wxds ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 05 No. 01. Jan-March 2026.Page# 3731-3742
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.21096112

The value of Corpus Linguistics for contemporary dictionary making: From Citations to Collocations

Falak Faridi
M.Phil Scholar, Department of English, University of Okara, Okara, Pakistan.
falakfaridi07@gmail.com
Muhammad Kamran Abbas Ismail (Corresponding Author)
Lecturer, Department of English, University of Okara.
Kamran.ch@uo.edu.pk
Fazeel iftikhar
M. Phil Scholar, Department of English, University of Okara, Okara, Pakistan.
fazeelgill123456@gmail.com

ABSTRACT
During the last 30 years, corpus linguistics has had a profound impact on lexicographers' work and has revolutionized the way in which dictionaries have been prepared from intuition to data. In the last 30 years, corpus linguistics has reshaped the work of the lexicographer from intuition to data. This article deals with the many aspects of dictionary making that have been influenced by the use of corpus linguistics, the use of the AntConc and LancsBox collocation analysis tools, and the use of computational methods to shape the dictionary processes of selection of headwords, writing definitions, extracting examples from the text and documenting patterns of phrases. Based on a comprehensive and extensive survey of the scholarly literature in the fields of lexicography, corpus linguistics, and computational approaches to language description, the study brings together the themes that emerged across the three main dimensions: corpus-driven methods in lexicographic work, the use of digital tools and technologies, and the creation of specialized and multilingual dictionaries. The article suggests two main advantages for the use of corpus-based methods: the enrichment of the descriptive content and the dictionary's currency, and the broadening of its pedagogical value for language learners as well as for specialised users. The results obtained from the literature survey show that the programs called AntConc and LancsBox can help lexicographers to extract statistically significant collocations and concordance lines, leading to better representative dictionaries of authentic language use. The article also mentions a gap in research with respect to the few studies on low-resource and indigenous languages that employ corpus methods. The study ends with a plea for a wider, more inclusive definition of corpus-assisted lexicography, which goes beyond the technological skills to the linguistic diversity.

Keywords: *Corpus Linguistics; Lexicography; Dictionary Making; Collocation Analysis; AntConc; LancsBox; Corpus-driven Methodology; Digital Dictionary*

1. Introduction

Since the use of computational corpus linguistics, the language - dictionary relationship has changed. Traditionally, dictionary building has been done with the help of citation slips, editorial judgment and the opinions of language scholars. This method created seminal reference works, but was, however, effectively constrained in terms of the breadth, representativeness and range of language use in contexts and registers. But with the advent of

large electronic corpora, and of powerful analytical software, the documentation of language has changed dramatically for the lexicographer.

With the advent of the corpus revolution in lexicography, a term very well coined by Krishnamurthy (2002), it has been possible to base the dictionary entries on millions of authentic instances of language in use. It is corpus linguistics that equips the lexicographer with resources for determining not only the meaning of words but also their behaviour, in the sense that it describes the other words that often co-occur with them, the syntactic structures in which they are found, and the pragmatic contexts in which they may be expected to occur. Kilgarriff and Jackson (2013) note that today's lexicographic work is guided chiefly by the corpus, in the process of choosing headwords, constructing definitions, and finding examples. Heuberger (2016) takes this one step further by claiming that corpus tools have established a 'true paradigm shift in English lexicography, affecting the way dictionaries are written, structured and developed. This shift has a practical implication in the sense that corpus data can not only inform the content of a dictionary, but also the prioritization of its entries as outlined by Walter (2010) and as detailed by Rees (2022) in his thorough description of the processes through which corpus data is converted to a publishable dictionary entry.

The main thesis of this article is that corpus linguistics has come to be an essential methodology for contemporary dictionary writing, offering unparalleled amounts of empirical precision, descriptive breadth and pedagogical pertinence. The corpus-based approach to dictionary construction is based on corpora and therefore shows actual language use and is not based on a normative ideal. And the availability of tools like AntConc and LancsBox has enabled corpus analysis to be operationalized in an efficient and reproducible manner. This article argues that without corpus methods, modern dictionaries would still not be good models of dynamic, living language.

There is a plausible counterargument, however, that corpus-driven approaches are empirically effective but also have the potential to privilege high-frequency and well-documented language variants in favor of those of low frequency, regional and minority variants. Corpus linguistics, as a field of study, has always been influenced by the corpus, and most of the bigger, multi-million-word corpora are of the mainstream, written and formal varieties of language (B Barlow, 2011). According to this perspective the corpus-based dictionary can exclude dialects, indigenous languages and specific technical registers which are not well represented in existing corpora.

The proportions of these two positions indicate that corpus linguistics can be transformative in the field of lexicography, but that its true power can only come to pass if corpora construction is not limited to the dominant languages and registers. So the question here is not really about the use of corpus methods, but about an increasing number of corpora used and the representation of linguistic diversity in the dictionaries generated by corpus data.

Although many studies have been conducted on corpus linguistics and lexicography, there is a gap in the research regarding a systematic investigation of the use of specific corpus analysis tools, such as AntConc and LancsBox, in the compilation of dictionaries, as well as the effects of these integrations on the lexicographic results for various dictionary types and target languages. Although various corpus-based studies have focused on the theoretical model of corpus lexicography or on the role of computer-based tools in lexicography, there is a lack of research that has integrated the three aspects of tool-specific analysis, dictionary genre and linguistic diversity.

The following research questions are therefore used to guide this study:

RQ1: What impact has corpus linguistics had on the method of contemporary dictionary-making, specifically in the selection of headwords, in the form of entries and in the choice of examples?
RQ2: How can corpus analysis tools like AntConc and LancsBox help identify the collocational patterns and incorporate them into lexicographic products?
RQ3: How and how much has the concept of corpus been used in the creation of specialized, multilingual and digital dictionaries and what are the challenges in expanding the same methodology to less well-represented languages?
The value of this study is that it could contribute to the theory and application of lexicography. The results can serve as a summary of the best practices in corpus-assisted lexicography for dictionary makers. The study demonstrates the usefulness of computational tools for corpus linguists' practical language documentation. The study highlights the educational significance of dictionaries based on real language data for language learners and teachers.

2. Literature Review

There are several different types of lexicographic methodology. The most prevalent is the corpus-driven approach, which is used in lexicographic practice.
The use of corpus in making dictionaries has been theorized and documented in the last few decades of research. The change stems from a belief that dictionaries must accurately represent the patterns of language use, rather than those that should be used, as defined by prescriptive criteria. Kilgarriff (2005) was one of the first writers to argue systemically for the inclusion of the corpus in the dictionary; he proposed a process in which all decisions made in the dictionary from what to include to how to label the usage are based on corpus evidence. Kilgarriff and Jackson (2013) expanded on this by describing comprehensively the use of corpora for dictionary data, from frequency to collocation through to the discovery of the meaning of words in specific contexts.
Dash and Ramamoorthy (2018) expanded the framework by looking at the real-world connection between corpus construction and dictionary making, and they showed how the choices of corpus size, genre balance, and time span have an impact on the lexicographic output generated by this corpus. The authors stress that a well-designed corpus is not just a large number of texts, but a well-designed, representative sample of language, the register, the domain and the community of users the dictionary is supposed to serve. It is strengthened by Kilgarriff, Rundell and Uí Dhonnchadha (2006), who developed a New Corpus for Ireland for the purpose of building a dictionary, which showed that the process of corpus development for lexicographic use entails making systematic choices over text selection, metadata annotation and copyright clearance.
There has been a special focus on the mechanics of corpus-assisted definition writing in the literature. Based on the corpus evidence, Summers (2014) suggests that lexicographers can include the most frequent and common meaning of a word and create definitions of words that more closely match the experiences of the users of common language. In a similar fashion, Walter (2010) show how concordance lines from large corpora can be used to capture nuances of meaning, register, and pragmatic function that otherwise would be hard or impossible to capture without access to large amounts of authentic data. Heuberger (2016) places these changes in the wider context of English lexicography and sees the shift from the citation slip towards corpus as a paradigm change that has fundamentally changed the definition of a "good" dictionary.
The automatic recognition of good dictionary examples has also been considered computationally. Kilgarriff, Husák, McAdam, Rundell, and Rychlý (2008) presented an algorithm

they called GDEX (Good Dictionary EXamples) that selects illustrative examples that are maximally informative and learner-friendly, automatically based on a set of criteria derived from a corpus. This is a typical example of the intersection of corpus linguistics and computational linguistics applied to the task of dictionary making, and it is among the most specific uses of the corpus methodology applied to dictionary making.

2.3 Oriental Lexicography

One of the major developments from which the operationalization of corpus linguistics in the field of lexicography has emerged is the development of dedicated corpus analysis software. Two of the most popular tools in contemporary research and applied lexicography are AntConc and LancsBox, which offer advanced interfaces for concordancing, frequency analysis and, most importantly, for extracting collocations for the lexicographer. Designed by Laurence Anthony, AntConc provides a set of tools such as concordance search, collocate analysis, n-gram extraction and keyword comparison between corpora. Advanced collocation and dispersion statistics are supplied by LancsBox developed at Lancaster University, including the GraphColl module to display collocational networks and highlight the most statistically significant lexical associates for any particular node word. All these, when used together, can help corpus-driven lexicography to go beyond mere frequency counts and provide a nuanced picture of the patterning of the lexical level and the functioning of phrases.

The general corpus tools and dictionary-making specific corpus tools are discussed at length by Kilgarriff and Kosem (2012). They contend that the best tools combine the qualities of providing frequency data, collocational statistics, and an interface for retrieval of examples in a single tool, thus reducing the cognitive load on the lexicographer when working under production constraints. In addition, Paquot (2012) presents another example in which integrated tools from the corpus and dictionary are used for pedagogical purposes, arguing that the LEAD dictionary-cum-writing aid is a blueprint for the development of a pedagogically useful, integrated corpus and dictionary language assistance tool for learners.

Collocation analysis is of great significance in the process of dictionary making. The tendency for some words to occur with certain others more often than by chance—called collocation—is a major organizing principle of the mental lexicon; dictionaries that don't reflect this collocational behavior do not leave users very well prepared to produce natural-sounding language. Kopřivová and Hnátková (2014) studied the relationship between phraseology and dictionary structure, and showed that phraseological patterns discovered by corpus analysis could be systematically used in dictionary entries in the form of collocational phrases, sub-entries, or examples of usage. Gantar (2006) further developed this argument for the dictionary and phraseology field by demonstrating that the phraseological approach of a corpus is more precise than the pre-corpus approach; that is, it is possible to more accurately identify and document fixed expressions, collocations, and idioms in and out of context than would have been possible before the introduction of the corpus approach.

The GDEX system mentioned by Kilgarriff et al. (2008) is an important example of computational solutions to the example sentence selection problem, which has been traditionally performed manually and inconsistently. GDEX exemplifies how digital technology can improve the quality and uniformity of the lexicographic product by automating the process of identifying good examples in accordance with corpus-derived criteria. Most recently, Rice and Zorn (2021) have investigated the use of corpus-based data for the construction of a dictionary in the context of sentiment analysis, and have demonstrated the ways in which collocational and semantic data from corpora can be utilized to create specialized lexicographic resources for computational purposes. It marks a progression in the scope of use of corpus-

based dictionaries from language reference to natural language processing and computational social science.

Lew (2013) takes the issue of "dictionaries and technology" in a more general direction by looking at the ways in which digital delivery platforms have transformed user expectations and usage, and how this affects lexicographic design. The shift from print to electronic dictionaries has provided opportunities to present additional features of the dictionary, such as dynamic display of updated entries, hyperlinked collocates, and frequency-based ranking of dictionary entries; it has also posed challenges in terms of how to display corpus evidence for non-specialist users.

2.4 Specialized, Multilingual and Digital Dictionaries

One of the most fruitful fields for the use of corpus linguistics in the field of lexicography has been the creation of specialized dictionaries for specific user groups, professional areas, and/or language learning. Bowker (2010) discusses the role of corpus linguistics in the context of specialized learner dictionaries, suggesting that the collections of learner corpora, consisting of L1 or L2 language produced by L2 students, can offer valuable information to the specialist dictionaries for learners. The learner-centeredness of corpus-based lexicography has important implications for the development of ESP and EAP dictionaries.

Prinsloo (2009) is optimistic about the future of corpora in dictionaries, forecasting that as more large and diverse corpora are made available, dictionaries will be more descriptively accurate, more frequently updated and more responsive to the needs of specific user groups. In some years this foresight was realised as can be seen in the development of corpus projects in specific fields of application and specialised lexicographic efforts reported in the literature. In contrast, Mascott (2017) looks at the dictionary as a specialized corpus, and proposes that the norms found in dictionaries can be investigated as a type of institutionalised language data, thus reversing the conventional corpus-to-dictionary workflow.

One of the areas of special research interest has been the development of multilingual and bilingual corpus-based dictionaries. In Orenha-Ottaiano (2016), the author presents the compilation of a printed and online corpus-based bilingual collocations dictionary, showing how parallel and comparable corpora can be utilized to uncover various collocational patterns which don't map easily across languages. This work emphasizes the specific problems involved in corpus-based bilingual lexicography such as the absence of corpus data for low-frequency but significant collocations and the necessity of corpus tools which are able to deal with parallel texts and cross-linguistic alignment.

The development of a digital dictionary for learners has been a growing trend and has been the subject of some recent research that covers the use of corpus approaches in online and mobile reference tools. Kuhn (2017) proposes a design for an online corpus-driven dictionary for college learners of the Portuguese language, and shows how one can leverage corpus data for the design of specific dictionary entries for a given learner population. This study, conducted by RU, Mahliatussikah, and Ahsanuddin (2024), presents empirical results that indicate that the students perceived that a corpus-based dictionary is more useful and authentic than a non-corpus dictionary especially in the aspect of understanding of idiomatic and phraseological language.

Zainudin, Jalaluddin, and Bakar (2014) analyze corpus and frame semantics in a lexicography class, and they argue that the use of corpus-based approach in the teaching-learning process of lexicography helps the students to enhance their competence in assessing dictionary entries and to understand the importance of empirical approach in the process of dictionary-making. This paper presents a case study on corpus-to-dictionary compilation for Balinese language,

which describes how the corpus method could be adapted to the language that has limited electronic resources as well as specialized cultural vocabulary. Although this work is an important example of the use of corpus linguistics in languages other than the dominant ones of the world, the author is aware that the scarcity of data and annotated corpora for low-resource languages pose significant challenges.

In their paper, (Apresjan, Mikulin, 2016) discusses the dictionary as a tool of linguistic research and look at how dictionary data can be used to enrich corpus research, which is not often done and needs further development. This is a two-way view of corpus-dictionary relationship, anticipating the increasing interest in the integration between two resources as an iterative rather than a pipeline process. The conceptual lexicography, as introduced by Hartmann and James (2002) in their Dictionary of Lexicography, has given the lexicographic basis for the definition of terms, which has in turn become the theoretical and theoretical-linguistic framework within which corpus-based approaches to dictionary making have been theorized and discussed.

The literature examined in the three themes provides a firm foundation of understanding about the potential of corpus linguistics to revolutionize modern lexicography. In parallel, it exposes a persistent research void: The systematic investigation of the systematic use of specific corpus tools (such as AntConc and LancsBox) in dictionary compilation processes, and the impact of this use on the quality of the lexicographic products in various types of dictionaries and target languages has been insufficiently developed. This article aims to fill this gap, summarizing the evidence and detailing the mechanisms that make corpus analysis tools significant in lexicographic practice.

3. Methodology

3.1 Research Design

The research design of this study was systematic literature review with qualitative approach. The research method used is systematic literature review and qualitative research design. The methodology chosen for this systematic review is a content area in which the research questions are mostly conceptual and theoretical in a sense that the collection of primary empirical data is not the main objective in the review, but rather the synthesis and critical evaluation of existing scholarship. The method enabled an overall picture to be obtained of the main themes, debates and findings in the literature consulted and it enabled gaps in the literature to be identified which are not apparent in any individual study.

3.2 Corpus of Sources

The materials examined in this research were based on corpus-based materials from journals, edited books, conference proceedings and book chapters in the area of corpus linguistics, lexicography and computational linguistics. Sources were chosen for study because they directly related to the intersection between corpus methods and the making of dictionaries. The reference corpus for this study included a total of 31 sources ranging from 2002-2024, in order to show the intellectual evolution of the field, from basic theoretical works to recent digital and multilingual uses. Each source that was included in the review was assessed in terms of methodological quality, theoretical significance, and usefulness with regard to the three research questions.

3.3 Data Collection and Analysis

Thematic coding and close reading were used to analyze the sources. For each source, the main theoretical contributions, methodological approaches, findings of the studies, specific tools or software mentioned, and the implications for corpus-based lexicography were coded. Corpus-driven methodology in lexicographic practice; Digital tools, collocation analysis and

computational lexicography; and Specialized, multilingual and digital dictionaries were the three major themes of the literature review which was organized by the coding process. Lexicographic and related corpus-based research was a particular focus in the study, and it was important to find references to AntConc and LancsBox, the two tools involved.

3.4 Tools: AntConc and LancsBox

The following two corpus analysis tools were selected for special consideration in the methodological approaches' synthesis: AntConc and LancsBox. AntConc (Anthony, 2022) is a free corpus analysis toolkit that offers concordancing, frequency analysis, collocate extraction and identification of n-grams, as well as keyword analysis. Its use in both research and applied lexicography makes it a point of reference for the use of corpus data in dictionary making contexts. LancsBox (Brezina et al., 2020) is an advanced collocation analysis system developed at Lancaster University, which offers a variety of association measures, such as MI, log-likelihood and the ΔP family of statistics, and has a GraphColl module for visualizing collocational networks. The ability of LancsBox to reveal the strength and directionality of collocational relationships is especially important for the lexicographers who are interested in dealing with the phraseological behavior of dictionary entries.

3.5 Ethical Considerations

This study did not require primary data collection with human subjects, so formal ethical approval was not required. All resources used have been publicly available academic publications and full reference has been given in line with academic integrity. In preparing this review, the publisher has not used any proprietary corpus data or unpublished materials.

4. Results

4.1 Thematic Synthesis: Corpus Methods and Lexicographic Practice

A systematic review of 31 sources revealed that there was found to be converging and consistent evidence that corpus linguistics is the backbone of modern dictionary-making in a variety of dictionary types, genres and contexts. The literature revealed three main ways corpus methods influence lexicographic practice: creating a list of headwords based on the frequency of words, using a corpus to write definitions, and automatically or semi-automatically extracting examples from the corpus.

The most common criterion for the selection of the headwords in modern dictionaries presented in the reviewed sources was corpus frequency. According to the above sources such as Walter (2010), Kilgarriff and Jackson (2013) and Rees (2022), corpus-based frequency lists can help lexicographers identify which words and meanings are salient at a particular point in time and thus enable dictionaries to focus on the most relevant vocabulary for their intended users. This is especially important in a learner dictionary in which words are selected according to their frequencies, which means that the words which children most commonly need to produce and most commonly find in a text are included.

Corpus-driven definition writing was reported by various sources to be able to provide more precise, natural and register-responsive definitions in the dictionary. Both Summers (2014) and Walter (2010) reported that corpus concordances can be used to infer what sorts of syntactic contexts words are typical of, what kinds of prosodies they carry, and where they occur in discourse, all of which help to word dictionary definitions. Kilgarriff et al. (2008) documented the production of statistically validated automatically selected examples in the GDEX system which performed better than manually selected examples on various quality measures such as length, frequency of vocabulary, and lack of ambiguous references.

4.2 Collocation Analysis in AntConc and LancsBox

One of the key points of this review is focused on the particular contribution by AntConc and LancsBox to collocational analysis in the field of lexicographic contexts. The studies reviewed consistently suggested that the collocation analysis tools are useful and can be used for identifying the most frequent and statistically significant collocational partners of the headword, for documenting the pattern of collocations that would not be noticed in manual analysis and for choosing representative collocational examples to be included in dictionary entries.

As Kilgarriff and Kosem (2012) noted, the ability to access integrated functions such as frequency, collocation and concordance, as offered by tools like AntConc and LancsBox, is crucially important in cutting down the time needed to do corpus analysis tasks in the dictionary compilation workflows. LancsBox's GraphColl module, specifically, gives the visual representation of the collocational network around any given word, providing the lexicographer with information not just about immediate collocates but also about secondary and tertiary collocational relationships which can offer a clearer picture of the semantic and pragmatic behavior of the headword. This may have implications for the documentation of multi-word units, idiomatic expressions and phrasal verbs in dictionaries.

From the literature reviewed, it was found that AntConc's collocate extraction function, which produces association measures such as Mutual Information (MI) and Log-likelihood scores, was of great value in identifying accidental co-occurrence and genuine collocation. In a study, Rice and Zorn (2021) proved that the dictionary resources built with such corpus-based collocation analyses would be more discriminatory than the dictionary resources created based on intuitive or frequency-based methods. Copřivová and Hnátková (2014) also found that the phraseological patterns revealed by corpus-based collocation analysis can be formally included in the dictionary entries, which will make the dictionary more useful to users who are looking for ways to understand word usage rather than its meaning.

4.3 Specialized, Multilingual and Digital Dictionary Development

The literature search on specialized, multilingual and digital dictionaries revealed that the application of corpus methods has opened up a wide variety of specialized, domain-specific and learner-oriented dictionaries beyond the scope of traditional monolingual general-purpose dictionaries. According to Bowker (2010), corpus methods can enable dictionary compilers of special purposes or disciplines to discover the kind of terminology, collocations and phraseological patterns that are typical of a particular field of professional or discipline with greater accuracy than could be achieved manually. Ningsih et al. (2022) reported that corpus methods can be applied successfully in preparing an e-dictionary for automotive domain that has limited resources in the corpus.

The evidence on digital dictionary development showed that the use of delivery systems such as the Internet and mobile devices is becoming more significant as means of delivering corpus-based lexicographic content. Lew (2013) reported that in the digital context, corpus-based dictionary can be enriched with dynamic features, for example, frequency list, interactive concordance, collocational networks, which could not be included in print versions of the dictionary. According to Kuhn (2017), corpus-driven design principles, such as providing access to frequency information to prioritize the entries, and providing authentic corpus examples to demonstrate how the item is used, greatly enhanced the perceived relevance and usability of an online learner dictionary for university learners of Portuguese.

Corpus methods in underrepresented and low-resource languages was found to be an active but difficult stream of development. In other words, the use of corpus methods with the Balinese language needed a lot of adaptation of the normal corpus-building and analysis

process to suit the scarcity of digital text source materials and the cultural specificity of the Balinese vocabulary as reported by Praminatih (2023).

5. Discussion

5.1 Summary of Key Findings

The results of this systematic review do support the main thesis of this article: Corpus linguistics has now become an essential methodological bedrock for modern dictionary compilation. In all three areas of analysis (corpus-driven methodology, digital tools and collocation analysis, and specialised and multilingual dictionaries), the literature presents a similar picture of change. The value of a dictionary informed by corpus evidence is clearly greater than their pre-corpus counterparts and the tools like AntConc and LancsBox have enabled lexicographers to implement corpus analysis in practical, efficient and repeatable ways.

5.2 Comparison to Previous Literature

The conclusions to this review are consistent with, and in some aspects extend, the arguments put forth in previous foundational reviews. The literature that followed Krishnamurthy (2002) bears out the continuation and intensification of this revolution, as well as its extension to other dictionary genres, languages and modes of delivery. The studies reviewed provide empirical support for the theoretical approach developed by Kilgarriff and Jackson (2013) in their understanding of corpora as the primary data source for dictionaries, in that each study demonstrates, in different domain, how corpus evidence benefits lexicographic outcomes.

The contribution of the present review is that it brings together the evidence from some isolated studies on tool mediated corpus lexicography, which has not been brought together in a unified framework. This review, by combining research results collected by the AntConc software and the research results collected by LancsBox software, with other general corpus lexicography research results, will make it clear that the greater the capacity of the software, the more the quality of the lexicography. By integrating the results of research conducted with the AntConc software and the results of research conducted with the LancsBox software, and other general corpus lexicography research results, this review will elaborate more clearly the relationship that is often implicit or assumed in the individual research results between the capacity of the software and the quality of the lexicography.

5.3 Interpretation of Findings

Convergent evidence of the transformative function of corpus linguistics in dictionary making can be interpreted in two complementary ways. The first is representational accuracy, because corpus methods give lexicographers access to larger and more varied collections of language than would be possible for any single compiler to gather, and so would allow them to represent the language in actual use and not as it prescriptive authorities think it should be used. The second one is the efficiency: for instance, the tools used to extract tools for collocation extraction, the example selection, the frequency ranking, etc. which would otherwise be too time-consuming if done by hand, make it possible to carry out corpus-based lexicography at scale.

The results of a specific contribution of collocation analysis with AntConc and LancsBox to the documentation of the phraseological behaviour in dictionaries is of particular significance. As shown by Kopřivová and Hnátková (2014) and Gantar (2006), one of the areas of language description where the corpus method adds the most value to the pre-corpus method is the description of phraseology, as these patterns are too subtle and too frequent to be reliably identified through introspection or manual analysis. LancsBox's GraphColl module is able to produce networks of the collocations that help the lexicographers to have a more

comprehensive picture of the behavior of a word not only for the writing of individual entries, but also with respect to the structure and interrelationship of related entries in a dictionary.

5.4 Limitations

There are a number of limitations to consider in this study. Firstly, as a systematic literature review, it relies on the quality and extent of the literature reviewed. An attempt was made to cover a representative cross section of scholarship but it may be possible that the review corpus does not reflect scholarly literature in the field. Secondly, the review is mainly restricted to English-language scholarship, which can lead to a certain perspective on corpus-based lexicographic practices in other research traditions that do not make English as their native language. Thirdly, the study itself contains no primary empirical data on the use of AntConc and LancsBox for actual dictionary compilation processes; conclusions about the use of these tools are based on reports in literature, and are not empirical or observational.

5.5 Theoretical Implications

On the theoretical level this research helps to build a framework for corpus-assisted lexicography as a whole, as it illustrates the correspondence between the features of tools and the tasks performed in them. The mapping of AntConc's collocate extraction function with the documentation of word behaviour in dictionary entries, for instance, suggests a direct relation between the quality of the documentation of the dictionary entry and the design of the software, implications that can be used to guide the design and evaluation of corpus tools. The success of corpus methods in specialized, multilingual and digital dictionary research suggests that the corpus revolution in lexicography is not limited to a restricted number of mainstream, monolingual general purpose dictionary genres, but extends to any and all dictionary genres.

5.6 Practical Implications

In practice, the results of this research are significant for lexicographers, the creators of a corpus database, teachers of foreign languages, and policy makers who are interested in language documentation. The evidence presented in this paper gives a good foundation for the use of AntConc and LancsBox as the standard tools for lexicographers to use in dictionary compilation processes, especially in the analysis of collocation and phraseology. The review demonstrates the pedagogical utility of corpus-based dictionaries for the language teacher and the language curriculum developer, as it enables language learners to have access to authentic language data in a structured reference format. The results suggest that corpus-based methods have much to offer to the documentation of underrepresented languages, including language agencies and policy makers, and draw attention to the necessity of investing in corpus building for low-resource languages before corpus-based lexicographic development can be seen to yield results.

5.7 Future Research Directions

This review highlights several specific avenues for research into an appropriate future direction. Empirical research on the real application of AntConc and LancsBox in dictionary compilation workflows would be helpful in providing evidence of the ways in which corpus analysis mediated through these tools has impacted the quality of lexicography, including in the form of case studies of lexicographic projects in which they have been used. Second, dictionary development applications that rely on the use of the corpus approach, as discussed by Praminatih (2023) and RU et al. (2024), would fill the greatest gap that was found in this corpus. Third, comparative research on the use of various collocation indicators, which are offered by AntConc and LancsBox (MI, log-likelihood, ΔP), to identify and prioritize collocates for lexicographic purposes would be useful guidance for lexicographers in the selection of an analytical approach.

6. Conclusion

In this article thirty-one academic sources of literature, ranging from theoretical beginnings to the latest developments in digital and multi-lingual dictionary production, have been examined systematically to describe the contribution of corpus linguistics to the making of contemporary dictionaries. The literature for all three thematic areas examined provides strong support for the main argument of this thesis: corpus linguistics is an essential methodological underpinning of modern lexicography. The answers to the three research questions are as follows: corpus linguistics has revolutionized dictionary making; it has provided the foundation for selection of heads for the dictionary, the writing of dictionary definitions, and the extraction of examples of sentences in which the heads are used; corpus-based methods have been productive for specialised, multilingual and digital dictionaries, but there are many challenges in adapting these to underrepresented and low-resource languages; corpus-based methods have been applied to dictionary making in a way that would be unsuccessful if done without them. For practitioners and policy makers, the key message here is that corpus construction and corpus tool building is not just a technical matter but a linguistic equity one as well. The advantages of the corpus lexicography will stay limited to some well-resourced world languages, until enough large well-designed electronic corpora will be available for all of them. This is one of the most important challenges for lexicographers in the twenty-first century.

References

Apresjan, V., & Mikulin, N. (2016). The Dictionary as a tool for language study. *The 17th EURALEX International Congress: Lexicography and Linguistic Diversity, Proceedings of. Ivane Javakhishvili Tbilisi State University* (pp. Tbilisi. 224–231).

Barlow, M. (2011). Corpus Linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, *16*(1), 3–44.

Bowker, L. (2010). The role of the Corpus Linguistics in the creation of learner's specialised dictionaries. *Specialized Dictionaries for Learners*, 155–168.

Dash, N. S., & Ramamoorthy, L. (2018). Structured and unstructured tasks for the making of a corpus and a dictionary. In *Utility and Application of Language Corpora* (pp. 121–138). Springer Singapore.

Gantar, P. (2006). Phraseology and dictionary approach, corpus approach. *Slavistična revija*, *54*, 161–162.

Hartmann, R. R. K., & James, G. (2002). *Dictionary of lexicography*. Routledge.

Heuberger, R. (2016). Corpora as game change: The emerging role of corpus tools in the dictionary-making and user community. *English Today*, *32*(2), 24–30.

Hoopes, N. A. (2018). The primary significance test of dictionaries, corpus linguistics and trademark genericide. *Tulsa Law Review*, *54*, 407.

Kilgarriff, A. (2005). Putting the corpus into the dictionary. In *Proceedings MEANING Workshop*.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress* (Vol. 1, pp. 425–432). Universitat Pompeu Fabra.

Kilgarriff, A., & Jackson, H. (2013). The use of corpora as sources of data for dictionaries. In *The Bloomsbury Companion to Lexicography* (pp. 77–96). Bloomsbury.

Kilgarriff, A., & Kosem, I. (2012). A selection of corpus tools for lexicographers (pp. 31–55).

Kilgarriff, A., Rundell, M., & Uí Dhonnchadha, E. (2006). A corpus development for the efficient preparation of a lexicography: The New Corpus for Ireland. *Language Resources and Evaluation*, *40*(2), 127–152.

Kopřivová, M., & Hnátková, M. (2014). From a dictionary to a corpus. For the *Phraseology in Dictionaries and Corpora* (pp. 155–168). Filozofska fakulteta Maribor.
Krishnamurthy, R. (2002). The transformation of EFL dictionaries. *Kernerman Dictionary News*, *10*, 23–27.
Kuhn, T. Z. (2017). *Proposal for an online Portuguese for university students dictionary based on a design approach to the Corpus-driven method*.
Lew, R. (2013). *Dictionaries and technology*.
Mascott, J. L. (2017). The dictionary as a special corpus. *BYU Law Review*, 1557.
Ningsih, Y., Syaief, A. N., Artika, K. D., & Herpendi, H. (2022). Developing multilingual automotive e-dictionary based on corpus linguistics. *Applied Science and Technology on Social Science 2021 (iCAST-SS 2021)* (pp. 962–967). Atlantis Press.
Orenha-Ottaiano, A. (2016). The preparation of an online and print bilingual collocations dictionary based on a corpus. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 6–10).
Paquot, M. (2012). The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool. *Electronic Lexicography*.
Praminatih, G. A. (2023). Corpus and junior dictionary of Balinese language. *Jurnal Kajian Bali (Journal of Bali Studies)*, *13*(1), 48–63.
Price, H. (2022). *The language of mental illness: Corpus linguistics and the construction of mental illness in the press*. Cambridge University Press.
Prinsloo, D. J. (2009). *The use of corpora in future dictionaries*.
RU, N. I., Mahliatussikah, H., & Ahsanuddin, M. (2024). The perception of students about the development of digital dictionary of Arabic idioms using corpus linguistics. Students perception about the development of digital dictionary of Arabic idioms using corpus linguistics. *Arabiyat: Journal of Arabic Education & Arabic Studies*, *11*(2).
Rees, G. (2022). Using corpora to write dictionaries. In *The Routledge Handbook of Corpus Linguistics* (pp. 387–404). Routledge.
Rice, D. R., & Zorn, C. (2021). Sentiment analysis dictionaries based on corpora of specialized vocabulary. *Political Science Research & Methods*, *9*(1), 20–35.
Summers, D. (2014). The importance of Dictionaries for language learning. *Vocabulary and Language Teaching* (pp. 111–125). Routledge.
Tankersley, D. C. (2018). However, corpus Linguistics: its use by the judge sua sponte is not adequate for statutory interpretation. *Mississippi Law Journal*, *87*, 641.
Walter, E. (2010). Using corpora to write dictionaries. In *The Routledge Handbook of Corpus Linguistics* (pp. 428–443). Routledge.
Zainudin, I. S., Jalaluddin, N. H., & Bakar, K. T. A. (2014). Corpus and frame semantics within a lexicography class and assessing the entries in a dictionary. *Procedia-Social and Behavioral Sciences*, *116*, 2316–2320.