


Review Article

Corpus Linguistics Today: A Qualitative Approach

Misnawati MISNAWATI^{1*} 
Sahril NUR² 

Saidna Zulfiqar Bin TAHIR³ 

Corresponding author: Misnawati_amir@yahoo.com

DOI: <https://doi.org/10.31963/rial.v2i1.4486>

Received: 27/10/2023

Revised: 04/12/2023,

Accepted: 24/12/2023

ABSTRACT

Corpus linguistics, the study of language form and function using computerized corpora, involves collecting extensive electronic texts to analyze language usage. It serves various objectives: qualitative analysis, exploring nuances in language use, and quantitative analysis, identifying patterns in word usage and collocations. Corpus linguistics tracks language variation over time, aids language teaching, supports lexicography, and contributes to discourse analysis. Its evolution grew from the 1950s to the present day, marked by technological advancements and theoretical integration. The article underscores the significance of qualitative analysis in providing contextual insights into language use. The systematic creation of a corpus involves steps like text selection, data collection, preprocessing, annotation, and quality control. Various qualitative analysis techniques, from discourse analysis to lexical semantics, offer diverse perspectives for studying linguistic phenomena. This article provides a concise overview of corpus linguistics, its evolution, the importance of qualitative analysis, corpus creation, and qualitative analysis techniques. Therefore, this article might contribute to corpus linguists, corpus researchers, and ESP teachers in starting, studying, and applying corpus.

Keywords: *Corpus linguistics, qualitative corpus approach, corpus today*

To cite this article: Misnawati., Nur, S., Tahir, S. Z. (2023). Corpus Linguistics Today: A Qualitative Approach. Research and Innovation in Applied Linguistics [RIAL], Vol 2 (1), 45-62.

<https://doi.org/10.31963/rial.v2i1.4486>

¹ Universitas Cahaya Prima, Bone, INDONESIA. Email: misnawati_amir@yahoo.com

ORCID: <https://orcid.org/0000-0003-3257-7464>

² Universitas Negeri Makassar, Makassar, INDONESIA. Email: sahrilfbsunm@unm.ac.id

ORCID: <https://orcid.org/0000-0002-6544-7378>

³ Universitas Iqra Buru, Maluku, INDONESIA. Email: saidnazulfiqar@yahoo.com

ORCID: <https://orcid.org/0000-0002-9533-5465>



INTRODUCTION

Corpus Linguistics has revolutionized our understanding and investigation of language, resulting in an explosion of empirical study encompassing practically all aspects of language. Primarily attributable to the information revolution of the late 20th century is the development and expansion of corpus and computational techniques within descriptive and applied linguistics (McCarthy & O'Keeffe, 2010). What began as an improvement in methodology, facilitated by an ever-increasing volume of data and more powerful, efficient computers for processing, has evolved into a paradigm shift with significant theoretical implications, necessitating a reevaluation of many established theoretical stances within the field (Tognini-Bonelli, 2010). Corpus linguistics is a branch of linguistics that relies on an extensive collection of authentic text samples, known as a corpus. These broad and diverse language collections enable linguists and scholars to examine language patterns, usage, and structure systematically and in a data-driven manner. This article examines the present state of corpus linguistics, focusing on qualitative approaches that increase our understanding of language beyond statistical analysis.

"Corpus linguistics today embraces qualitative approaches for deeper linguistic insights."

UNDERSTANDING CORPUS LINGUISTICS

Corpus Linguistics and Its Primary Objectives

Corpus Linguistics has become a term referring to a wide range of activities and approaches (Hunston & Su, 2019). Collecting enormous quantities of text in electronic format appears to be a feature all these entities share. Thus, Corpus Linguistics could be described succinctly as "the study of the form and function of language through the examination of computerized corpora" (Partington et al., 2013). However, a corpus is not just a simple collection of texts. Instead, a corpus seeks to represent a language or some part of a language (Biber et al., 2010).

The expansive characterization presented here encompasses a substantial corpus of scholarly investigations conducted through methodologies examining a specific search term and its surrounding context (e.g., keyword-in-context or concordance lines). These methodologies involve the computation of relative frequency (e.g., collocation studies) and the annotation of various linguistic features, such as word class, grammatical function, or semantic class, to facilitate frequency calculations based on these categories (Goźdz-Roszkowski, 2021). Furthermore, the field of Corpus Linguistics in the present day has experienced significant diversification, manifesting in several forms contingent upon the researcher's perspective on the status and use of corpus material. The treatment of corpus evidence by researchers has emerged as a significant concern, leading to the adoption of several techniques (Tognini-Bonelli, 2010) based on the primary objectives of corpus linguistics itself as follows:

- a. Descriptive/ Qualitative analysis: Corpus linguistics seeks to characterize language usage in real-world contexts. It attempts to accurately represent how words, phrases, and grammatical structures are used in various written and spoken discourse types. It is essential for exploring aspects of language, such as nuanced word meanings, pragmatic behaviors, and contextual usage, that quantitative analysis alone cannot reveal (Biber et al., 2010).
- b. Quantitative Analysis: One of the primary goals is to conduct quantitative analyses to identify patterns and tendencies in language utilization. This involves calculating the frequency of words, analyzing collocations, and analyzing distributional patterns (Sinclair, 1991).
- c. Variation and Change: Corpus linguistics enables linguists to monitor variation and change in language over time. By comparing corpora from various periods, researchers can identify shifts and developments in the language (Conrad, 2004; Crawford & Csomay, 2016).
- d. Linguistic Research: Corpus linguistics is an indispensable instrument for evaluating linguistic hypotheses. Linguists utilize Corpora to investigate a vast array of linguistic phenomena, from syntax and semantics to pragmatics and discourse analysis (Illie & Norrick, 2018)
- e. Language Teaching and Learning: Corpora are increasingly used to provide authentic examples of language use in language teaching and learning. The exposure of learners to authentic language in context can improve their comprehension and language proficiency (Friginal, 2018; Gavioli, 2005; O'Keeffe et al., 2007; Timmis, 2015).
- f. Lexicography: Corpora are essential in lexicography, assisting dictionary compilers with providing accurate and current definitions, examples, and collocations for words and phrases (Fuertes-Olivera, 2018; Kennedy, 1998).
- g. Stylistics and discourse analysis: It aids in analyzing how language is employed for particular purposes, such as in literary texts, academic writing, and advertisements (Crawford & Csomay, 2016; Hoey et al., 2007).

At its core, corpus linguistics is a collaboration between empirical research and the depth of rigorous linguistic analysis. It allows linguists to explore language in a structured and evidence-based way, uncovering patterns, variations, and insights critical to our understanding of how language works in practice."

The Evolution of Corpus Linguistics in Linguistic Research

Corpus Linguistics has a long history dating back to the early 20th century. Its evolution in linguistic research has been a transformative journey, significantly reshaping how we

understand and analyze language. Over the decades, corpus linguistics has evolved from a relatively niche methodology to a cornerstone of modern linguistics. This evolution can be delineated through several key phases:

- a. Pre-modern concordance (1950s): Corpus linguistics stems from the emergence of a (pre-modern) concordance in the late 1950s to the practical needs of scholars in the study of the Bible (O'Keeffe & McCarthy, 2010). Concordance is an alphabetical index of the principal words in a book or the works of an author with their immediate contexts.
- b. Early Beginnings (1960s-1970s): Corpus linguistics as a text analysis method based on electronic tools can be traced back to the 1960s and 1970s when researchers began to compile and analyze extensive text collections. One of the pioneering projects in Corpus Linguistics was the Brown Corpus created in the 1960s at Brown University and Lancaster-Oslo-Bergen (LOB) corpora, which contained two collections of 1 million words and 500 sample texts of American and British English by Henry Kucera and W. Nelson Francis. This seminal text corpus marks the beginnings of quantitative linguistic analysis Francis & Kucera 1979; Johannson et al. 1978 as cited by Gavioli (2005).
- c. Growth of Computer Technology (1980s): The advent of computer technology was a game-changer for corpus linguistics. The development of more powerful computers allowed the creation of larger, more diverse corpora and advanced analytical tools. In 1980, Sinclair, in his earliest exploratory years of corpus analysis that was to culminate in the ground-breaking COBUILD project, started to collect a corpus of texts on the computer for dictionary making and language study. The compilers of the Collins Cobuild English Language Dictionary (1987) had daily access to a corpus of approximately 20 million words (O'Keeffe & McCarthy, 2010).
- d. Theoretical integration (1990s): In the 1990s, corpus linguistics benefited from technological progress and is very much ascending. By the time computers could be used by anyone, the traditions of searching through texts for all examples of a specific part of the language, writing dictionaries based on proven usage, and analyzing language based on actual informant data became well established. Then, researchers began to use corpora to test and refine linguistic hypotheses, leading to a more nuanced understanding of language (O'Keeffe & McCarthy, 2010). This age was launched by some corpora, like the Bank of English (BoE) (1991) and the British National Corpus (BNC) (1995).
- e. Specialized Corpora (2000s): The 21st century saw the development of specialized corpora, focusing on specific linguistic aspects like spoken language, historical language, or learner language. These corpora enriched linguistic research in various

subfields. A large and growing number of highly specialized corpora are being created for a particular purpose. One of them is the Academic Word List (AWL) corpus by Coxhead (2000) (Coxhead & Nation, 2001).

- f. Interdisciplinary Collaborations (2010s): Corpus linguistics became increasingly interdisciplinary, with collaborations between linguists, computer scientists, and other fields. This synergy led to the development of advanced software tools and applications, further broadening the scope of research. There are, however, a growing number of corpora available in other languages as well. Some are monolingual corpora - collections of text from one language- and multilingual corpora with the exact text in several languages (Taylor & Marchi, 2018).
- g. The Digital Age (Present): In the digital age, corpora are more accessible and extensive than ever. Online resources and collaborative efforts have democratized corpus linguistics, making it a fundamental resource for linguistic research in academia and industry. More and more universities offer courses in corpus linguistics and use corpora in their teaching and research. The number and diversity of corpora being compiled are significant, and corpora are used in many projects. More information can be found by browsing the net and reading journals and books. The electronic mailing list Corpora can be a good starting point for someone who wishes to learn about what goes on within the field of corpus linguistics at the moment (Absalom, 2021; Mitkov, 2022).

The development of corpus linguistics demonstrates its transformation from a novel methodology to a fundamental aspect of linguistic research. It has made the study of language in diverse contexts possible, casting light on linguistic patterns, variation, and change. The future of corpus linguistics in linguistic research is even more promising as technology advances and new applications are developed.

QUALITATIVE APPROACHES IN CORPUS LINGUISTICS

Qualitative Research Methods in the Context of Corpus Linguistics

Corpus analysis can be broadly categorized as consisting of qualitative and quantitative analysis. Qualitative approaches play a pivotal role in corpus linguistics by offering a deeper, more nuanced understanding of language. It is a methodology for in-depth investigations of linguistic phenomena grounded in authentic communicative situations (Hasko, 2012). The aim of qualitative analysis is a complete, detailed description. The main disadvantage of qualitative approaches to corpus analysis is that their findings cannot be extended to broader populations with the same degree of certainty that quantitative can analyze. The research findings are not tested to discover whether they are statistically

significant or due to chance. Qualitative approaches, on the other hand, enable researchers to explore the 'how' and 'why' behind linguistic phenomena (Bennet, 2010).

In quantitative research, we classify features, count them, and even construct more complex statistical models to explain what is observed. Findings can be generalized to a larger population, and direct comparisons can be made between two corpora so long as valid sampling and significance techniques have been used (Ludeling & Kyto, 2009). Thus, quantitative analysis allows us to discover which phenomena are likely to be genuine reflections of the behavior of a language or variety and which are merely chance occurrences (Adamou, 2019). However, the picture of the data that emerges from quantitative analysis is less rich than that obtained from qualitative analysis. Quantitative analysis is, therefore, an idealization of the data in some cases. Also, quantitative analysis tends to sideline rare occurrences.

Thus, doing qualitative research in the context of corpus linguistics is particularly significant for several reasons:

- a. **Contextual Understanding:** Qualitative approaches allow researchers to consider the contextual factors influencing language use. This is essential for comprehending the subtleties of language, such as idiomatic expressions, metaphors, and cultural influences (Aijmer & Altenberg, 2014).
- b. **Semantic Analysis:** Qualitative methods enable the examination of semantic nuances, helping researchers identify connotations, pragmatic meanings, and speaker intentions that quantitative analyses may overlook (Bennet, 2010).
- c. **Fine-Grained Patterns:** By closely examining individual instances of language, qualitative approaches unveil fine-grained patterns, variations, and exceptions that quantitative analysis might homogenize. This analysis should consist of complete, naturally occurring texts (oral or written) whose origins and provenance are well documented (Hasko, 2012).
- d. **Interdisciplinary Insights:** Qualitative corpus analysis is often interdisciplinary, drawing from sociolinguistics, discourse analysis, and pragmatics. This interdisciplinary perspective enriches our understanding of language in different social and cultural contexts (Flowerdew, 2009).
- e. **Hypothesis Generation:** Qualitative approaches can be a valuable starting point for generating hypotheses, which can later be tested quantitatively. They help researchers form more precise research questions and hypotheses based on real-world language data (Bennet, 2010).

In conclusion, the significance of qualitative approaches in corpus linguistics rests in their capacity to reveal the underlying complexities of language, thereby fostering a more

comprehensive and contextual understanding of linguistic phenomena. Together, they contribute to an exhaustive investigation of language in its natural environment.

The Distinction Between Quantitative and Qualitative Approaches in the Context of Corpus Linguistics

As mentioned in the previous section, we can assume that the distinction between quantitative and qualitative approaches in research lies in their methods and goals. Quantitative corpus research gathers and analyzes numerical data to identify patterns, relationships, frequency, and statistical significance. It often involves surveys, experiments, and large-scale data collection, aiming for generalizability and objectivity. Lakoff (1987, as cited by Lewandowska-Tomaszczyk, 2007) states that a qualitative approach is proposed when we need to study the four types of Commensurability criteria, they are (1) truth-conditional criteria (classical translatability); (2) criteria of use; (3) framing criteria; and (4) conceptual organization criteria.

In contrast, qualitative corpus research explores a phenomenon's depth, context, and richness. Quantitative linguistic criteria consist of the: (1) frequencies: (i) in general language, (ii) in a context-specific language variety, (2) quantitative distributional facts; (3) sentence length; (4) type/token ratio; (5) lexical density (low frequency-high frequency); and (6) naturalness (frequency and contextual preferences) (Lewandowska-Tomaszczyk, 2007). It employs interviews, observations, and content analysis to gather non-numerical, textual, or visual data. Qualitative research seeks to understand the 'how' and 'why' behind human behavior, emphasizing interpretation, meaning, and context.

To choose what method is suitable for our study, we can go back clearly and pay attention to the research question of our study and what type of data we need. While quantitative research provides statistical rigor and generalisability, qualitative research offers insight into the intricacies of human experience, allowing researchers to uncover subjective perspectives and cultural nuances, word meanings based on intonation and speech, as well as by genre and environment. Often, by combining quantitative and qualitative methods, a mixed methods approach can provide a more comprehensive understanding of complex phenomena.

The Benefits of Qualitative Approaches in the Context of Corpus Linguistics

Qualitative approaches are valuable in corpus linguistics studies for several reasons:

- a. Contextual Insights: Qualitative methods permit researchers to investigate language in its natural context, revealing the complexities of how words, phrases, and grammatical structures are used in various texts and spoken discourse. This contextual awareness is essential for comprehending language's meaning and ramifications (Hardt-Mautner, 1995).

- b. Nuanced analysis: Qualitative methods enable linguists to analyze language at a granular level, revealing subtle nuances in meaning, connotation, and pragmatic application. This level of specificity is frequently difficult to attain through quantitative analysis alone (Adamou, 2019).
- c. Pragmatic interpretation: Qualitative analysis is particularly effective for interpreting pragmatic aspects of language, such as tone, intention, and speaker attitudes. Researchers can determine how sarcasm, politeness, and other subtle pragmatic signals are communicated through language (Rühlemann & Clancy, 2018).
- d. Uncovering patterns: By engaging in qualitative analysis, linguists can identify patterns that may not be readily apparent in quantitative data. These patterns may include idiosyncratic usages, stylistic preferences, and context-specific language variants (Hasko, 2012).
- e. Rich data exploration: Qualitative methods facilitate rich data exploration, allowing researchers to interpret 'how' and 'why' linguistic phenomena. This in-depth investigation frequently results in broader language comprehension (Hasko, 2012).
- f. Interdisciplinary insights: Qualitative corpus linguistics frequently intersects with other disciplines, such as sociolinguistics, discourse analysis, and stylistics, resulting in insights from multiple disciplines. This cross-pollination enriches the analysis by introducing a variety of perspectives and methods (Lewandowska-Tomaszczyk, 2007).
- g. Generation of Hypotheses: Qualitative analysis is essential in developing hypotheses and research questions. It can help linguists develop specific hypotheses to evaluate with larger datasets, guiding subsequent quantitative research (Partington & Marchi, 2015).
- h. Cultural and Societal Context: Qualitative approaches in corpus linguistics permit language study in various cultural and societal contexts. Researchers can determine how language reflects and influences cultural and social practices (Taylor & Marchi, 2018).
- i. Fine-Grained Research Qualitative analysis can reveal subtle variations in language, even within particular dialects or registers. This level of specificity is essential for a comprehensive analysis of language usage (Hasko, 2012).
- j. Linguistics discovery: Qualitative approaches frequently result in linguistic discoveries, such as discovering previously undocumented language phenomena, emergent neologisms, or shifts in language usage (Prasetya et al., 2020).

In conclusion, qualitative approaches in corpus linguistics extend the field's capabilities beyond statistical analysis, providing a more comprehensive and context-rich understanding of language in real-world settings.

BUILDING AND ANNOTATING CORPORA

The Process of Creating a Corpus for Qualitative Analysis

Creating a corpus for qualitative analysis is an organized and systematic endeavor involving several vital steps. The steps arranged here are according to the author's reading for several articles and books (Ahsanuddin et al., 2022; Coxhead, 2018; Goźdz-Roszkowski, 2021; Mcenery & Hardie, 2012; Nation, 2016; O'Keeffe & McCarthy, 2010; Toriida, 2016), they are:

- a. **Defining Research Objectives:** The process begins with clearly defining the research objectives and questions. Researchers must establish what they aim to explore or understand through the corpus.
- b. **Text Selection:** Selecting appropriate texts is critical. Researchers must choose texts that align with the research focus. This may involve decisions about sources, such as written documents, spoken language transcripts, online discussions, or a combination. In this step, we have to decide on some features to support our corpus; they are (1) text types (be it written or spoken text), (2) Geographical divisions (the distinction between US and UK English), (3) Age-related material (lists will be intended for young children, teenager, adult, etc.), (4) Language learning situation (English is learned as a second language or as a foreign language), (5) corpus size (for token or type), and (6) Proportion of text types in a corpus (how much the amount of written or spoken text will be).
- c. **Data Collection:** Data collection methods can vary depending on the sources selected. Researchers might manually transcribe spoken language, extract text from digital sources, or curate existing written texts. Data should be collected systematically to maintain consistency.
- d. **Corpus Compilation and Text Preprocessing:** Once the data is collected, it must be organized and compiled into a corpus. This includes structuring the data to facilitate analysis, often with attention to details like file formats, metadata, and organization into sub-corpora. Here, we can compile manually or assisted by software. There are so many kinds of software we can use, like Anthony's software (<https://www.laurenceanthony.net/software.html>), Range (<https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs>), WordSmith's tool (<https://www.lexically.net/wordsmith/downloads/>), or Sketch Engine (<https://www.sketchengine.eu/>). Then, the preprocessing steps begin to proofread

the text after analysis. This can involve removing irrelevant information, standardizing formats, and tokenizing (splitting text into words or phrases).

- e. **Annotation:** Annotation is a crucial step in the process. Researchers add metadata and linguistic information to the corpus, such as part-of-speech tags, syntactic structures, semantic categories, and other linguistic features. Annotation enhances the corpus's usability and assists in qualitative analysis.
- f. **Quality Control:** Rigorous quality control is essential to ensure the accuracy and reliability of the corpus. This includes verifying transcription accuracy, resolving annotation discrepancies, and ensuring the corpus's consistency.
- g. **Validation:** After the corpus goes through several stages, validation becomes essential to produce an accurate and valid corpus. Validation can be done by asking for expert help to see the relevance of the word list to the field of research focus. Some researchers use the semantic scale by [Dang \(2020\)](#) or directly consult with experts in the field of corpus studies.
- h. **Documentation:** Documenting the corpus creation process is essential for transparency and replicability. Researchers should maintain records of data sources, data cleaning procedures, annotation guidelines, and any decisions made during corpus compilation.
- i. **Storage and Access:** The corpus should be stored in a format that allows easy access and analysis. Digital repositories or corpus management systems are often used to organize and manage the corpus effectively.
- j. **Ethical Considerations:** Researchers must address ethical considerations, such as obtaining necessary permissions for data use, ensuring participant privacy, and adhering to copyright and data protection regulations.

Creating a qualitative corpus requires careful planning, attention to detail, and adherence to best practices. It provides the foundation for in-depth linguistic analysis, enabling researchers to explore the nuances and complexities of language use within their chosen context.

The Considerations for Corpus Selection and Compilation

When conducting qualitative corpus analysis, careful consideration of corpus selection and compilation is paramount to the success of the research. Several vital considerations ([Biber et al., 2010](#); [Coxhead, 2018](#); [Nation, 2016](#); [Poole, 2018](#); [Timmis, 2015](#)) should be addressed:

- a. **Representativeness:** The selected corpus should represent the language or context under investigation. Researchers must ensure that the corpus reflects the language's

diversity and nuances. For example, in sociolinguistic research, it's essential to include data from various social groups and regions to capture linguistic variation.

- b. **Balance between Diversity and Specificity:** Achieving a balance between corpus diversity and specificity is crucial. While a diverse corpus is valuable for capturing a wide range of language use, a corpus that is too broad may dilute the study's focus. Researchers need to determine the appropriate scope based on their research objectives.
- c. **Source and Genre Selection:** Decisions regarding the sources and genres included in the corpus are pivotal. Researchers should consider the appropriateness of text sources, such as newspapers, interviews, literary works, or social media posts. The selected sources should align with the research questions and the linguistic phenomena of interest.
- d. **Temporal Considerations:** Temporal considerations are essential for historical linguistic research. Researchers should decide on the corpus's periods, considering linguistic changes and diachronic analysis objectives. For contemporary linguistic analysis, recent data is often preferred to reflect current language usage.
- e. **Ethical and Legal Aspects:** Ethical considerations, including data privacy, informed consent, and adherence to copyright laws, must be addressed during corpus compilation. Researchers should ensure they have the necessary permissions to use and share the data.
- f. **Annotation Guidelines:** If the linguistic annotation is part of the process, clear guidelines should be established. These guidelines define how linguistic features will be marked in the corpus and provide consistency in the data, ensuring that it can be used effectively for qualitative analysis.
- g. **Corpus Size:** The size of the corpus is an important consideration. Researchers must determine the corpus's size based on the research objectives and available resources. A smaller, well-annotated corpus may be more effective for in-depth analysis, while a larger corpus may offer broader insights.
- h. **Sample Size:** Determining the appropriate sample size is critical for qualitative analysis. Researchers should decide how many texts or instances to include in the analysis, balancing the need for depth and richness of data with practical constraints.

By addressing these considerations, researchers can compile a corpus that is well-suited to qualitative analysis, facilitating a more in-depth exploration of the linguistic phenomena under investigation

QUALITATIVE ANALYSIS TECHNIQUES IN CORPUS LINGUISTICS

In corpus linguistics, qualitative analysis methods play a pivotal role in uncovering the depth and richness of language use. Researchers have developed various techniques to explore linguistic phenomena qualitatively. Here are some of the qualitative analysis methods commonly used:

- a. **Discourse Analysis:** Discourse analysis focuses on the structure and organization of language in context. Researchers examine how language constructs meaning within conversations, texts, or spoken discourse. This method helps uncover rhetorical strategies, speech acts, and patterns of communication (Mahlberg, 2014; Mautner, 2009; Partington & Marchi, 2015).
- b. **Content Analysis:** Content analysis is employed to examine textual data content systematically. Researchers categorize and analyze the occurrence of specific words, themes, or linguistic features within the corpus. This method often identifies recurring patterns or themes in language (Kreuter & Kreuter, 2021).
- c. **Thematic Analysis:** Thematic analysis involves identifying and analyzing recurring themes and patterns within the corpus. Researchers categorize data into themes or categories to understand the underlying structures of the language and the topics being discussed (Vaismoradi et al., 2016).
- d. **Pragmatic Analysis:** Pragmatic analysis delves into the use of language in context, focusing on speech acts, implicatures, and the intended meaning behind utterances. Researchers aim to uncover how language functions in real-life interactions (Jucker, 2013; Rühlemann & Clancy, 2018; Staples & Fernández, 2019).
- e. **Qualitative Coding:** Qualitative coding involves systematically labeling and categorizing text segments based on specific criteria. Researchers use coding to identify patterns, themes, or relationships within the corpus. This method is often used in thematic and content analysis (Biber et al., 2010).
- f. **Stylistic Analysis:** Stylistic analysis explores the linguistic choices made by authors or speakers, examining elements like figurative language, tone, register, and rhetorical devices. Researchers aim to understand how style contributes to meaning (Mahlberg, 2017).
- g. **Critical Discourse Analysis (CDA):** Critical discourse analysis focuses on the societal and power structures that underlie language use. Researchers examine how language perpetuates or challenges social norms, ideologies, and power relationships (Cheng, 2013; Hardt-Mautner, 1995; Mautner, 2009).

- h. **Frame Analysis:** Frame analysis investigates how language constructs frames or interpretative schemata for understanding events, issues, or concepts. Researchers examine how language shapes perception and frames public discourse (Davis, 1995).
- i. **Narrative Analysis:** Narrative analysis explores the structure and content of narratives within the corpus. Researchers examine how stories are constructed, identifying narrative elements and storytelling techniques (Cheng, 2014).
- j. **Conversation Analysis:** Conversation analysis studies the organization and structure of spoken or written conversations. Researchers investigate turn-taking, repair sequences, and other conversational features to understand how interactions negotiate meaning.
- k. **Lexical Semantics:** Lexical semantics involves qualitatively exploring word meanings, connotations, and semantic relations within the corpus. Researchers uncover the nuances of word usage, such as polysemy, metaphor, and metonymy (Geeraerts, 2019).

These qualitative analysis methods offer a diverse toolkit for corpus linguists to explore and interpret language in various forms and contexts. The choice of method depends on the research objectives and the nature of the linguistic data being analyzed.

CONCLUSION

This article comprehensively examines corpus linguistics's core principles and evolutionary path within its specific field. The discipline of corpus linguistics focuses on systematically collecting significant digitalized texts to study language. The all-encompassing concept of a corpus encompasses diverse research endeavors to portray and comprehend many aspects of language usage. The field has evolved, transitioning from rudimentary concordances in the 1950s to sophisticated computational tools and expansive digital corpora in the present period. The field of linguistics has integrated it as a fundamental element, yielding practical observations regarding patterns, diversity, and changes within language.

The article highlights the main objectives of corpus linguistics, which entail utilizing qualitative research methods. Qualitative analysis explores the nuanced contextual elements of language usage, delving into the underlying mechanisms and rationales contributing to manifesting linguistic phenomena in various real-world situations. The qualitative methodology offers a thorough understanding of language within its contextual framework, delving into the complexities of semantics, pragmatic acts, and the underlying intentions of speakers that cannot be fully captured through quantitative analysis alone. These approaches facilitate a comprehensive understanding of language.

Furthermore, the article outlines the systematic process of creating a corpus for qualitative analysis. This systematic undertaking involves formulating explicit research objectives, selecting texts that align with the study's goals, collecting and organizing data, annotating the corpus, ensuring quality, verifying findings, documenting processes, and addressing ethical considerations. A robust foundation is formed to undertake comprehensive linguistic analysis by meticulously adhering to these processes. The essay additionally examines the variables implicated in choosing and assembling a corpus. It emphasizes guaranteeing its representativeness, picking suitable sources and genres, considering temporal elements, addressing ethical considerations, and defining the corpus size. Researchers adhere to these standards while constructing corpora that align with their research objectives.

The article offers valuable insights into the many qualitative analysis approaches employed in corpus linguistics. A range of approaches can be employed for analysis, encompassing discourse analysis, thematic analysis, content analysis, and stylistic analysis. Each technique provides a unique framework for analyzing and understanding the intricacies of language in different contexts. This essay emphasizes the significance of qualitative approaches within corpus linguistics. These approaches thoroughly investigate language's intricacies and nuances, enabling a thorough understanding of linguistic phenomena.

ACKNOWLEDGEMENT

No funding bodies associated with this study

CONFLICT OF INTEREST

No conflict of interest reported by the author(s)

REFERENCES


- Absalom, M. (2021). Digital corpora: Language teaching and learning in the age of big data. In *Innovative language pedagogy report* (pp. 97–101). Research-publishing.net. <https://doi.org/10.14705/rpnet.2021.50.1242>
- Adamou, E. (2019). Corpus linguistic methods. In *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)* (pp. 638–653). De Gruyter. <https://doi.org/10.1515/9783110435351-052>
- Ahsanuddin, M., Hanafi, Y., Basthomi, Y., Taufiqurrahman, F., Bukhori, H. A., Samodra, J., Widiati, U., & Wijayati, P. H. (2022). Building a corpus-based academic vocabulary list of four languages. *Pegem Journal of Education and Instruction*, 12(1), 159–167. <https://doi.org/10.47750/pegegog.12.01.15>
- Aijmer, K., & Altenberg, B. (2014). *English corpus linguistics*. Routledge. <https://doi.org/10.4324/9781315845890>


- Bennet, G. R. (2010). Using corpora in the language learning classroom: Corpus linguistics for teachers. *Language Value*, 2, 140–143.
- Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed, pp. 548–570). Oxford University Press.
- Cheng, W. (2013). Corpus-based linguistic approaches to critical discourse analysis. *The Encyclopedia of Applied Linguistics*, 1353–1360.
- Cheng, W. (2014). Corpus analyses of professional discourse. In *The Routledge handbook of language and professional communication* (pp. 13–25). Routledge.
- Conrad, S. (2004). *Corpus variety: Corpus linguistics, language variation, and language teaching* (pp. 67–85). <https://doi.org/10.1075/scl.12.08con>
- Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives* (First Edit). Routledge.
- Coxhead, A., & Nation, P. (2001). *The specialized vocabulary of English for academic purposes*. Cambridge University Press. <https://doi.org/https://doi.org/10.1017/CBO9781139524766.020>
- Crawford, W. J., & Csomay, E. (2016). *Doing corpus linguistics* (First publ). Routledge.
- Dang, T. N. Y. (2020). Corpus-based list in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 288–303). Routledge
- Davis, K. A. (1995). Qualitative theory and methods in applied linguistics research. *Tesol Quarterly*, 29(3), 427–453.
- Flowerdew, J. (2009). Corpora in language teaching. In *The Handbook of Language Teaching* (pp. 327–350). Wiley-Blackwell. <https://doi.org/10.1002/9781444315783.ch19>
- Friginal, E. (2018). *Corpus linguistics for English teachers: New tools, online resources, and classroom activities* (First publ). Routledge.
- Fuertes-Olivera, P. A. (2018). *The Routledge handbook of lexicography*. Routledge.
- Gavioli, L. (2005). *Exploring Corpora for ESP Learning*. John Benjamins Publishing Company.
- Geeraerts, D. (2019). Lexical semantics. *Cognitive Linguistics-A Survey of Linguistic Subfields*, 15, 25.
- Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, 34(5), 1515–1540. <https://doi.org/10.1007/s11196-021-09860-8>
- Hardt-Mautner, G. (1995). “Only connect”. *Critical discourse analysis and corpus linguistics*. Citeseer.


- Hasko, V. (2012). Qualitative corpus analysis. In *The Encyclopedia of Applied Linguistics*. Wiley. <https://doi.org/10.1002/9781405198431.wbeal0974>
- Hoey, M., Mahlberg, M., Stubbs, M., & Teubert, W. (2007). *Text, discourse and corpora: Theory and analysis*. Continuum.
- Hunston, S., & Su, H. (2019). Patterns, constructions, and local grammar: A case study of 'evaluation.' *Applied Linguistics*, 40(4), 567–593. <https://doi.org/10.1093/applin/amx046>
- Illie, C., & Norrick, N. R. (2018). *Pragmatics and its Interfaces*. John Benjamins Publishing Company.
- Jucker, A. H. (2013). Corpus pragmatics. *Handbook of Pragmatics*, 17, 1–18.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. Pearson Education Limited.
- Kreuter, J., & Kreuter, J. (2021). The tools for empirical analysis – The method of qualitative content analysis. *Climate Engineering as an Instance of Politicization: Talking Tomorrow's Technology – Framing Political Choice?*, 165–186.
- Lewandowska-Tomaszczyk, B. (2007). Cognitive corpus studies: A new qualitative & quantitative agenda for contrasting languages. *MFU CONNEXION*, 1(1), 1–26.
- Ludeling, A., & Kyto, M. (2009). *Corpus linguistics an international handbook* (Vol. 2). Walter de Gruyter.
- Mahlberg, M. (2014). Corpus linguistics and discourse analysis. *Pragmatics of Discourse*, 215–238.
- Mahlberg, M. (2017). Corpus stylistics. In *The Routledge handbook of stylistics* (pp. 378–392). Routledge.
- Mautner, G. (2009). Corpora and critical discourse analysis. *Contemporary Corpus Linguistics*, 32–46.
- McCarthy, M., & O'Keeffe, A. (2010). Historical perspective what are corpora and how have they evolved? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3–13). Routledge.
- Mcenery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Mitkov, R. (2022). *The Oxford handbook of computational linguistics*. Oxford University Press.
- Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.208>
- O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics* (First Edition). Routledge. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs>

- O’Keeffe, A., McCarthy, M. J., & Carter, R. A. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.
- Partington, A., Duguid, A., & Charlotte, T. (2013). *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins.
- Partington, A., & Marchi, A. (2015). Using corpora in discourse analysis. In *The Cambridge handbook of English corpus linguistics* (Vol. 1, pp. 216–234). Cambridge University Press.
- Poole, R. (2018). *A guide to using corpora for English language learners*. Edinburgh University Press Ltd.
- Prasetya, E. P., Ekawati, A. D., Nugraha, D. S., Marzuq, A., & Darlis, T. S. (2020). Corpus linguistics, language corpora and language teaching. *English Journal*, 14(2), 75–83.
- Rühlemann, C., & Clancy, B. (2018). Corpus linguistics and pragmatics. *Pragmatics and Its Interfaces*, 241–266.
- Sinclair, J. (1991). *Corpus, concordance, collocation* (Describing). Oxford University Press.
- Staples, S., & Fernández, J. (2019). Corpus linguistics approaches to L2 pragmatics research. *The Routledge Handbook of Second Language Acquisition and Pragmatics*, 241–254.
- Taylor, C., & Marchi, A. (2018). *Corpus approaches to discourse: A critical review* (1st Edition). Routledge. <https://doi.org/https://doi.org/10.4324/9781315179346>
- Timmis, I. (2015). *Corpus linguistics for ELT* (First publ). Routledge.
- Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 14–27). Routledge.
- Toriida, M.-C. (2016). Steps for creating a specialized corpus and developing an annotated frequency-based vocabulary list. *TESL Canada Journal/Revue TESL Du Canada*, 34(11), 87–105. <https://doi.org/http://dx.doi.org/1018806/tesl.v34i1.1255>
- Vaismoradi, M., Jones, J., Turunen, H., & Snelgrove, S. (2016). *Theme development in qualitative content analysis and thematic analysis*.

ABOUT THE AUTHORS

	Author's Biography			
	Name	Misnawati		
	Affiliation	Universitas Cahaya Prima		
	Email	<i>misnawati@uncapi.ac.id</i>	Scopus ID	57211372221
	Google Scholar	<i>https://bit.ly/3SJVNSP</i>	WoS ID	-
	Research Interest	<i>She is a lecturer at Universitas Cahaya Prima. Her research interest falls within the area of teaching and learning method, MALL, Corpus linguistics</i>		

	Author's Biography			
	Name	Sahril Nur		
	Affiliation	Universitas Negeri Makassar		
	Email	<i>sahrilfbsumm@unm.ac.id</i>	Scopus ID	57214247081
	Google Scholar	<i>https://bit.ly/3Sq0yj5</i>	WoS ID	-
	Research Interest	<i>Sahril Nur is a full professor in English Department, Universitas Negeri Makassar. His research interest includes learning psychology, learners' anxiety, motivation, and technology-enhanced language learning</i>		

	Author's Biography			
	Name	Saidna Zulfiqar bin Tahir		
	Affiliation	Universitas Iqra Buru		
	Email	<i>saidnazulfiqar@yahoo.com</i>	Scopus ID	57195480233
	Google Scholar	<i>https://bit.ly/3I4ouRf</i>	WoS ID	-
	Research Interest	<i>Saidna Zulfiqar bin Tahir is a lecturer at Universitas Iqra Buru. His research interest includes translanguaging, Multilingual, Curriculum and Material Development</i>		