Data in Brief 36 (2021) 107013

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

House building tips (HBT) corpus dataset as a resource to discover Malay architectural ingenuity and identity[1]

Muhamad Fadzllah Zaini[a], Anida Sarudin[a,*], Mazura Mastura Muhammad[b], Zulkifli Osman[a], Husna Faredza Mohamed Redzwan[a], Muhammad Anas Al-Muhsin[c]

[a] Department of Malay Language and Literature, Faculty of Language and Communication, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia
[b] Department of English Language and Literature, Faculty of Language and Communication, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia
[c] Department of Modern Language and Literature, Faculty of Language and Communication, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia

ARTICLE INFO

Article history:
Received 19 November 2020
Revised 12 March 2021
Accepted 23 March 2021
Available online 27 March 2021

ABSTRACT

House Building Tips is the title of a classic text containing historical information on early house construction in Malay communities. These tips were written by a scholar with knowledge of house construction through observation of the surrounding environment. In Malaysia, written sources or records of house construction are scarce and underexposed. As such, this research was conducted to guarantee the written legacy of the construction of Malay houses. The purpose of this paper is to introduce a statistical data source of house building tips that is laden with Malay ingenuity and identity. The wordlists generated from this study can become a source of reference for the field of Malay architecture. Accordingly, this study utilises the quantitative method by applying the Linguistic Corpus Statistical Approach; these data utilise specific corpus development procedures, begin-

* Department of Malay Language and Literature, Faculty of Language and Communication, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia
E-mail addresses: mfadzllah@fbk.upsi.edu.my (M.F. Zaini), anida@fbk.upsi.edu.my (A. Sarudin), mazura@fbk.upsi.edu.my (M.M. Muhammad), zulkifli@fbk.upsi.edu.my (Z. Osman), husna.faredza@fbk.upsi.edu.my (H.F. Mohamed Redzwan), anas@fbk.upsi.edu.my (M.A. Al-Muhsin).

[1] DATASET HOUSEHOLD BUILDING TIPS (Original data) (TABLEAU PUBLIC)

https://doi.org/10.1016/j.dib.2021.107013

ning with text collection, scanning and cleaning processes, text annotation, and data storing in plain text. Next, the data analysis procedure utilises a corpus software, LancsBox, to generate specialised wordlists. The bubble graphs are developed based on these wordlists through the Tableau software, and illustrate the most used lexical items with the raw and relative frequency values. This facilitates searches for, and the reading of, architectural words and architectural word references. These data represent written sources that need to be preserved and become points of reference concerning Malay architectural ingenuity and identity.

Specification Table

Subject	Linguistics
Specific subject areas	Corpus Linguistics and Computational Linguistics.
Type of data Raw	Tables & Figures Visualisation Graphs
How data were acquired	The data were collected from seven classical Malay manuscripts about house building tips.
Data format	Raw and Analysed dataset
Parameters for data collection	Raw and Relative Frequency Wordlists with Corpus Linguistics Parameters (Normalisation per 10,000 words)
Description of data collection	A total of seven House Building Tips manuscripts were extracted and developed into corpus data. Total tokens for this data are 14648, 2079 types and Type/Token Ration (TTR) of 14.19.
Data source location	Institution: Sultan Idris Education University City: Tanjong Malim, Perak. Country: Malaysia
Data accessibility	https://public.tableau.com/profile/muhamad.fadzllah.zaini/
Related research article	Zaini, M. F., Sarudin, A., Muhammad, M. M., & Abu Bakar, S. S. [5]. Representatif Leksikal Ukuran sebagai Metafora Linguistik berdasarkan Teks Klasik Melayu (Representatives of Lexical *Ukuran* as Linguistics Metaphors Based on Malay Classical Texts). *GEMA Online® Journal of Language Studies*, 20(2), 168–187. https://doi.org/10.17576/gema-2020-2002-10

Value of the Data

- This dataset is useful in the construction of Malay architectural corpora.
- This dataset explores the Malay identity in the sphere of Malay architecture.
- This dataset can be utilised to determine lexical collocations in architectural discourse.
- This dataset is invaluable for linguists and lexicographers in generating architectural terms.
- This dataset can be utilised to observe metaphorical representations in the sphere of Malay architecture through the use of lexis.

1. Data Description

The *House Building Tips* (HBT) manuscript corpus is a specialised body of work encompassing seven manuscripts under the title *House Building Tips*. Baker, Hardie and McEnery [1] explained that a specialised corpus is specifically designed for a particular research project. An example

of a specialised corpus is the Malay Concordance Project (MCP), which is based at the National Australian University (NAS). The main feature of the MCP is that it is considered a specialised corpus of classical Malay texts, comprising 165 texts and 5.8 million words, including 140,000 verses. The MCP corpus does not represent a sample of a larger historical corpus.

Like the MCP, HBT is a specialised corpus that has been extracted from the archives of the National Centre of Malay Manuscripts, located at the Malaysian National Library. This centre has an estimated 4,884 original manuscripts that were handwritten between the 16th and the early 20th century. The knowledge captured in these Malay manuscripts covers various subjects, such as literature, history, religious texts, medicine, law, the constitution, charms and tips. One of these subjects is *House Building Tips*. Hence, the HBT manuscript corpus does not represent a sample of a larger historical corpus but can be a specific domain or genre, and is designed to represent a manuscript [2].

Ding Choo Ming [3] explains that the *House Building Tips* manuscripts are steeped in knowledge of the procedures of building the houses of the Malay community. Such information is invaluable and is considered part of the "Community Heritage Inventory" in the sphere of architecture [4]. Hence, these manuscripts have been selected to uncover the knowledge of our forefathers, particularly the ancient knowledge on building houses that they contain.

These manuscripts were transliterated from *Jawi* to romanised Malay by Abdul Rahman Al-Ahmadi, who was the Visiting Professor at the National Library of Malaysia from 1997 to 1998. Jawi is a writing system based on Arabic scripts that is used for writing the Malay languages, and other languages such as Acehnese and Banjarese. This transliteration was later revised by Associate Professor Dr Muhammad Anas Al-Muhsin in 2019 as part of the requirement of FRGS research grant 2019-0036-108-02. Each manuscript was then coded (MSS coding) as official information and referenced in the National Library [5]. During the corpus construction process, several aspects were taken into account, namely corpus annotation, retainment of borrowed words and corpus storage in *Notepad*, in order to facilitate data generation through LancsBox. The annotation process was used to differentiate lexical classes within the texts of the manuscripts, and all the transliterated manuscripts were stored in Notepad (.txt) format.

For this research, each text that has been acquired is a part of an open collection that can be used for research and publication purposes. The researchers acquired a copy of the HBT text, which was bought from the National Malay Manuscript Centre in the National Library of Malaysia. This publication will be presented to the centre as a record of the research, as stipulated in the National Library Malay Manuscript Policy under 'No. 7.7 Use and Research'. The *House Building Tips* manuscript is one of the titles from the centre's open collection.

The HBT corpus annotation was conducted through the *engine.malaynlp* website (generated on April 25, 2019) which is run by the Faculty of Science and Technology, National University of Malaysia. Part-of-speech (POS) tagging was performed whereby the transliterated manuscripts were uploaded onto the raw data space and the newly tagged manuscripts were then fed onto the tagging space provided. POS tagging was a fundamental step in ensuring that the dataset was ready to be analysed and classified using corpus linguistics analysis tools and visualisation graphs.

As aforementioned, this dataset encompassed classical Malay manuscripts under the title of *House Building Tips* (HBT), including the Book of *Tajul Muluk*, the Book of *Abu Mas'yar*, and manuscripts entitled MSS714, MSS1415, MSS1521, MSS1849, and MSS2001. The book of Abu Masyar Al-Faliki was written by Syekh Abu Hayyillah al-Marzuqi, a Muslim astrologer who was also known as Abu Ma'shar al-Balkhiy or Abu Ma'shar al-Falaqi. The Book of*Tajul Muluk* was translated by Faqir Tuan Haji Wan Hassan Ibn Sheikh Tuan Ishak Fatani. However, there were no author entries that had been recorded in the National Library of Malaysia on the other manuscripts, namely MSS 741, MSS1415, MSS1521, MSS1849, and MSS2001. Table 1 shows the information for the indexed manuscripts.

The representativeness of a specialised corpus, at the lexical level at least, can be measured by the degree of 'closure' or 'saturation' of the corpus [2]. Hence, this dataset utilised the simple random sampling technique [5], where chunks of samples were extracted from the initial, middle, and end of each text [2,6]. Additionally, all tables, figures and illustrations were eliminated.

Table 1
Information on the HBT manuscript corpus.

No. Manuscript	Title	Author / Officer / Copywriter	Date/Year	Location Manuscript	The First Year of Publication
Book of *Abu Ma'syar*	*House Building Tips*	Author; Syekh Abu Hayyillah al-Marzuqi	Not specified	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
MSS741	*House Building Tips*	Manuscript Officer: Nor Azela Abdul Samad	[BETWEEN 1600 AND 1800]	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
MSS1415	*House Building Tips*	Manuscript Officer: Nor Azela Abdul Samad	Not specified	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
MSS1521	*House Building Tips*	Manuscript Officer: Nor Azela Abdul Samad	[BETWEEN 1600 AND 1800]	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
MSS1849	*House Building Tips*	Manuscript Officer: Nor Azela Abdul Samad	Not specified	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
MSS2001	*House Building Tips*	Manuscript Officer: Nor Azela Abdul Samad	[BETWEEN 1600 AND 1800]	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000
Book of *Tajul Muluk*	*House Building Tips*	Copyist; Faqir Tuan Haji Wan Hassan Ibn Sheikh Tuan Ishak Fatani	Not specified	National Library of Malaysia	Abdul Rahman Al-Ahmadi; 2000

Table 2
Technical information of the manuscripts.

No.	Manuscript	File Size (bytes)
1	Book of *Abu Ma'syar*	1986
2	MSS741	21814
3	MSS1415	4361
4	MSS1521	6147
5	MSS1849	4870
6	MSS2001	1454
7	Book of *Tajul Muluk*	110174

Table 3
Basic statistics of the *House Building Tips* (HBT) Corpus.

Manuscript	Tokens	Types	Type/Token Ratio
Book of *Abu Ma'syar*	145	84	57.93
MSS741	3465	673	19.42
MSS1415	658	189	28.72
MSS1521	910	363	39.89
MSS1849	728	212	29.12
MSS2001	212	79	37.26
Book of *Tajul Muluk*	8526	1619	18.98

Therefore, for the Book of *Tajul Muluk* and the Book of *Abu Mas'yar*, the sample data were taken from clauses 87, 90, 91, 95, 96, 97, 99, 100, and 103. The Book of *Abu Mas'yar* was sampled from clause 13. For manuscripts MSS714, MSS1415, MSS1521, MSS1849, and MSS2001, samples were extracted from the initial, middle, and end of the manuscripts.

Table 1 illustrates the content of the HBT Corpus which refers to the file size of the data during the data generation process. The technical information of the HBT content was generated based on the file size (refer to Table 1).

Table 2 shows that the biggest file size was the Book of *Tajul Muluk*, with 110,174 bytes, followed by MSS741 with 21,814 bytes. The smallest file size was MSS2011 with only 1,454 bytes. The file size was determined based on the setting in the document properties. The results of the raw data were then generated and tabulated through LancsBox (refer to Table 3). These basic statistics were based on the measurements that were pre-set in LancsBox.

Table 3 displays the composition of the HBT Corpus with the number of words (tokens), the number of different words (types) and the type/token ratio (TTR). 'Tokens' are the total number of words in a corpus, whereas 'types' refers to the different types of lexical items in a corpus. In addition, the type/token ratio (TTR) shows the relationship between the total number of types and the total number of tokens in the manuscripts. The TTR in Table 2 was representative of the content quality of the HBT Corpus with the overall word counts compared to the different word types [7].

Below is the formula for TTR:

$$\frac{number\ of\ Types}{number\ of\ Tokens}$$

Hence, the TTR for MSS741 manuscript was measured below:

$$\frac{673}{3465} = 0.1942279\ or\ 19.42$$

The TTR is a number of types (unique words) in a text, which is divided by the number of tokens. A high TTR suggests that a text is lexically diverse, whereas a low TTR suggests that there is an extensive repetition of lexical items in a file [1].

Tables 4 and 5 show how the type calculation was performed within the manuscript itself (Table 4) as well as within whole or entire manuscripts (Table 5). For manuscripts MSS741 and MSS1415, the number of types was five. For the entire manuscripts, the number of types was

Table 4
Type calculation method within manuscripts.

Manuscript	W1	W2	W3	W4	W5	Types	Total Types	Token
MSS741	rumah	Tanah	tempat	dibuat	dusun	5	5	5
Manuscript	W1	W2	W3	W4	W5	Types	Total Types	Token
MSS1415	rumah	Jadi	baik	dan	hormat	5	5	5

*[W1= Word 1, W2= Word 2, W3= Word 3, W4= Word 4, W5= Word 5].

Table 5
Type calculation method within entire manuscripts.

Manuscript	W1	W2	W3	W4	W5	Types	Total Types	Total Token
MSS741	*rumah* 1	*tanah* 2	*tempat* 3	*dibuat* 4	*dusun* 5	5	9	10
MSS1415	*rumah* 1	*jadi* 6	*baik* 7	*dan* 8	*hormat* 9	4		

*[W=Word].

nine, since the lexical type of *rumah* (house) was not recalculated, and thus was considered as only one type. The total number of tokens was ten (total number of words).

Absolute Frequency (AF) and Relative Frequency (RF) of the wordlist were represented in the form of bubble graphs using the Tableau software. The graph displays large and small bubble visualisations based on these word frequencies. Before generating the bubble graphs, data of the word frequency list is stored in the in the form of .cvs file to facilitate the generation process using the Tableau software. This .cvs file contains wordlist information. The overall display of HBT data is shown and represented in a bubble graph visualisations. The HBT data can be accessed via https://public.tableau.com/profile/muhamad.fadzllah.zaini/.

2. Experimental Design, Materials and Methods

In the construction of the wordlists from the HBT Corpus, the quantitative design was utilised by applying the Frequency Wordlist approach [6]. Based on the wordlists, two forms of frequency lists were generated: the Absolute (or Raw) Frequency (AF) and the Relative (or normalised) Frequency (RF). AF is a simple statistical form of frequency that is generated by observing the actual occurrence of a specific word in a corpus. It is useful to resort to AF when researchers are looking at a single corpus. For example, AF can be used to determine the most frequent temporal (days, weeks, months, etc) nouns in the MSS741 manuscript. This will yield interesting findings, since ancient Malay architecture emphasised the importance of time, day and month (temporal factors) before the construction of a house or building. RF, on the other hand, refers to a normalised lexical frequency, and it is used to compare between corpora, especially those of different sizes. The following is the formula to measure RF:

$$(relative\ frequency\ (of\ a\ specific\ word)) = \frac{(raw\ frequency\ of\ a\ specific\ word)}{(total\ token\ in\ the\ corpus)} \times basis\ of\ normalisation$$

For example, the RF for the temporal noun *bulan* (month) was calculated as below:

$$relative\ frequency\ (bulan) = \frac{125}{3465} \times per\ 10{,}000\ words$$

$$relative\ frequency\ (bulan) = 360.750361$$

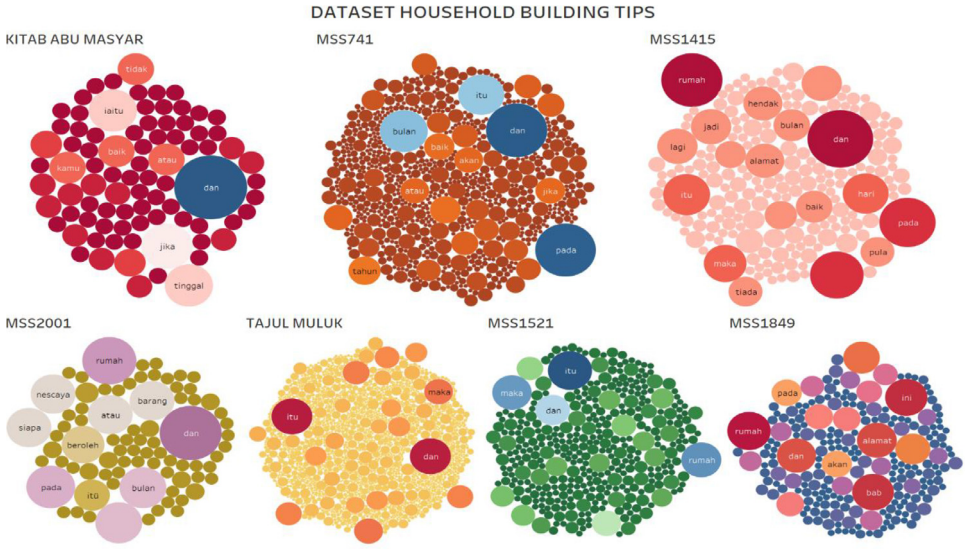

Fig. 1. HBT corpus dataset.

The LancsBox software pre-set the use of RF at 'per 10,000' words/tokens.

The generated wordlists of each manuscript were saved in Excel format (.xlsx). This simplified the process of generating visualisation graphs on Tableau. The information that was yielded from the visualisation graphs were divided into three main elements: lexical, AF and RF. Using the Tableau software, the visualisation graphs were presented in packed bubbles that displayed the differences between the manuscripts, based on colours and degrees of frequency.

As previously mentioned, the total number of word types for the Book of *Abu Masyar* was 84 words, MSS741 (673 words), MSS1415 (189 words), MSS1521 (363 words), MSS1849 (212 words), MSS2001 (79 words) and Book of *Tajul Muluk* (1619 words). These word types were plotted based on the three elements (lexical, AF and RF), according to the degrees of frequency. The use of colours on the packed bubbles was also prominent, as it indicated the distinctive frequency of information of each manuscript.

Fig. 1 displays the HBT dataset generated through LancsBox, and represented in the form of visualisation graphs using Tableau. The bigger bubbles show the high-frequency lexical items, while the smaller bubbles indicate low-frequency lexical items.

The purpose of the study is to examine the lexical variations based on classical Malay manuscripts, namely *House Building Tips* (HBT). The HBT Corpus is a lexical resource which is invaluable for an extensive range of research on classical lexical variations and the national identity of Malay architecture. This dataset emphasises two aspects, namely the significance of architectural trends, and the architectural lexical trend that was applied in the architecture industry in the past. Such knowledge can be evaluated by examining the present needs of architectural trends pertaining to national identity [8]. This dataset enables architects and literary scholars to refer to the scenario or phenomenon of previous civilisations with various form of applied knowledge [9]. The construction of this data is in accordance with the specialised corpus data proposed by Biber, Conrad and Reppen [10].

The use of this data is significant in several contexts:

1. National Identity Architecture Dictionary
2. Knowledge List of Malay Architecture
3. Malaysian Architects Board Reference based on Architectural Trends

4. Reference for Malay language corpora such as the Malay Concordance Project, Sketch Engine (Malay Text Corpora), UKM-DBP, Leipzig German (Malay Language Media Corpus), and the Malay Speech Corpus (UK Essays).

The three aforementioned aspects provide a novel way of tracking the history and civilisation of historic Malay architecture. Notably, the HBT corpus dataset can be a reference for the current generation and allow the focus of national identity, not only in the design but also the selection of the materials, as well as the importance of the household in performing the daily routines of life.

CRediT Author Statement

Muhamad Fadzllah Zaini: Generation and tabulation using corpus linguistics software, Visualisation, Writing - Original draft preparation; **Anida Sarudin:** Text Annotation; **Mazura Mastura Muhammad:** Conceptualisation, Writing - Reviewing and Editing; **Zulkifli Osman:** Data Scanning; **Husna Faredza Mohamed Redzwan:** Text Collection; **Muhammad Anas Al Muhsin:** Transliteration.

Ethics Statement

The work not involves data collected from social media platforms, animal experiments, and the use of human subjects.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work that is reported in this article.

Acknowledgements

This research has been conducted based on the Fundamental Research Grants Scheme (FRGS/1/2018/WAB04/UPSI/02/5) that was granted by the Ministry of Education of Malaysia. Much appreciation goes to the National Malay Manuscript Centre, the National Library of Malaysia.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107013.

References

[1] P. Baker, A. Hardie, T. McEnery, A Glossary of Corpus Linguistics, Edinburgh University Press, Edinburgh, 2006.
[2] T. McEnery, R. Xiao, Y. Tono, *Corpus-Based Language Studies An Advanced Resource Book* (First Edit), Routledge, New York, 2006.
[3] Ding Choo Ming, *ManuskripMelayu: Sumber Maklumat PeribumiMelayu* (Four), UniversitiKebangsaan Malaysia Press, Bangi, 2016.
[4] A.N. Farhana, A. Faizah, A.A. Shah, Heritage place inventory: A tool for establishing the significance of places, J. Des. Built Environ. 15 (1) (2015) 15–23, doi:10.22452/jdbe.vol15no1.3.

[5] M.F. Zaini, A. Sarudin, M.M. Muhammad, S.S. Abu Bakar, RepresentatifLeksikalukuransebagaimetaforalinguistikberdasarkan teks klasikmelayu (representatives of lexical ukuran as linguistics metaphors based on malay classic text), GEMA Online® J. Lang. Stud. 20 (2) (2020) 168–187, doi:10.17576/gema-2020-2002-10.
[6] V. Brezina, Statistics in Corpus Linguistics, Cambridge University Press, 2018 In(First), doi:10.1017/9781316410899.
[7] V. Davis, Types, tokens, and hapaxes: a new heap's law abstract, Glottotheory 9 (2) (2019) 113–129, doi:10.1515/glot-2018-0014.
[8] H. Baharum, Landskapsebagai mekanisme pertahanan diri berdasarkan karya sastera melayu tradsional, PENDETA J. Malay Lang. Educ. Lit. 6 (2015) 244–259.
[9] F. Soleimani, Alignment of accountability and language planning and policy in the third millennium, Asian J. Eng. Lang. Pedagogy 8 (1) (2020) 8–17.
[10] D. Biber, S. Conrad, R. Reppen, *Corpus Linguistics Investigating Language Structure and Use* (Fifth), Cambrige University Press, New York, 2006.