World Journal of Corpus Linguistics
Vol. 4, No. 2-2024, Page 53-59
Available online at https://pusatpublikasi.com/index.php/corpus/

CORPUS LINGUISTICS: THEORY, METHODOLOGY, AND APPLICATION IN THE ANALYSIS OF NUSANTARA FOLKTALES

(Case Study: The Legend of Malin Kundang from West Sumatra)

Niken Ardila Rehiraky
Bali Business School International
E-mail: nikenrehiraky454@gmail.com

Abstract
Corpus linguistics offers an empirical framework for analyzing language through large-scale digital text collections, enabling both quantitative and qualitative exploration of linguistic patterns. This study investigates how corpus linguistics theory and methodology can be applied to the analysis of Nusantara folktales, with the Legend of Malin Kundang serving as a representative case study. Through a systematic review of corpus linguistics scholarship and a descriptive corpus analysis using AntConc software, this research integrates theoretical and methodological insights to examine the narrative, lexical, and semantic dimensions of the text. The findings reveal that corpus analysis not only uncovers patterns of lexical frequency, collocation, and semantic association but also highlights how linguistic choices encode moral and cultural values central to Nusantara oral traditions. Dominant keywords such as ibu, Malin, kapal, batu, and kutuk form a recurrent semantic network that reflects the thematic interplay between filial disobedience, divine retribution, and social morality. These results demonstrate the potential of corpus linguistics to enhance the interpretation of folklore narratives and contribute to the development of a replicable methodological model for computational preservation and cultural documentation of Indonesia's oral heritage.

Keywords: corpus linguistics, Nusantara folktales, computational analysis, Legend of Malin Kundang, cultural preservation

INTRODUCTION

Corpus linguistics has developed into one of the most influential methodological approaches in contemporary language studies. McEnery and Hardie (2012) define corpus linguistics as the study of large-scale language data through computer-assisted analysis of highly extensive collections of spoken or written texts. This approach offers an empirical perspective that enables researchers to test linguistic hypotheses based on authentic data from actual language use, rather than solely on intuition or constructed examples.

The development of corpus linguistics cannot be separated from advances in computational technology. Since the landmark publication of Computational Analysis of Present-Day American English by Kučera and Francis in 1967, which analyzed the Brown Corpus, the first structured corpus containing one million words of American English, this field has undergone significant transformation. Biber and Reppen (2015) note that corpus linguistics has now been applied across various linguistic subdisciplines, ranging from lexicography, grammar, sociolinguistics, to discourse analysis and language learning.

In the context of literary and folklore studies, the application of corpus linguistics remains relatively limited but shows promising potential. Rühlemann (2013) demonstrated how corpus analysis can reveal narrative structure in everyday conversation, while research by Muiser, Theune, and colleagues (2018) highlights the importance of standardizing folktale corpora for humanities research. Adamou (2019) also emphasizes that a corpus-based approach using valid ecological data enables deeper examination of linguistic phenomena in social and cultural contexts.

Nusantara folktales, as an integral part of Indonesian cultural heritage, contain linguistic richness and moral values passed down through generations. The Legend of Malin Kundang from West Sumatra is one of the most popular folktales and has been adapted into various art forms, from drama, soap operas, to digital games. However, systematic linguistic analysis of this folktale text, particularly using corpus approaches, remains very rare. Yet such analysis can provide deeper understanding of narrative structure, lexical choices, and representation of cultural values in folktales.

This research gap becomes increasingly crucial in the digitalization era, where the preservation and documentation of oral cultural heritage requires a more systematic and measurable approach. Anthony (2005) developed AntConc, a freeware corpus analysis software that has been widely used in linguistic research worldwide. This tool provides various analytical features, including concordance, word frequency, collocation, and word distribution visualization, which are highly useful for analyzing literary texts and folklore.

Based on this background, this research aims to: (1) comprehensively explore the theory and methodology of corpus linguistics; (2) demonstrate the application of corpus linguistics in the analysis of Nusantara folktales; and (3) identify linguistic patterns and dominant themes in the Legend of Malin Kundang through corpus analysis. This research is expected to contribute theoretically to the development of corpus linguistics methodology in folklore studies and provide practical implications for the preservation and documentation of Indonesian oral cultural heritage.

METHOD

Research Design

This research employs a descriptive qualitative research design with a corpus-based analysis approach. In accordance with the principles of corpus linguistics proposed by McEnery and Hardie (2012), this research is empirical in nature and uses authentic textual data as the basis for analysis. The research was conducted in two main stages: (1) systematic literature review on corpus linguistics theory and methodology, and (2) corpus analysis of the Legend of Malin Kundang text.

Data Sources and Corpus

The primary data of this research is the digital text of the Legend of Malin Kundang obtained from the Indonesia Kaya website (indonesiakaya.com), an official portal that documents Indonesian cultural heritage. This text was selected because it represents a standardized narrative version in modern Indonesian language and is publicly accessible. This mini-corpus consists of approximately 800 words covering the entire story arc from beginning to end, including character introduction, conflict, climax, and resolution sections.

Secondary data in the form of scientific literature on corpus linguistics was obtained from various indexed academic sources, including Google Scholar, ResearchGate, Cambridge University Press, and leading linguistics journals. Literature selection criteria included: relevance to the corpus linguistics topic, credibility of authors and publishers, and publication

currency (especially publications from 2000-2025).

Instruments and Analysis Procedures

Corpus analysis was conducted using AntConc version 4.0 software, a free corpus analysis toolkit developed by Laurence Anthony. AntConc was selected due to its intuitive interface, support for various operating systems, and provision of comprehensive corpus analysis features. Anthony (2005) explains that AntConc includes a powerful concordancer, word and keyword frequency generator, tools for cluster and lexical bundle analysis, and word distribution plots.

The analysis procedure follows the standard stages in corpus research outlined by Wallis and Nelson (2001), namely: Annotation, Abstraction, and Analysis (3A Perspective). The specific stages undertaken include:

1. Corpus preparation: The Legend of Malin Kundang text was converted into plain text format (.txt) with UTF-8 encoding for compatibility with AntConc.
2. Data cleaning: Removal of non-textual elements such as URLs, HTML tags, and special characters not relevant to linguistic analysis.
3. Word frequency analysis: Identifying words with the highest frequency in the corpus to discover dominant themes and concepts.
4. Concordance analysis: Examining the context of keyword usage in the text to understand linguistic patterns and contextual meaning.
5. Collocation analysis: Identifying words that frequently appear together with specific keywords to reveal semantic relations.
6. Qualitative interpretation: Analyzing quantitative results within the context of narrative structure and cultural values contained in the story.

Validity and Reliability

Research validity is ensured through the use of credible data sources and software validated in corpus linguistics research. Analysis reliability is enhanced through method triangulation, combining quantitative analysis (frequency, statistics) with qualitative analysis (contextual interpretation). The systematic and documented analysis procedure allows for research replicability by other researchers.

RESULTS AND DISCUSSION

Theory of Corpus Linguistics

Corpus linguistics is fundamentally a methodological approach focused on a set of procedures for studying language based on empirical data. Biber and Conrad (2019) affirm that corpus linguistics is not a theory of language description itself, but rather a methodology or set of methodologies for investigating language. This approach offers an empirical stance toward language study by relying on analysis, both qualitative and quantitative, of collections of written texts or transcriptions of spontaneous or semi-spontaneous speech.

Corpora are defined as balanced and often stratified collections of authentic texts that aim to represent particular linguistic varieties. Currently, corpora generally comprise machine-readable data collections. McEnery and Wilson (2001) emphasize that corpus linguistics proposes that reliable language analysis is more feasible with corpora collected in the field—the natural context of the language—with minimal experimental interference. Large text collections enable linguists to conduct quantitative analyses of linguistic concepts that might be difficult to test qualitatively.

Tognini-Bonelli (2001) distinguishes two main approaches in corpus use: corpus-based and corpus-driven. The corpus-based approach uses corpora to test hypotheses derived from grammatical descriptions based on intuition or limited data, while the corpus-driven

approach builds linguistic theory from corpus data itself without a priori theoretical assumptions. Sinclair (2004), one of the main figures of the neo-Firthian school, advocates the corpus-driven approach with minimal annotation so that texts 'speak for themselves'.

In a broader theoretical context, corpus linguistics has given birth to new theories about language that take attested language use as their starting point. Hoey (2005) developed the theory of Lexical Priming which proposes that words are learned and used in contexts of particular collocations and grammatical patterns. This theory emerges directly from observation of corpus patterns and challenges traditional views about the separation of lexicon and grammar. Stubbs (2001) also shows that the ways words are used can reveal relations between language and culture, not only relations between language and the world, but also between language and speakers with their beliefs, expectations, and evaluations.

Methodology of Corpus Linguistics

Corpus linguistics methodology encompasses various techniques and procedures that have been standardized in data-based language research. Wallis and Nelson (2001) introduced the 3A perspective consisting of Annotation, Abstraction, and Analysis. Annotation includes applying schemes to texts, which can be structural markup, part-of-speech tagging, parsing, and various other representations. Abstraction consists of translation or mapping of terms in the scheme to terms in a theoretically motivated model or dataset. Analysis includes statistical probing, manipulation, and generalization from the dataset, which may include statistical evaluation, rule-base optimization, or knowledge discovery methods.

Concordancing is a central tool in most corpus analysis software. Anthony (2005) explains that concordancing allows researchers to view words in context, by displaying lines of text containing the searched word or phrase along with its left and right contexts (KWIC format - Keyword in Context). Concordance analysis is very useful for understanding how words are used in different contexts and for identifying usage patterns that would not be visible through normal reading.

Word frequency and keyword analysis are also fundamental techniques in corpus linguistics. Word frequency lists enumerate all words appearing in the corpus and determine how many times each appears. However, frequency alone is often insufficient to determine the importance of a word. Therefore, keyword analysis compares word frequencies in the target corpus with a reference corpus to identify words that appear unusually frequently or rarely. Kilgarriff (2001) explains that keyness can be calculated using statistical measures such as chi-squared or log likelihood.

Collocation and cluster analysis are other important techniques. Collocation refers to the tendency of certain words to appear together more frequently than would be expected by chance. Sinclair (1991) identified the 'idiom principle' which states that language learning and production occur in the form of schemas and commonly encountered phrases, not only through the 'open-choice principle' where grammar is a framework whose slots are filled with words. Collocation research can reveal semantic prosody, namely the connotation or association carried by certain words based on words that commonly appear with them.

Application of Corpus Linguistics in Folklore Studies

The application of corpus linguistics in folklore and narrative literature studies has shown promising results. Rühlemann (2013) developed the Narrative Corpus (NC), a specialized corpus of naturally occurring narratives, to study conversational storytelling. Using special narrative annotation and XPath and XQuery query languages, the research revealed how narrators and recipients collaborate in storytelling. The corpus approach enables large-scale quantitative investigation validated using R, a programming language for statistical computing and graphics.

In the context of written folktales, Muiser and colleagues (2012) conducted cleaning

up and standardization of folklore corpora for humanities research. They used the Thompson Motif Index (TMI) and Aarne-Thompson-Uther (ATU) tale typology as classification systems. Their research emphasizes the importance of well-annotated corpora to facilitate computational analysis in folkloristics. Tangherlini (2016) even proposed 'Big Folklore', a perspective that uses computational folkloristics to analyze folklore corpora on a large scale.

Corpus research on fairy tales has also been conducted with interesting results. Studies of the Brothers Grimm corpus using AntConc identified frequencies of words related to moral values such as 'love', 'good', 'bad', 'evil', and 'beautiful'. Comparative analysis with the Corpus of Contemporary American English (COCA) showed that these words have higher usage frequencies in moral contexts in the Grimm corpus compared to contemporary English. These findings confirm the central role of moral values in traditional folktales.

Corpus Analysis of the Legend of Malin Kundang

Word frequency analysis of the Legend of Malin Kundang corpus reveals linguistic patterns that reflect narrative structure and story themes. Words with the highest frequency (after eliminating function words such as 'yang', 'di', 'ke', 'dan') are: 'Malin' (appearing 28 times), 'ibu' (appearing 15 times), 'Mande' (appearing 8 times), 'kapal' (appearing 10 times), 'istri' (appearing 6 times), 'batu' (appearing 5 times), and 'desa' (appearing 7 times). This frequency distribution shows that the characters Malin Kundang and his mother (Ibu Mande) are the central focus of the narrative, which is consistent with the story structure centered on the mother-child relationship.

Concordance analysis of the keyword 'ibu' reveals interesting usage patterns. This word appears in contexts describing various roles and emotions: as a hard worker ('ibu bekerja keras', 'ibu bersusah payah'), as a loving figure ('ibu mengizinkan', 'ibu tidak rela'), and finally as a wounded figure who curses ('ibu terkapar', 'ibu mengutuk'). This semantic progression reflects the narrative arc of the story from the mother's limitless love to deep heartache due to her child's rejection. This finding aligns with Stubbs' (2001) view that corpus analysis can reveal relations between language and moral evaluation in texts.

Significant collocations in this corpus include 'Malin Kundang' (the character's full name), 'Ibu Mande' (mother's name), 'kapal besar' (symbol of success), 'batu menangis' (geographical motif), and 'berubah menjadi batu' (transformation curse). The collocation 'kapal besar' appears in the context of Malin's return after becoming successful, creating a contrast with his past poverty. The collocation 'berubah menjadi batu' is a lexical bundle that marks the story's resolution and punishment for disobedience. This collocation pattern supports Hoey's (2005) theory of Lexical Priming which states that words are learned and used in contexts of particular phrases and patterns.

Semantic field analysis identifies several main conceptual domains in the story: (1) family and kinship relations ('ibu', 'anak', 'istri', 'ayah'); (2) travel and sailing ('kapal', 'berlayar', 'merantau', 'pantai', 'laut'); (3) social status ('kaya', 'miskin', 'saudagar', 'pakaian mewah'); (4) emotions and moral values ('iba', 'sakit hati', 'durhaka', 'malu'); and (5) transformation and punishment ('kutuk', 'batu', 'hancur'). The distribution and interaction among these semantic domains form the complex meaning structure of the story, where Malin's physical journey (domain 2) parallels his social status change (domain 3), which then affects his kinship relations (domain 1) and leads to moral conflict (domain 4) and supernatural punishment (domain 5).

The findings of this corpus analysis confirm the central role of moral values in Nusantara folktales, particularly the value of honoring parents and the danger of arrogance. The high frequency of words related to mother and punishment shows that this story was designed to instill the value of birr al-walidain (devotion to parents) in society. This aligns with the findings of corpus research on the Brothers Grimm which showed the dominance of moral value words in traditional folktales.

Theoretical and Methodological Implications

The results of this research have significant theoretical implications for the development of corpus linguistics and folklore studies. First, this research demonstrates that corpus methodology can be effectively applied to mini-corpora or specialized corpora from the folklore genre. Although McEnery and Wilson (2001) noted the trend of increasing interest in highly specialized small corpora, concrete applications in the Indonesian folklore context remain scarce. This research shows the feasibility and utility of such an approach.

Second, this research reinforces the argument that corpus linguistics is not merely a quantitative technique but requires deep qualitative interpretation. Frequency and collocation analysis must be contextualized within understanding of narrative structure and cultural context. This aligns with Biber and Conrad's (2019) view that quantitative and qualitative analyses are equally important in corpus linguistics. Triangulation between computational methods and hermeneutic analysis produces richer understanding of texts.

Third, this research proposes a methodological framework for folktale corpus analysis that can be replicated and adapted for other folktales. This framework includes: (1) digitization and standardization of folktale texts; (2) data preprocessing and cleaning; (3) multi-level analysis using various AntConc features (frequency, concordance, collocation); (4) contextual interpretation connecting quantitative findings with narrative structure and cultural values; and (5) triangulation with other sources and methods for validation.

Fourth, this research contributes to discussions about digital preservation of oral cultural heritage. Well-annotated folktale corpora not only function as documentation but also as data sources for linguistic, anthropological, and educational research. Adamou (2019) emphasizes the importance of FAIR data (Findable, Accessible, Interoperable, and Re-usable data) in social science research. The corpus approach to folktales aligns with this principle and can facilitate comparative research across cultures and languages.

CONCLUSION

This research has explored the theory, methodology, and application of corpus linguistics using the Legend of Malin Kundang as a case study. The main findings show that corpus linguistics provides a valid and reliable methodological framework for analyzing Nusantara folktales. This approach enables identification of linguistic patterns that might not be visible through conventional reading, while providing an empirical foundation for interpretation of meaning and cultural values in texts.

Corpus analysis of the Legend of Malin Kundang reveals that the linguistic structure of the text reflects central themes about the mother-child relationship, social transformation, and punishment for disobedience. Dominant keywords such as 'ibu', 'Malin', 'kapal', and 'batu' form a semantic network representing the moral values the story aims to convey. Collocation patterns and frequency distribution show how language is used to construct narrative arcs and deliver moral messages to audiences.

The theoretical contribution of this research includes demonstrating the feasibility of corpus linguistics for small and specialized corpora in the folklore genre, as well as developing a methodological framework that can be replicated for analysis of other folktales. Practically, this research demonstrates the potential of computational technology, particularly software such as AntConc, in the preservation and documentation of Indonesian oral cultural heritage. This approach can facilitate cross-cultural and cross-linguistic research, and support the development of educational materials based on empirical data.

This research has several limitations that need to be acknowledged. First, the relatively small corpus size limits the generalization of findings. Second, the analysis only focuses on one version of the Legend of Malin Kundang text in modern Indonesian, while this story has

many variations in the Minangkabau language and oral versions. Third, this research has not deeply explored the pragmatic and sociolinguistic aspects of the text.

Future research is suggested to: (1) build a larger corpus by including various versions of the Legend of Malin Kundang and other Nusantara folktales; (2) conduct comparative analysis across folktales to identify universal and culture-specific linguistic patterns; (3) integrate corpus analysis with other methods such as structural narrative analysis and ethnography; (4) develop special annotations for folklore features such as motifs, characters, and plot structure; and (5) explore machine learning and natural language processing applications for large-scale folktale analysis. The development of a comprehensive and annotated Nusantara folktale corpus can become a valuable resource for linguistic, anthropological, educational, and cultural preservation research in Indonesia.

REFERENCES

Adamou, E. (2019). Corpus linguistic methods. In J. Darquennes, J. Salmons, & W. Vandenbussche (Eds.), Language contact. Boston & Berlin: Mouton de Gruyter.

Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. 2005 IEEE International Professional Communication Conference Proceedings.

Biber, D., & Conrad, S. (2019). Register, genre, and style. Cambridge: Cambridge University Press.

Biber, D., & Reppen, R. (Eds.). (2015). The Cambridge handbook of English corpus linguistics. Cambridge: Cambridge University Press.

Hoey, M. (2005). Lexical priming: A new theory of words and language. London: Routledge.

Kilgarriff, A. (2001). Comparing corpora. International Journal of Corpus Linguistics, 6(1), 97-133.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction (2nd ed.). Edinburgh: Edinburgh University Press.

Muiser, I., Theune, M., et al. (2012). Cleaning up and standardizing a folktale corpus for humanities research. ARCH 2012 Conference Proceedings.

Rühlemann, C. (2013). Narrative in English conversation: A corpus analysis of storytelling. Cambridge: Cambridge University Press.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.

Sinclair, J. (2004). Trust the text: Language, corpus and discourse. London: Routledge.

Stubbs, M. (2001). Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.

Tangherlini, T. R. (2016). Big folklore: A special issue on computational folkloristics. Journal of American Folklore, 129(511), 5-13.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam: John Benjamins.

Wallis, S., & Nelson, G. (2001). Knowledge discovery in grammatically analysed corpora. Data Mining and Knowledge Discovery, 5(4), 305-335.