DOI: https://doi.org/10.24297/jal.v12i.9122

Understanding Malay Corpora: A Content Analysis of 15 Malay Corpora

Jowati Juhary[1], Erda Wati Bakar[2], Mardziah Shamsudin[3], Asniah Alias[4]

[1,2,3,4]Language Centre, National Defence University of Malaysia, Malaysia

Abstract

In recent years, corpus research has grown in importance, particularly in Malaysia and for the Malay language, the country's official language. Texts and transcriptions of talks for a range of settings make up a corpus. The Malay language, which is Malaysia's native and official language, is the focus of this short paper. The objectives of this paper are to identify the features and types of Malay Corpora, as well as the needs for a military-oriented Malay Corpus. The methodology used in this short paper consists solely of content analysis of pertinent texts on the establishment of Malay language corpora. Preliminary findings suggest that there are at least 15 Malay corpora in existence and that some of the features in these corpora overlap. Further, the researchers argue for the need for a Malay Corpus for Military Operations since the existing corpora do not fully cater for this type of corpus.

Keywords: concordance; corpus; Malay corpora; multimodal corpora.

Introduction

The Malay language is part of the Austronesian language family, which also includes Malagasy, Tagalog, and Pilipino (O'Grady & Archibald, 2000). The language is spoken in Malaysia, Brunei, Indonesia, Singapore, and southern Thailand. The written form of the Malay language in the past was Jawi, an adapted Arabic script; however, Malaysia, Indonesia, and Brunei collaborated to "produce a more uniform form of the language utilising Roman alphabet" (Tan et al., 2009).

The Malay language has been the official and national language of Malaysia since its independence in 1957. The status and functions of the language are explicitly stated in Article 152 of the Malaysian Constitution. Despite this, due to allegedly its restrictions on syntax, morphology, pragmatics, and vocabulary, the Malay language has not been used in all domains of expertise, academic, or research. Some researchers argue that because of colonial influence, English, the country's second language, has become more prominent in a variety of fields of study. Despite being a dominant language in the region, Knowles and Zuraidah (2006) asserted that Malay is "one of the least known to current linguists in the western world" and maybe "the least regulated."

The choice of corpus becomes crucial not just for corpus builders but also for corpus users when establishing a corpus because the range of questions that can be investigated is determined by the composition of the corpus. This study serves as a foundation for understanding the existing literature in the subject as part of a larger initiative that will eventually establish and develop a Malay Corpus for Military Operations. Malay is the tenth most widely spoken language worldwide. Because there are little or minimal digital resources on the Malay language, many studies on language and corpus focus on the English language (Nasiroh Omar et al., 2017). Therefore, this paper attempts to understand and identify what constitutes Malay Corpora.

To address the objectives and research questions, this paper is divided into four main sections, including this introduction, which consists of a short discussion on the methodology adopted in this paper and its research questions. The main findings of this paper are presented in the second section, where selected reports and past research on Malay corpora are examined and discussed. The third section answers the research questions posed, and then it provides some directions for the next course of action on completing the bigger research on a Malay Corpus for Military Operations. A brief conclusion closes this paper.

As this is a short paper, content analysis is used as the primary research approach. The first stage requires the researchers to search for documents on Malay Corpora written by local and international scholars. The researchers choose to use only Google Scholar at this stage. Next, the main findings are identified by the researchers to facilitate discussions and to answer the research questions at the end of this paper.

CC BY

Two research questions will be answered, and these include,

a. What are the types and features of Malay Corpora?

b. Why is there a need to develop a Malay Corpus for Military Operations in Malaysia?

Understanding the Debates on Malay Corpora

This section is divided into three sub-sections that are interrelated to Malay Corpora. These include the discussions on the corpus, linguistics, and Malay Corpora.

2.1. What is a Corpus?

A corpus is a collection of naturally occurring language text selected to represent a state or variant of a language (Sinclair, 1991). A corpus, as defined by Bjorkenstam (2013), is a collection of natural language that contains not only texts but also transcriptions of talks or signs. While most available corpora are text only, a rising number of multimodal corpora, such as sign language corpora, are becoming available. A multimodal corpus is "a computer-based collection of language and communication-related material drawing on more than one sensory modality or more than one production modality" (Allwood, 2008), where sensory modalities include sight, hearing, touch, smell or taste, and production modalities, for example, speech, signs, eye gaze, body posture, and gestures. Therefore, a multimodal corpus is a collection of videos and/or audio recordings of people communicating, and in a variety of scenarios and contexts.

However, any collection of audio and video cannot be simply labelled a corpus. Bjorkenstam (2013) further claimed that as a result, audio-visual content must be selected and filtered first, and metadata should be used to characterise the information. Second, the content should be analysed and reported consistently, with transcriptions and notes. A corpus, in theory, is a collection of language production samples chosen to be representative of a language (or sub-language) rather than a set of data collected at random. The sampling of the corpus defines how representative a corpus is for a certain research subject. To generate a generic corpus, language samples from men and women of all ages from various regions of the area where the language is spoken must be collected and included.

2.2. Linguistics

Linguistics is the scientific study of language, where language elements including syntax, morphology, pragmatics, semantics, and vocabulary are observed. Linguistics researchers examine all these aspects to see how the language is organised by the original speaker or the individual who makes the typical and systematic utterances (Shahidi A. Hamid, Kartini Abd. Wahab & Sa'adiah Ma'alip, 2018). The way the language is arranged can be seen through the corpus data. What is meant by corpus data? Corpus data are the collection of language or texts data. Language data can be in the form of oral or written or both which are saved in the computer and functioned as the language sample for linguistics research. There are three important aspects in explaining the corpus data: authentic data, electronic data, and oversized data (see Sinclair & Renouf, 1988; Francis, 1992; Kennedy, 1998; Tognini-Bonelli, 2001).

The term 'authentic' refers to data that are acquired from actual communications between people. These data are crucial in linguistics study because they indicate how a language behaves when it is utilised. The corpus data used as a sample will allow researchers to gain a true image of a language to create a language model or formula. The corpus data, on the other hand, is kept in an electronic format. Corpus data are always associated with the machine-readability phrase rather than electronic phrases. This means that the machine, that is the computer, oversees most of the corpus data control. The usage of computers has increased the amount of data that can be stored. For instance, Brown Corpus is the first corpus data produced in the form of machine-readable. It has recorded an amount of one million words which have been collected from 500 textbooks, and each text consists of 2,000 words (Garside, Leech & McEnery, 1997).

A few aspects need to be taken into considerations when developing a corpus (Biber, 1993; McEnery, Xiao & Tono, 2006) such as,

a. texts selection;

b. the number of texts sufficient for the research; and

c. the size of corpus data that needs to be developed.

By these three considerations, researchers and scholars are reminded that text selections must include samples that are comprehensive in terms of gender, locality, and scenarios of the interlocutors. Further, depending on the research questions and objectives of the research, the sample size can be determined. No one research can cover all language samples! Finally, the size of the corpus data relies on the research questions and objectives explained earlier. Developing a corpus is done in stages because language is dynamic; it keeps growing as speakers keep using the language and when transfer or borrowing of words continue to occur.

2.3. Malay Corpora

This sub-section begins with a historical account of Malay Corpora in Malaysia. The agency responsible for this effort is *Dewan Bahasa dan Pustaka* (DBP), a government institution in charge of the planning and monitoring of the use of the Malay language in the country. DBP has its corpus database, known as the DBP Corpus database, which is built in stages. The DBP Corpus database was originally developed as an archive of texts, according to Rusli Abdul Ghani, Norhafizah Mohamed Husin and Chin (2006), with a method for processing selected texts to provide concordances as well as statistical information on word frequencies and total number. In addition, the DBP Corpus database contains 115,530 news articles, 1,981 magazine articles, 703 literary texts, 663 books, 128 working papers, and 36 "ephemeral" materials and 118,913 texts. Most newspaper texts (86,885 or 75%) were from *Berita Harian*, and 17,539 (15%) were from *Utusan Malaysia*. There were also texts from smaller newspapers such as *Harian Metro*, *Berita Minggu*, *Harakah* and *Metro Ahad*. Despite these statistics, the corpus details (such as the dates, the size, and the number of words) were not mentioned on the DBP Corpus website (Chung et al., 2019).

Nonetheless, according to Hajar Abdul Rahim (2014), with the expansion of Malay Corpora and its role as a Malay language resource centre, DBP has expanded corpus-based research, not only for systematic analysis and description of Malay linguistics but also for Malay language pedagogy, in collaboration with corpus linguists from within and outside the country. She went on to claim that the DBP Corpus database has produced new descriptions of the Malay language, which has influenced the content, material creation, and syllabus design for Malay language teaching and learning in the country.

Furthermore, according to a bibliographic analysis undertaken by Siti Aiesha Joharry and Hajar Abdul Rahim (2014) based on published works between 1996 and 2012, the use of the corpus technique in language research in Malaysia is on the rise. They also discovered that corpus research in Malaysia is concentrated on five areas: (a) English language use in Malaysia; (b) Malaysian English language learners; (c) Malaysian textbook material; (d) Malay language description and lexicography; and (e) corpora development. According to Siti Aiesha Joharry and Hajar Abdul Rahim (2014), the development of other types of corpora, such as the Malay language learner and pedagogic corpora, will allow research in the Malay language to expand beyond language description, translation, and lexicography to include Malay language learning and teaching issues, which are currently underserved.

Hajar Abdul Rahim (2014) reported that up until 2008, the DBP Corpus database comprised 128 million words which were compiled in 10 sub-corpora representing different genres of texts. These include books, magazines, newspapers, translations, ephemerals, drama, poems, material cards, traditional texts, and school textbooks. In 2017, the number of words in the DBP Corpus database has increased to 135 million words, and the sub-corpora and the number of words for each sub-corpus are illustrated in Table 1.

Table 1. The DBP Corpus database (Official Website for DBP, 2021)

Sub-Corpora	Types of Materials	Number of Words
Books	novels, academic books, general reading, textbooks	31,580,305

Magazines	general and covers various fields	14,406,888
Newspapers	daily, tabloids, Sunday editions	80,029,347
Translations	academic books and general readings	2,021,191
Ephemerals	flyers, brochures, advertisements	290,207
Drama	bound form	404,176
Poems	bound form	116,428
Material Cards	collected cards for the development of *Kamus Dewan*	3,130,641
Traditional Texts	Old text collection and folklores	2,825,329
Textbooks	primary and secondary school textbooks	1,095,726

One of the outputs of the DBP Corpus database, according to some, is the release of a dictionary, which is regarded as an important success in Malaysian corpus-based Malay lexicography. Other areas of pedagogy, however, can benefit from these corpora. Figure 1 depicts Johansson's (2009) diagrammatic description of corpora's usage in language teaching and learning, which shows how corpora can fully contribute to language pedagogy.

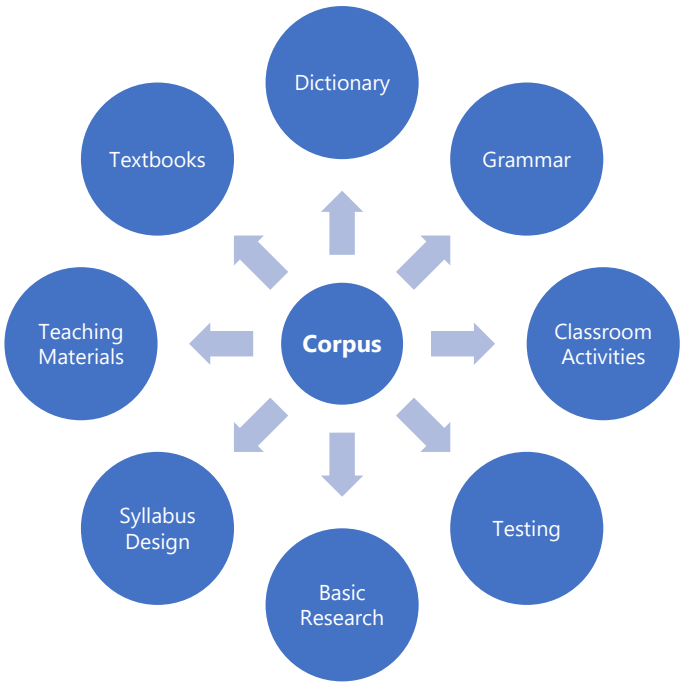

Figure 1. Uses of corpora in language teaching and learning (Johansson, 2009: 40)

Furthermore, Normi Sham Awang Abu Bakar (2020) in her study concluded that there are six Malay text corpora, including sealang.net, https://glosbe.com, mcp.anu.edu.au, MyBaca.org., https://www.lexilogos.com/english/malay_dictionary.htm, and prpm.dbp.gov.my (DBP). Each one of these is further explained in Table 2. Notwithstanding this, Normi Sham Awang Abu Bakar also argued that these various corpora of Malay text are not integrated, where some words are not included or missing in some corpora.

CC BY

Table 2. Malay Corpora (Normi Sham Awang Abu Bakar, 2020)

Corpus	Context(s)	Application	Provider
sealang.net	General	Collocation analysis Concordance Provide with examples Sentences from multiple sources	CRCL and the University of Wisconsin-Madison Centre for Southeast Asian Studies (CSEAS)
https://glosbe.com	General/Education	Online dictionary Almost all languages available Millions of translations	Glosbe
mcp.anu.edu.au	Education	Concordance	Australian National University
MyBaca.org	Education	To find words whether it is capable in starts, end or middle of the sentences	School of Educational Studies
https://www.lexilogo.com/english/malay_dictionary.htm	Education	Malay - English dictionary	Lexilogos Project
prpm.dbp.gov.my	Education	Define and provide the detailed meaning of words	DBP

In addition to the Malay corpora identified by Normi Sham Awang Abu Bakar (2020), Chung et al. (2019) have also identified other corpora that are discussed by other researchers (see Table 3).

Table 3. Malay Corpora (Chung et al., 2019)

Corpus	Descriptor(s)	Websites
MalaysianWac (or zsmWac) Corpus, Sketch Engine	Non-tagged raw texts	https://www.sketchengine.eu/zsmwac-malaysian-corpus/
Malay Practical Grammar Corpus (MPGC)	A section of the DBP Corpus database	Not available
MALEX (MALay LEXicon)	A list of Malay lexicons (with morphology information) with English translations	Not available

In concluding this section, the researchers opine that Malay Corpora are indeed rich and that DBP as the responsible agency to ensure effective use of the Malay language in Malaysia has a massive task ahead. The main question remains whether these corpora are sufficient, or if more corpora should be developed, what is the best strategy to ensure incidents of missing words are avoided, and that all can be integrated.

CC BY

Results and Discussions

This section focuses on answering the research questions posed earlier in this paper. At the same time, before closing this section, future directions of Malay Corpora will also be discussed.

3.1. Research Question 1 – What are the types and features of Malay Corpora?

Based on the literature discussed earlier and based on the DBP Corpus database (see Figure 2) and SEAlang Library Malay Corpus database (see Figure 3) that are available online, it can be summarised that Malay Corpora have been extensively developed and will continue to be expanded due to social, economic, and political shifts in Malaysia. For example, it is estimated that SEAlang contains about 2.5 million words collected from the web using a crawler (Chung et al., 2019). Given the above analysis, at least at this stage, there are nine types of Malay Corpora documented (see Tables 2 and 3). The researchers argue that there are other Malay Corpora, which might not have been properly documented, and some may be the sub-types of the existing Malay Corpus types.

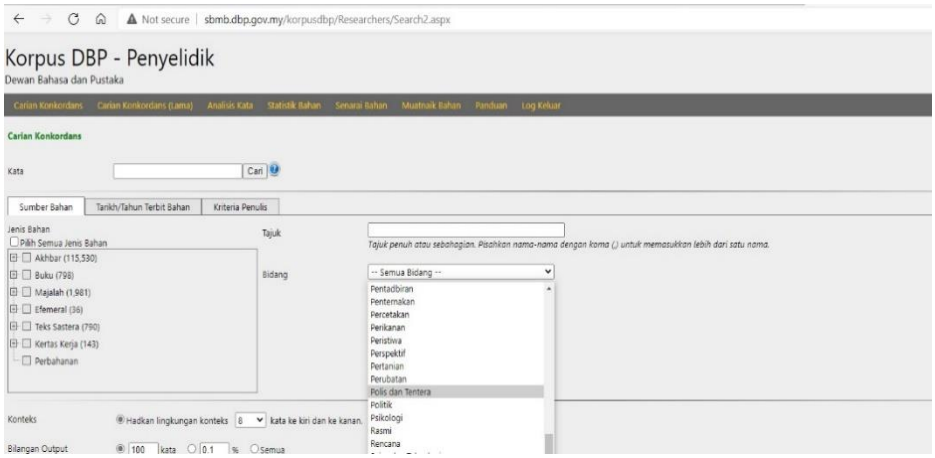

Figure 2. The DBP Corpus database

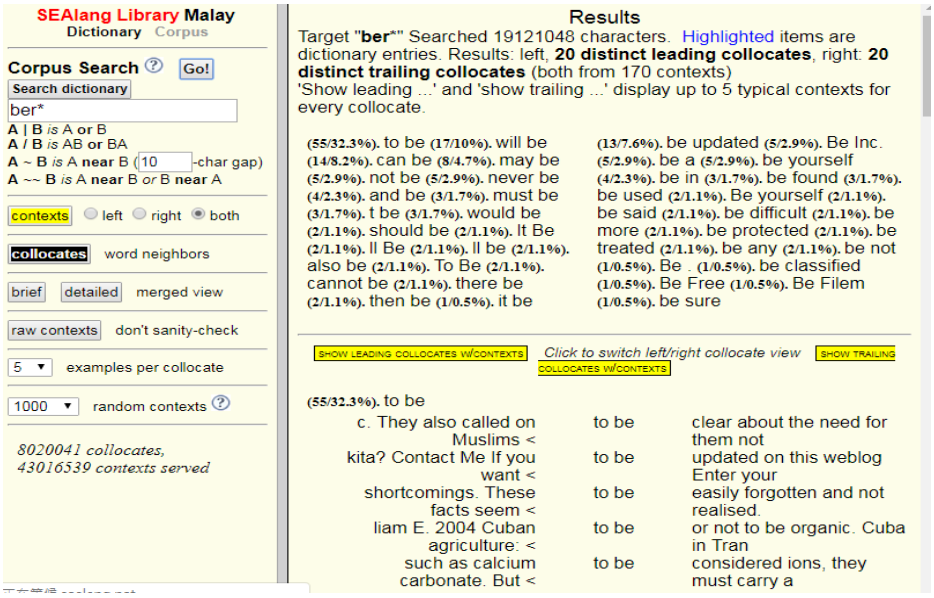

Figure 3. SEAlang Library Malay Text Corpus

Apart from the DBP Corpus database, Siti Syakirah Sazali, Nurazzah Abdul Rahman and Zainab Abu Bakar (In Press) argued that there are other but not annotated corpora available such as from the Institute of Language and Literature. These corpora provide multi-domains such as newspaper excerpts, magazines, novels and many more. Another corpus but also not annotated is Mutiara Hadith UiTM that provides translated Quran and Hadith

CC BY

documents publicly (see Table 4). In addition, there are annotated existing corpora on the terrorism-related corpus, news and biomedical articles, and Twitter excerpts; however, these are not publicly available for they are built as a form of experiments in natural language processing.

Table 4. Other Existing Malay Corpora (Siti Syakirah Sazali, Nurazzah Abdul Rahman & Zainab Abu Bakar, In Press)

Researchers	**Domain(s)**	**Publicly Available**	**Annotated?**
The Institute of Language and Literature	Multiple domains	Yes	No
Mutiara Hadith UiTM	Malay translated Quran and Hadith documents	Yes	No
Zamin et al. (2012)	Terrorism-related journalistic articles	No	Yes
Alfred et al. (2013)	News articles, Biomedical articles	No	Yes
Noor Ariffin and Tiun (2018)	Malay Twitter Excerpts	No	Yes
Azizan et al. (2019)	Durian-related documents	No	Yes

Therefore, based on Table 4, it could be deduced that some corpora are annotated, and some are not. At the same time, some corpora are not publicly available, which makes it even challenging to analyse and understand the existing Malay Corpora. Regardless of this, there are a total of 15 Malay Corpora, four are not available for the public and two are not annotated (see Tables 2, 3 and 4).

Further, the features of Malay Corpora include the identified contexts, which are for general and educational purposes, and their applications, ranging from providing meanings (dictionary) to generating concordance and sentence analysis. The researchers argue that the contexts could be further categorised into other sub-contexts such as research for educational purposes and politics, economy and social for general purposes. This further sub-categorising allows for a more focused and directed work and understanding of Malay Corpora.

3.2. Research Question 2 – Why is there a need to develop a Malay Corpus for Military Operations in Malaysia?

The second research question appears to be easy to answer but poses a tricky scenario to linguists alike. Based on Figure 2, there exists the domain for Police and Military in the DBP Corpus database, but not Military Operations per se. As argued earlier, the contexts could have been extended to include more contexts or sub-contexts to reflect the dynamics of the Malay language itself. At this stage, the researchers could only argue on the significance of developing a Malay Corpus for Military Operations. There are at least two reasons for developing the corpus in the context or sub-context of defence and security. Firstly, as the unknown challenges and threats in the world today increase, there is an urgent need to conduct relevant research in this area. The report of this research must also be written in the Malay language to ensure that Malaysians are updated with the current defence and security issues. This is where the corpus for Military Operations becomes significant.

Secondly, in the advent of Industrial Revolution 4.0 (IR4.0), the defence and security industry need a substantive corpus in the Malay language to enrich the understanding and strengthen the knowledge of Malaysians and international scholars interested in the Malay language alike. Most of the reports and documents on scientific findings are written in English, highlighting the assumption that the Malay language is a sub-language. The Malay language is rich with its terms and vocabulary, and by developing a corpus targeting Military Operations, other areas other than research and providing meaning such as the pedagogy of military training can also use the Malay language effectively due to this corpus.

CC BY

3.3. What Next?

Much is yet to be done in terms of strengthening the Malay Corpus for Military Operations; having the Police and Military as one of the domains in the DBP Corpus database is inadequate due to the unknown challenges in this era and the future. The researchers and others interested in this area of research should ensure that the development of the Malay Corpus for Military Operations include all areas of defence and security as well as military operations other than war (MOOTW). Small scales corpus development on the Malay Corpus for Military Operations can be the starting point for this effort, followed by collaborating and integrating all resources on the corpus for Military Operations. Given the tedious processes involved, the researchers are adamant that only with strong and continuous support from various authorities including DBP, the government and Malaysians, the Malay Corpus for Military Operations can be developed and later used for various purposes.

Conclusions

Based on the discussions in this paper, it is evident that there are at least 15 Malay Corpora. Out of these, four are not made public for reasons unknown to the researchers, and two are not annotated. The researchers argue that there is no harm in developing a new corpus; this will only enrich the language itself because speakers, users and researchers alike get the benefits of various Malay Corpora. In addition, the researchers also argue that the features of Malay Corpora such as contexts could be further categorised into other sub-contexts. The other feature, which is applications of the corpora has been beneficial and significant especially in developing dictionaries.

To conclude, the researchers opine that there is a need to develop the Malay Corpus for Military Operations for the use of the country. Based on the discussions and analysis of the existing corpora, it is evident that there is no specific corpus that could enhance the understanding of military terms in the Malay language especially one that focuses on military operations. It is high time that this Malay Corpus for Military Operations is developed since Malaysia has massive experience in international military operations such as ones in Somalia and Lebanon under the United Nations, and these experiences must be documented in the Malay language since it is the national language and for future reference, locally and internationally.

Conflicts of Interest

There is no conflict of interest.

Funding Statement

The short paper is funded by the National Defence University of Malaysia, under the short-term research grant, UPNM/2020/GPJP/SSI/5.

Acknowledgements

The researchers would like to thank the National Defence University of Malaysia (NDUM) and its research wing, the Centre for Research Management and Innovation, for the grant received to conduct research on Malay Corpora, particularly investigating the Military Malay Corpus in Malaysia.

References

1. Allwood, J. (2008). Multimodal Corpora. In A. Lüdeling, & Merja, K. (Eds.), *Corpus Linguistics: An International Handbook* (pp. 207-225). Berlin: Mouton de Gruyter.
2. Biber, D. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243-257.
3. Bjorkenstam, K.N. (2013). What is a Corpus and Why are Corpora Important Tools? Retrieved on March 19, 2021, from https://nordiskateckensprak.files.wordpress.com/2014/01/knb_whatisacorpus_cph-2013_outline.pdf.
4. Chung, S. F., Shih, M. H., Nomoto, I. H., & Moeljadi, D. (2019). An Annotated News Corpus of Malaysian Malay. *NUSA: Linguistic Studies of Languages in and around Indonesia*, *67*, 7-34.

5. *(The) DBP Corpus database*. (2021). http://lamanweb.dbp.gov.my/index.php/pages/view/76?mid=61. DBP: Kuala Lumpur.
6. Francis, W.N. (1992). Language Corpora. In J. Svartvik. (Ed.), *Directions in Corpus Linguistics* (pp.17-32). Proceedings of Nobel Symposium 82. Berlin; New York, NY: Mouton de Gruyter.
7. Garside, R., Leech, G., & McEnery, T. (Eds.). (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Routledge.
8. Hajar Abdul Rahim. (2014). Corpora in Language Research in Malaysia. *Kajian Malaysia*, *32*(1), 1-16.
9. Johansson, S. (2009). Some Thoughts on Corpora and Second-Language Acquisition. In K. Aijmer. (Ed.), *Corpora and Language Teaching* (pp. 33-44). Amsterdam: John Benjamins.
10. Kennedy, G.D. (1998). *An Introduction to Corpus Linguistics*. London; New York, NY: Longman.
11. Knowles, G., & Zuraidah Mohd Don. (2006). *Word Class in Malay: A Corpus-based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
12. McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London; New York, NY: Routledge.
13. Nasiroh Omar, Ahmad Farhan Hamsani, Nur Atiqah Sia Abdullah, & Siti Zaleha Zainal Abidin. (2017). Construction of Malay Abbreviation Corpus Based on Social Media Data. *Journal of Engineering and Applied Sciences*, *12*(3), 468-474.
14. Normi Sham Awang Abu Bakar. (2020). The Development of an Integrated Corpus for Malay Language. In R. Alfred, Y. Lim, H. Haviluddin, & C. On. (Eds.), *Computational Science and Technology. Lecture Notes in Electrical Engineerin*g, 603, (pp. 425-433). Singapore: Springer.
15. O'Grady, W., & Archibald, J. (2000). *Contemporary Linguistic Analysis: An Int*roduction. Toronto: Addison Wesley Longman.
16. Rusli Abdul Ghani, Norhafizah Mohamed Husin, & Chin, L.Y. (2006). Pangkalan Data Korpus DBP: Perancangan, Pembinaan dan Pemanfaatan. In Zaharani Ahmad. (Ed.), *Aspek Nahu Praktis Bahasa Melayu* (pp. 21-25). Bangi: Universiti Kebangsaan Malaysia Press.
17. Shahidi A. Hamid, Kartini Abd. Wahab, & Sa'adiah Ma'alip. (2018). *Kesinambungan Linguistik Melayu*. Bangi: Penerbit Universiti Kebangsaan Malaysia.
18. Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
19. Sinclair, J.M., & Renouf, A. (1988). A Lexical Syllabus for Language Learning. In R. Carter, & M. McCarthy. (Eds.), *Vocabulary and Language Teaching* (pp. 140-160). London; New York, NY: Longman.
20. Siti Aiesha Joharry, & Hajar Abdul Rahim. (2014). Corpus Research in Malaysia: A Bibliographic Analysis. *Kajian Malaysia*, *32*(1), 17-43.
21. Siti Syakirah Sazali, Nurazzah Abdul Rahman, & Zainab Abu Bakar. (In Press). Characteristics of Malay Translated Hadith Corpus. *Journal of King Saud University – Computer and Information Sciences*. https://doi-org.libproxy.upnm.edu.my/10.1016/j.jksuci.2020.07.011.
22. Tan, T., Xiao, X., Tang, E.K., Chng, E.S., & Li, H. (2009). MASS: A Malay Language LVCSR Corpus Resource. *Oriental COCOSDA International Conference on Speech Database and Assessments*, pp. 25-30.
23. Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam; Philadelphia, PA: John Benjamins.

CC BY