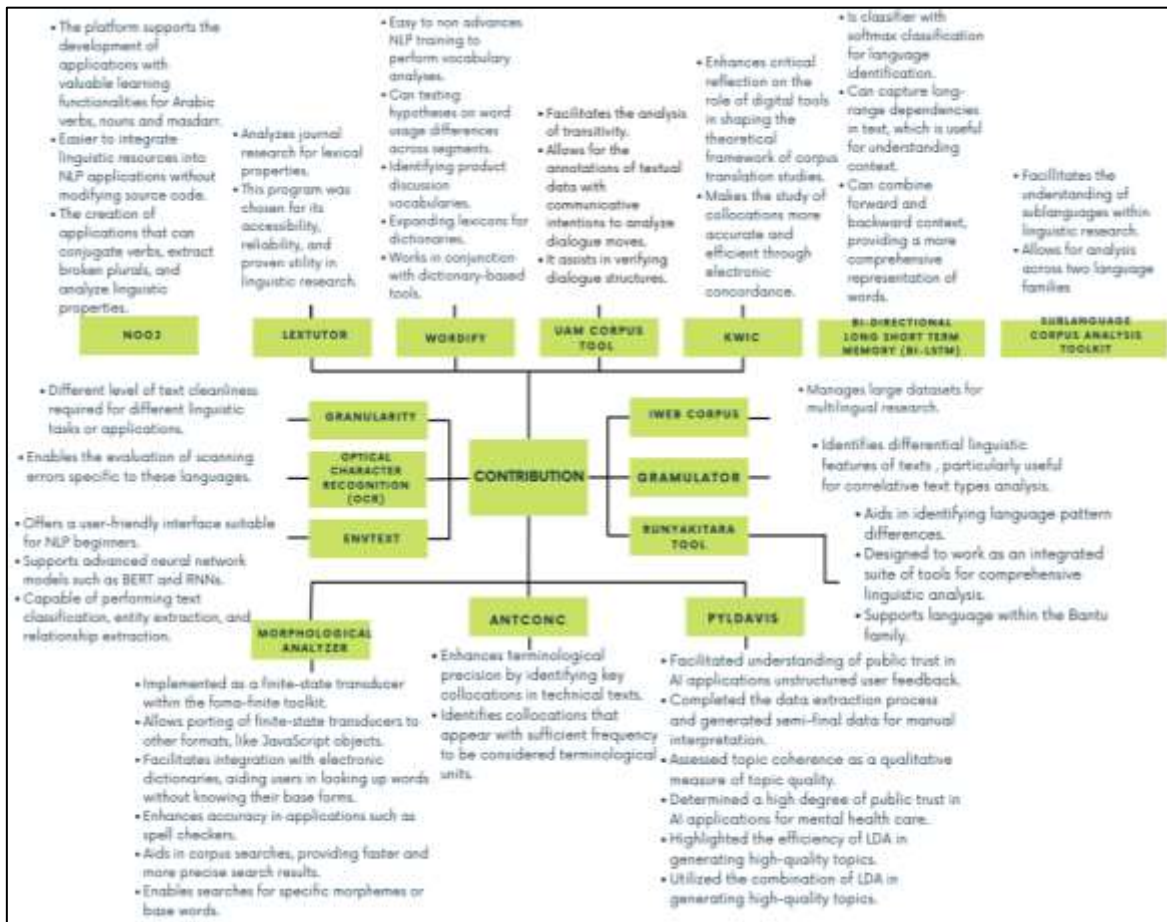


Figure 3. Contributions of computational tools based on the selected articles

From these results, we can observe the continued development and enhancement of user-friendly open-source corpus tools that cater to the evolving needs of researchers. The integration of advanced technologies such as machine learning and NLP into corpus tools enables more sophisticated analyses. Accessibility and open data initiatives are promoted to make diverse corpora more accessible to researchers. This can be achieved through the platform's development. This facilitates the sharing and collaborative creation of corpora, thereby reducing barriers to researchers' entry. Furthermore, joint efforts have bridged the gap between existing tools, new machine learning AI, and model approaches, promoting a holistic understanding of language. A new advancement in current research is multimodal and multilingual adaptation. The development of tools has recognized the increasing complexity of linguistic research. However, these contributions have limitations.

3.3. Limitations of computational tools

This section discusses the limitations and vulnerabilities of the tools used for data analysis, as highlighted in the 15 articles. These articles mention the constraints either in the methods section, under specific subheadings related to the tools, or in the results section. Figure 5 provides an overview of the limitations of computational tool-based research. This figure shows that researchers need to investigate

before using the tool [62], which may prevent the acquisition of accurate data and findings capable of answering the study question.

Among the implications of the weaknesses of this corpus tool, researchers have found it difficult to gain a full understanding of language in the desired context. For example, Wordify adapted statistical logistic regression methods. This reduces its ability to detect complex languages across the board, which may result in incomplete and imperfect analysis. In addition, the limited ability of tools to analyze data in large corpus sizes makes it difficult and slows down the process of bulk data acquisition, as researchers must manually organize corpus data in minimal amounts before it is included in the corpus tool. Examples include the SketchEngine and SubCAT. In addition, the lack of volume of lemmatization, vocabulary in several languages in the application, database management, broken forms in several languages, language resources, application management, and others are also mentioned in Figure 5. Therefore, it is important for researchers to understand each weakness of this tool in depth so that the analysis process does not take a long time. However, in addition to the weaknesses of these tools, there are also weaknesses that arise because of the shortcomings of the researchers themselves. In conclusion, each user should equip themselves with sufficient skills and knowledge of the tool as well as perform corpus tool testing as a preliminary preparatory step before performing linguistic corpus research.

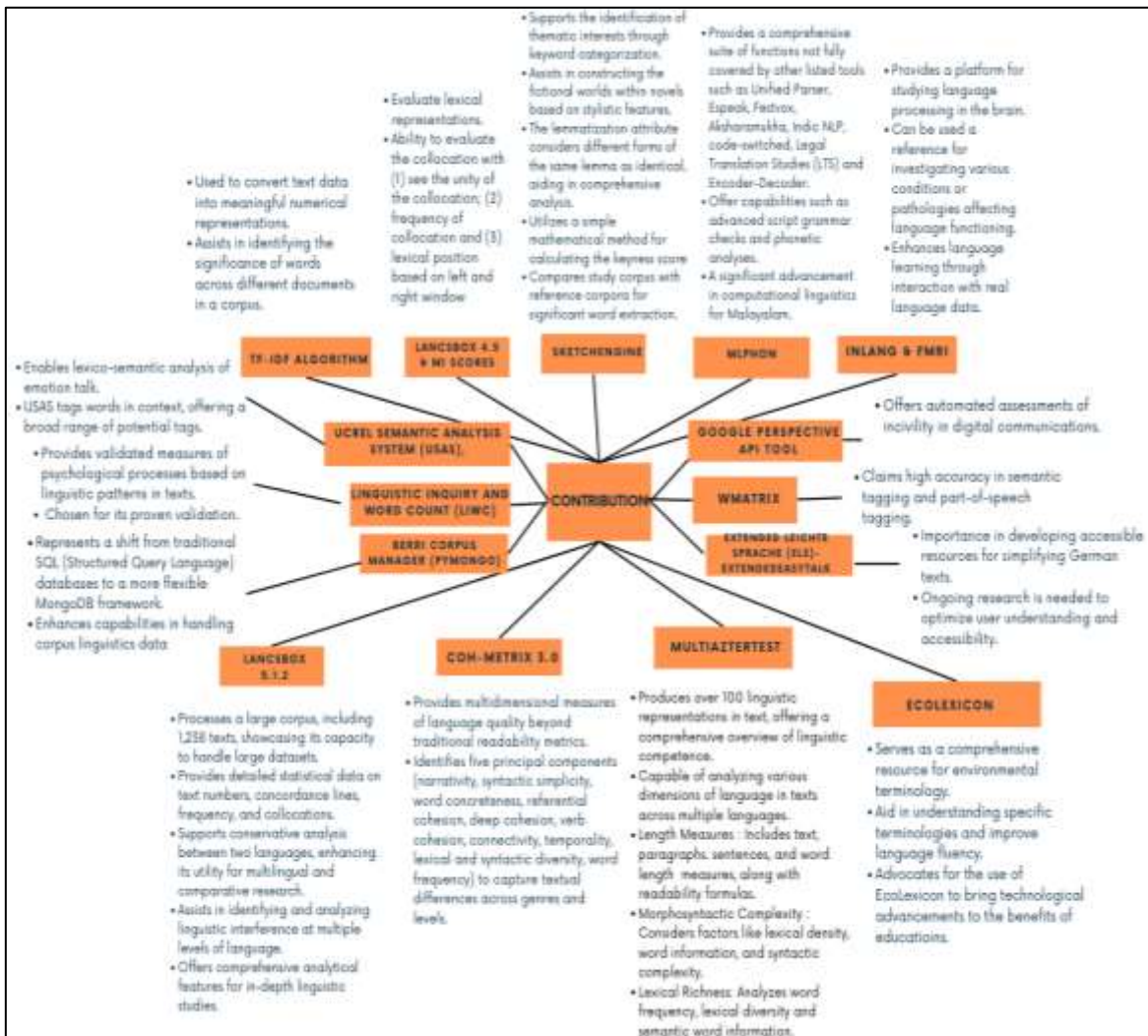


Figure 4. Contributions of computational tools based on selected articles

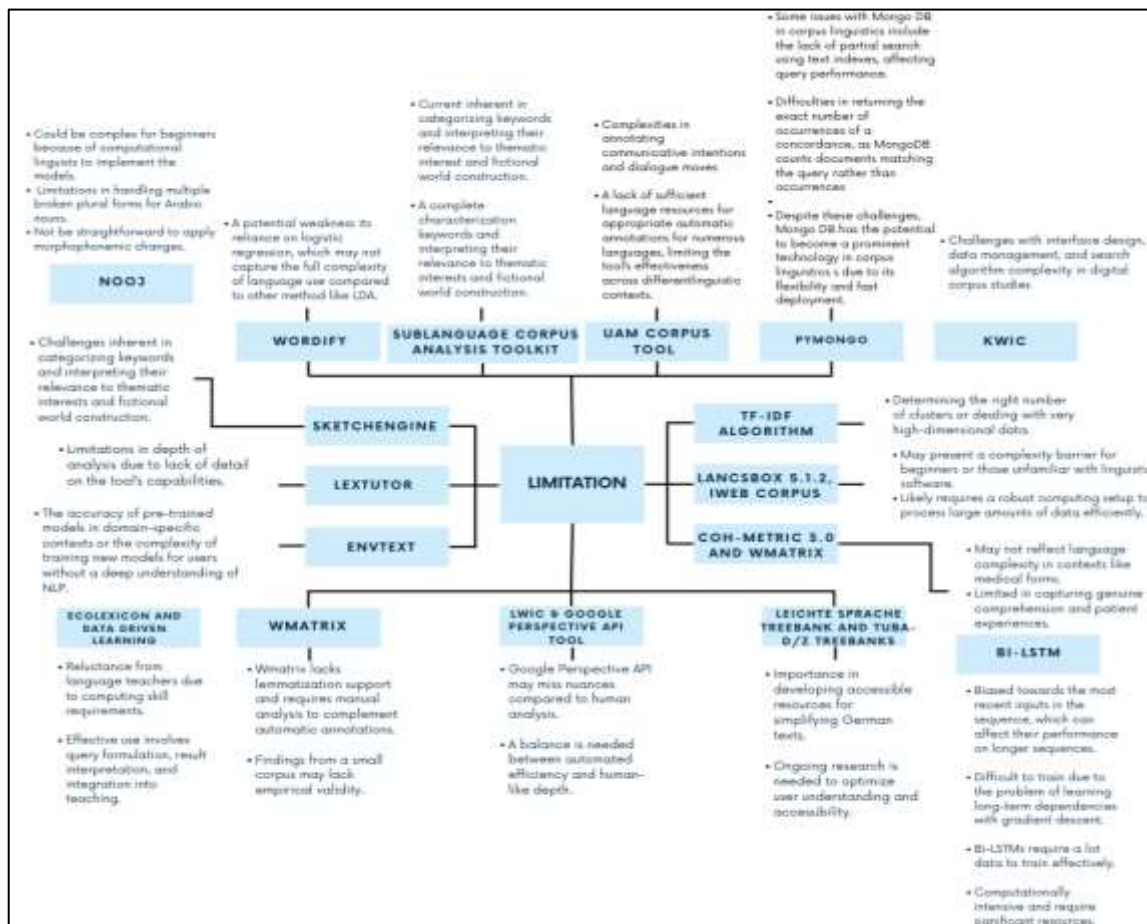


Figure 5. Limitations of computational tools based on the selected articles

4. CONCLUSION

In conclusion, this SLR presents the use of computational tools most recently used by researchers in linguistics. The variety of uses of computational tools by researchers in the field of language signifies that these researchers not only use such tools but also combine them with other technological tools to assist in the analysis of their studies. This study shows that language researchers have successfully explored and learned to use advanced applications through AI technology tools and other facilities, such as DBMS, IDEs, web scrapping and data extraction tools, speech processing tools, and statistical analysis software.

In addition, the combination of various theories, language models, and NLP models has led to new and diverse discoveries in this field. This shows excellent reform, and the data findings are more accurate than those obtained with the limited use of a single computational tool. However, there are disadvantages in terms of the features of the computational tools. This is most notable in the language selection section of the NLP tool. This makes it difficult for users to analyze their data using applications, and they have no choice but to use existing facilities with limited features. To obtain better data results, they require knowledge and skills in the field of AI algorithms or additional knowledge from machine learning programs and other technology tools as a complement to their data analysis.


Therefore, this field still requires the further production of new and advanced projects to upgrade existing computational tools, as several constraints still limit users from analyzing data. However, while the availability of additional tool support as a complement is seen to help overcome existing cocoons, this cocoon, in terms of this facility, must be overcome from time to time to facilitate the user. This will further enhance the motivation and ability of researchers to continue studying languages in various fields and encourage interdisciplinary collaboration among linguists, computer scientists, and experts from related fields to foster the development of innovative tools and methodologies. It is hoped that the availability of exploration in computational tools will help researchers see the potential of computational tools available today and increase the number of studies using computational tools in different fields and needs, leading to improvements and updates for each computational tool in corpus research.

REFERENCES





- [1] X. Lu, *Computational methods for corpus annotation and analysis*, vol. 9789401786454. Springer New York Heidelberg Dordrecht London, 2014. doi: 10.1007/978-94-017-8645-4.
- [2] P. M. Davies, R. J. Passonneau, S. Muresan, and Y. Gao, "Analytical techniques for developing argumentative writing in STEM: a pilot study," *IEEE Transactions on Education*, vol. 65, no. 3, pp. 373–383, 2022, doi: 10.1109/TE.2021.3116202.
- [3] F. Goyak, M. M. Muhammad, M. F. Zaini, W. MM. A. Ibrahim, and A. Gunsuh, "Diachronic analysis of the profane words in English song lyrics: a computational linguistics perspective," *Malaysian Journal of Music*, vol. 11, no. 1, pp. 1–19, 2022, doi: 10.37134/mjm.vol11.1.2.2022.
- [4] W. W. Lun *et al.*, "Analysis of covid-19 related phrases using corpus-based tools: dualisms language & technology," *Journal of Positive School Psychology*, vol. 6, no. 3, pp. 5034–5044, 2022.
- [5] C. Minjie, W. W. Lun, Y. Guojie, C. K. S. Singh, W. Mihat, and Y. S. May, "Global intellectual trend of corpus linguistics studies among scholars in social sciences from September 2013 – September 2021," *Asian Journal of University Education*, vol. 19, no. 4, pp. 613–632, 2023, doi: 10.24191/ajue.v19i4.24615.
- [6] S. Goźdź-Roszkowski, "Corpus linguistics in legal discourse," *International Journal for the Semiotics of Law*, vol. 34, no. 5, pp. 1515–1540, 2021, doi: 10.1007/s11196-021-09860-8.
- [7] X. Cheng and X. Shi, "A corpus-based study of the discursive construction of corporate identities by Chinese and American banks," *Contrastive Pragmatics*, vol. 3, no. 2, pp. 313–335, 2021, doi: 10.1163/26660393-12340008.
- [8] D. Hovy, S. Melumad, and J. J. Inman, "Wordify: a tool for discovering and differentiating consumer vocabularies," *Journal of Consumer Research*, vol. 48, no. 3, pp. 394–414, 2021, doi: 10.1093/jcr/ucab018.
- [9] C. Baden, C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden, "Three Gaps in computational text analysis methods for social sciences: a research agenda," *Communication Methods and Measures*, vol. 16, no. 1, pp. 1–18, 2022, doi: 10.1080/19312458.2021.2015574.
- [10] S. M. Maci and M. Sala, *Corpus linguistics and translation tools for digital humanities : research methods and applications*, vol. 3, 2023. doi: 10.1515/jccall-2023-0007.
- [11] V. Brezina, *Statistics in corpus linguistics: a practical guide*. 2018. doi: 10.1017/9781316410899.
- [12] B. E. Sibarani, "What do we know about balanced scorecard and its benefit? a systematic literature review," *Jurnal Dinamika Akuntansi dan Bisnis*, vol. 10, no. 1, pp. 133–148, 2023, doi: 10.24815/jdab.v10i1.29351.
- [13] D. Sandra, J. Segers, and R. Giacalone, "How organizations can benefit from entrainment: a systematic literature review," *Journal of Organizational Change Management*, vol. 36, no. 2, pp. 233–256, 2022, doi: 10.1108/JOCM-01-2022-0023.
- [14] A. Panayi, K. Ward, A. Benhadji-Schaff, A. S. Ibanez-Lopez, A. Xia, and R. Barzilay, "Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews," *Systematic Reviews*, vol. 12, no. 1, pp. 1–11, 2023, doi: 10.1186/s13643-023-02351-w.
- [15] P. Martin-Rodilla and M. Sánchez, "Software support for discourse-based textual information analysis: a systematic literature review and software guidelines in practice," *Information (Switzerland)*, vol. 11, no. 5, 2020, doi: 10.3390/INFO11050256.
- [16] J. Antidze, N. Gulua, and I. Kardava, "The software for composition of some natural languages' words," *Lecture Notes on Software Engineering*, vol. 1, no. 3, pp. 295–297, 2013, doi: 10.7763/Inse.2013.v1.64.
- [17] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, 2021, doi: 10.1136/bmj.n71.
- [18] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ (Clinical research ed.)*, vol. 339, 2009, doi: 10.1136/bmj.b2700.
- [19] D. Vojnović, "Key noun+noun collocations in the language of tourism: a corpus-based study of English and Serbian," *ELOPE: English Language Overseas Perspectives and Enquiries*, vol. 18, no. 2, pp. 51–68, 2021, doi: 10.4312/ELOPE.18.2.51-68.
- [20] J. Buts and H. Jones, "From text to data: mediality in corpus-based translation studies," *Monografias de Traducción e Interpretación (MonTI)*, no. 13, pp. 301–329, 2021, doi: 10.6035/MonTI.2021.13.10.
- [21] C. Laske, "Corpus linguistics: the digital tool kit for analysing language and the law," *Comparative Legal History*, vol. 10, no. 1, pp. 3–32, Jan. 2022, doi: 10.1080/2049677X.2022.2063510.
- [22] T. M. G. Roberto, "Corpus linguistics in language teaching," *REVISTA FOCO*, vol. 16, no. 7, p. e2631, Jul. 2023, doi: 10.54751/revistafoco.v16n7-089.
- [23] A. Sarudin, H. F. M. Redzwan, and A. A. Tan, "The development of a medicinal nadir glossary through the # lancsbox 6 . 0 of wizard software," *Malaysia Journal of Invention and Innovation*, vol. 1, no. 1, pp. 73–84, 2022.
- [24] Abhay A. Dande and Dr. M. A. Pund, "A review study on applications of natural language processing," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 10, no. 2, pp. 122–126, 2023, doi: 10.32628/ijrsrset2310214.
- [25] F. Shah-Mohammadi and J. Finkelstein, "Combining NLP and Machine learning for differential diagnosis of COPD exacerbation using emergency room data," *Studies in Health Technology and Informatics*, vol. 305, pp. 525–528, 2023, doi: 10.3233/SHTI230549.
- [26] A. Ş. Özbay and Z. Gürsoy, "Computerized corpus as a tool for educational technology and learning in the analysis of four-word recurrent expressions," *Journal of Educational Technology and Online Learning*, vol. 6, no. 1, pp. 249–272, 2023, doi: 10.31681/jetol.1186346.
- [27] T. Gaillat *et al.*, "A data repository for the management of dynamic linguistic datasets," in *A data repository for the management of dynamic linguistic datasets*, 2021.
- [28] P. T. Duncan, H. Torrence, T. Major, and J. Kandybowicz, "Managing data for theoretical syntactic study of underdocumented languages," in *The Open Handbook of Linguistic Data Management*, 2022, pp. 523–530.
- [29] X. Q. Xia, M. McClelland, and Y. Wang, "TabSQL: a MySQL tool to facilitate mapping user data to public databases," *BMC Bioinformatics*, vol. 11, pp. 2–5, 2010, doi: 10.1186/1471-2105-11-342.
- [30] R. Németh and J. Koltai, "Natural language processing: the integration of a new methodological paradigm into sociology," *Intersections East European Journal of Society and Politics*, vol. 9, no. 1, pp. 5–22, 2023, doi: 10.17356/ieejsp.v9i1.871.
- [31] K. Manohar, A. R. Jayan, and R. Rajan, "Mlphon: a multifunctional grapheme-phoneme conversion tool using finite state transducers," *IEEE Access*, vol. 10, pp. 97555–97575, 2022, doi: 10.1109/ACCESS.2022.3204403.
- [32] S. L. R. Schreiner, L. Schwartz, B. Hunt, and E. Chen, "Multidirectional leveraging for computational morphology and language documentation and revitalization," *Language Documentation and Conservation*, vol. 14, pp. 69–86, 2020.
- [33] F. Katshemererwe, A. Caines, and P. Buttery, "Building natural language processing tools for Runyakitara," *Applied Linguistics Review*, vol. 12, no. 4, pp. 585–609, 2021, doi: 10.1515/applirev-2020-2004.

- [34] A. I. Martínez-Hernández, "Using a free corpus tool for time-efficient feedback on english as a foreign language reports," *Miscelanea*, vol. 66, no. 2022, pp. 13–39, 2022, doi: 10.26754/ojs_misc/mj.20227356.
- [35] R. T. Viera, "Lexical richness of abstracts in scientific papers in anglophone and non-anglophone journals," *3L: Language, Linguistics, Literature*, vol. 28, no. 2, pp. 224–239, 2022, doi: 10.17576/3L-2022-2802-15.
- [36] P. M. McCarthy, N. W. Kaddoura, A. Al-Harthy, A. M. Thomas, N. D. Duran, and K. Ahmed, "Corpus analysis on students' counter and support arguments in argumentative writing," *Pegem Eğitim ve Öğretim Dergisi*, vol. 12, no. 1, pp. 256–271, 2022, doi: 10.47750/pegegog.12.01.27.
- [37] T. Isaacs, J. Murdoch, Z. Demjén, and F. Stevenson, "Examining the language demands of informed consent documents in patient recruitment to cancer trials using tools from corpus and computational linguistics," *Health (United Kingdom)*, vol. 26, no. 4, pp. 431–456, 2022, doi: 10.1177/1363459320963431.
- [38] H. Stevens, M. E. Rasul, and Y. J. Oh, "Emotions and incivility in vaccine mandate discourse: natural language processing insights," *JMIR Infodemiology*, vol. 2, no. 2, pp. 1–13, 2022, doi: 10.2196/37635.
- [39] Z. Liu and H. Liu, "The construction of china's national image from transitivity perspective—a case study of fighting COVID-19: China in action," *Theory and Practice in Language Studies*, vol. 11, no. 11, pp. 1421–1427, 2021, doi: 10.17507/tpls.1111.09.
- [40] O. Yaskorska-Shah, "Data-driven dialogue models: applying formal and computational tools to the study of financial and moral dialogues," *Studies in Logic, Grammar and Rhetoric*, vol. 63, no. 1, pp. 185–208, 2020, doi: 10.2478/slgr-2020-0034.
- [41] B. S. M. Moustafa, "A comparative corpus stylistic analysis of the thematization and characterization in Gordimer's My Son's Story and Coetzee's Disgrace," *Open Linguistics*, vol. 16, no. 1, pp. 1–18, 2022, doi: 10.1080/19312458.2021.2015574.
- [42] A. Granados, A. Lorenzo-Espejo, and F. Lorenzo, "A portrait of academic literacy in mid-adolescence: a computational longitudinal account of cognitive academic language proficiency during secondary school," *Language and Education*, vol. 37, no. 3, pp. 288–307, 2023, doi: 10.1080/09500782.2022.2079951.
- [43] I. P. Temnikova *et al.*, "Sublanguage corpus analysis toolkit: a tool for assessing the representativeness and sublanguage characteristics of corpora," *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 1714–1718, 2014.
- [44] H. Bi, B. Li, Y. Qiu, and M. Change, "EnvText: a Chinese text mining tool for environmental domain with advanced bert model," *Software Impacts*, vol. 17, pp. 1–4, 2023, doi: 10.1016/j.simpa.2023.100559.
- [45] E. Roger *et al.*, "Unraveling the functional attributes of the language connectome: crucial subnetworks, flexibility and variability," *NeuroImage*, vol. 263, no. October, p. 119672, 2022, doi: 10.1016/j.neuroimage.2022.119672.
- [46] H. Sanjurjo-González, "Berri corpus manager: a corpus analysis tool using MongoDB technology," *Frontiers in Artificial Intelligence and Applications*, vol. 328, pp. 166–173, 2020, doi: 10.3233/faia200619.
- [47] C. I. López-Rodríguez, "Emotion at the end of life: semantic annotation and key domains in a pilot study audiovisual corpus," *Lingua*, vol. 277, 2022, doi: 10.1016/j.lingua.2022.103401.
- [48] A. B. Ormanova and M. L. Anafinova, "A linguistic interference in information space terms: a corpus-based study in Kazakh," *Theory and Practice in Language Studies*, vol. 12, no. 12, pp. 2497–2507, 2022, doi: 10.17507/tpls.1212.04.
- [49] M. F. Zaini *et al.*, "Geometric lexical representative perspectives: the impact of threshold values through #Lancsbox Software," *AIP Conference Proceedings*, vol. 2644, 2022, doi: 10.1063/5.0104817.
- [50] W. W. Lun, M. M. Muhammad, W. Mihat, M. A. Rahman, M. S. Y. Shak, and L. M. Chew, "Using technologised computational corpus-driven linguistics study on the vocabulary uses among advanced malaysian upper primary school English as a second language learners (ESL) in northern region," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 31, no. 1, pp. 298–314, 2023, doi: 10.37934/ARASET.31.1.298314.
- [51] I. Blanchete and M. Mourchid, "The use of Arabic linguistic resources to develop learning applications," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 29, no. 1, pp. 562–571, 2023, doi: 10.11591/ijeecs.v29.i1.pp562-571.
- [52] M. Z. Ansari, T. Ahmad, M. M. S. Beg, and N. Bari, "Language lexicons for Hindi-English multilingual text processing," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 641–648, 2022, doi: 10.11591/ijai.v11.i2.pp641-648.
- [53] B. Li, "Integrating linguistic theory and neural language models," *ArXiv Preprint*, 2022. [Online]. Available: <http://arxiv.org/abs/2207.09643v1>
- [54] P. Yao, M. Kosmajac, A. Waheed, K. Guzhva, N. Hervieux, and D. Barbosa, "NLP workbench: efficient and extensible integration of state-of-the-art text mining tools," *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 18–26, 2023, doi: 10.18653/v1/2023.eacl-demo.3.
- [55] Y. Shan, M. Ji, W. Xie, K.-Y. Lam, and C.-Y. Chow, "Public Trust in artificial intelligence applications in mental health care: topic modeling analysis," *JMIR Human Factors*, vol. 9, no. 4, pp. 1–12, 2022, doi: 10.2196/38799.
- [56] D. J. Prinsloo, E. Taljard, and M. Goosen, "Optical character recognition and text cleaning in the indigenous South African languages," *Stellenbosch Papers in Linguistics Plus*, vol. 64, pp. 165–187, 2022, doi: 10.5842/64-1-867.
- [57] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, pp. 776–784, 2022, doi: 10.11591/ijece.v12i1.pp776-784.
- [58] K. Harbusch and I. Steinmetz, "A computer-assisted writing tool for an extended variety of leichte sprache (easy-to-read German)," *Frontiers in Communication*, vol. 6, no. January, pp. 1–22, 2021, doi: 10.3389/fcomm.2021.689009.
- [59] M. Rudneva, "Corpus-driven ESP pedagogy: a preliminary case study," *Journal of Teaching English for Specific and Academic Purposes*, vol. 8, no. 3, pp. 241–248, 2020, doi: 10.22190/JTESAP2003241R.
- [60] A. Ahnaf, H. M. Mahmudul Hasan, N. S. Sworna, and N. Hossain, "An improved extrinsic monolingual plagiarism detection approach of the Bengali text," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 4, pp. 4256–4267, 2023, doi: 10.11591/ijece.v13i4.pp4256-4267.
- [61] W. Layang and L. Zhaoxia, "A review on the weakness of UAM corpus tool," *EAS Journal of Humanities and Cultural Studies*, vol. 4, no. 5, pp. 214–218, 2022, doi: 10.36349/easjhcs.2022.v04i05.004.
- [62] G. G. Rafjabova, "Corpus technologies in teaching academic writing," *Foreign Languages in Uzbekistan*, vol. 1, no. 1, pp. 92–103, 2023, doi: 10.36078/1679549918.
- [63] E. Osipova and E. Bagrova, "Corpus linguistic technology as a tool to improve creative thinking in the interpretation of English language idioms," *Lecture Notes in Networks and Systems*, vol. 345. Springer International Publishing, pp. 948–962, 2022. doi: 10.1007/978-3-030-89708-6_76.





BIOGRAPHIES OF AUTHORS

Wan Nur Aida Sakinah Wan Jusoh     is doctoral candidate in Malay Language Education of the Universiti Pendidikan Sultan Idris, Malaysia (UPSI) and was earning a Master in the same field at Universiti Kebangsaan Malaysia (UKM). Since 2015, she has served as a Malay Language teacher and was received Hadiah Latihan Persekutuan (HLP) scholarship from Ministry of Education Malaysia (KPM) in 2021. Her research focuses on pedagogy in teaching and learning language, discourse analysis and corpus linguistic. She can be contacted at email: csakinah00@gmail.com.







Norfaizah Abdul Jobar     obtained her Ph.D. and Master in Malay Language Education from Universiti Pendidikan Sultan Idris, Malaysia (UPSI) in 2017 and 2010. She is senior lecturer at the same university at Faculty of Languages and Communication. Her research interests include pedagogy in teaching and learning, essay writing, and discourse analysis in language education. She can be contacted at email: norfaizah.aj@fbk.upsi.edu.my.



Md Zahril Nizam Md Yusoff     is a lecturer at the Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris, Perak, Malaysia (UPSI). He pursued his graduate studies, earning a Master of Literature in Malay Language at Universiti Kebangsaan Malaysia (UKM), and received a Ph.D. in Malay Language study from Universiti Sains Malaysia (USM). His research interests include critical discourse analysis (CDA) in linguistic forensic and corpus linguistics. He can be contacted at email: mdzahril@fbk.upsi.edu.my.



Hanifah Binti Mahat     is an associate professor and senior lecturer from the Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Perak (UPSI). She pursued her graduate studies in Master of Science in Technology Education at Universiti Putra Malaysia (UPM) and received a Ph.D. degree in Geography at Universiti Pendidikan Sultan Idris, Malaysia (UPSI). She is an expert in geography education and technology education. She can be contacted at email: hanifah.mahat@fsk.upsi.edu.my.