PERSATUAN LINGUISTIK MALAYSIA
Bahasa Sendi Tamadun Bangsa

www.plm.org.my

JURNAL LINGUISTIK Vol. 25 (1) Mei 2021 (066 - 079)

The Development of the Malaysian Hansard Corpus: A Corpus of Parliamentary Debates 1959-2020

[1]Imran Ho Abdullah, [2]Anis Nadiah Che Abdul Rahman & [3]Azhar Jaludin
[1]imranho@ukm.edu.my, [2]anisnadiahcar@gmail.com, [3]azharj@ukm.edu.my

Centre for Language and Linguistics, Faculty of Social Science and Humanities, Universiti Kebangsaan Malaysia

Tarikh terima : 10 April 2021
Received
Terima untuk diterbitkan : 14 Mei 2021
Accepted
Tarikh terbit dalam talian : 30 Mei 2021
Published online

Abstract

Parliamentary corpora are pertinent language resources of various subject matters. The applications of parliamentary corpora enable researchers to analyse data from different perspectives including linguistics, political sciences, computational linguistics or history. The availability of parliamentary corpora in specific languages facilitates towards the analysis of that particular language. In Malaysia, parliamentary corpus has not been developed and Malaysian Hansard portal could not be comprehensively used to analyse linguistics patterns, semantic shift or discourse analysis. Malaysian Hansard Corpus (MHC) was initially created to provide a comprehensive and digitally accessible parliamentary documents written in the Malay and English language, compiled from Malaysian Parliamentary Reports (Malaysian Hansard). MHC contains documents from the House of Representative (*Dewan Rakyat*) in Malaysian Parliament. The dates of the documents ranges from Parliament 1 (1959) to the present's Parliament 14 (2019). This initial version of digital collection comprises of 167,513,039 tokens (word count) gathered from 3,684 scanned files in PDF format. The corpus is readily accessible for public use in txt and PDF formats and has its own distinctive typology. The availability of this corpus will promote more statistical analyses and hypothesis testing, enable the testing of occurrences, frequencies or validating the rules of linguistic within an explicit language dominion.

Keywords: Corpus development; corpus linguistics; parliamentary corpus; Malaysian Hansard; Malaysian Hansard Corpus

Perkembangan Hansard Corpus Malaysia: Kumpulan Perbahasan Parlimen 1959-2020

Abstrak

Korpus parlimen merupakan satu sumber bahasa yang mempunyai perkaitan dengan banyak perkara. Aplikasi korpus parlimen membolehkan para penyelidik untuk menganalisis data dari pelbagai perspektif termasuklah linguistik, sains politik, linguistik komputer mahupun sejarah. Ketersediaan korpus parlimen dalam bahasa tertentu dapat memudahkan analisis dalam bahasa tersebut. Di Malaysia, korpus parlimen belum lagi dikembangkan dan portal Hansard Malaysia tidak dapat digunakan secara komprehensif bagi menganalisis pola linguistik, pertukaran semantik ataupun analisis wacana. Malaysian Hansard Corpus (MHC) pada mulanya dicipta untuk menyediakan dokumen parlimen yang komprehensif dan boleh diakses secara digital yang ditulis dalam bahasa Melayu dan Inggeris, yang disusun dari Laporan Parlimen Malaysia (Hansard Malaysia). MHC mengandungi dokumen dari Dewan Rakyat di Parlimen Malaysia. Tarikh dokumen yang dikumpul adalah di antara Parlimen 1 (1959) hingga Parlimen 14 yang terkini (2020). Versi awal koleksi digital ini terdiri daripada 167,513,039 token (jumlah perkataan) yang dikumpulkan daripada 3,684 fail yang diimbas dalam format PDF.

Korpus ini mudah diakses untuk kegunaan umum dalam format txt dan PDF dan mempunyai tipologi tersendiri. Keberadaan korpus ini akan membolehkan dan memberi galakan bagi penulisan analisis-analisis kuantitatif dan ujian hipotesis, ujian kejadian, kekerapan ataupun mengesahkan peraturan linguistik dalam satu domain bahasa yang eksplisit.

Kata kunci: *pembinaan korpus; linguistik korpus; korpus parlimen; Hansard Malaysia; Korpus Hansard Malaysia*

1.0 Introduction

A parliament is a legislative body of government and serves as a platform to represent the body of voters, examine bills and pass them into laws and policies, amend laws and policies, and administer the government. However, the task of parliament may show a discrepancy from country to country. In Malaysia, the Parliament of Malaysia embodies the government's democratic structure and echoes the people's aspirations through representatives elected in the election known as the Member of Parliament (henceforth MP). (Official Portal of Parliament of Malaysia, 2020). The Parliament of Malaysia comprises of three components; His Majesty the King, Senate and House of Representative.

The discussions, agreements and debates of the MPs take part in the Parliament (the building) during Parliamentary proceedings, also known as parliamentary debates. Parliamentary debates are recorded and transcribed by parliamentary administrators and are officially published as Parliamentary proceedings called Official Report or Hansard (Official Portal of Parliament of Malaysia, 2020). In Malaysia, Hansards are housed in Official Portal of Parliament of Malaysia.

Hansard contain parliamentary debates of the MPs for over a period of time. In the case of Malaysia, from Parliament 1 in 1959 to Parliament 14 (still on going in 2020). Hansard offers an access to data regarding sentiments and preferences of the MPs over the time. According to Blätte and Blessing (2018), parliamentary debates are invaluable to study language variations across political domains and language change. Due to that, Hansard captivates the interest from various of backgrounds to use it.

Hansards are normally made available publicly in a country's official Parliament's website and can be utilised for different purposes. For example, UK Parliament (https://hansard.parliament.uk/), Parliament of Canada (https://www.ourcommons.ca/documentviewer/en/35-2/house/hansard-index), Parliament of Australia (https://www.aph.gov.au/parliamentary_business/hansard) and the Official Portal of Parliament Malaysia (https://www.parlimen.gov.my/hansard-dewan-rakyat.html?uweb=dr&)

Despite its existence and accessibility, Hansard transcripts can be massive. British's House of Common for example, has historical contents from 1803 and contains a fair amount of words. Hansards may have been digitised but the availability in parliamentary website is insufficient for different type of exploitations. According to Abercrombie and Batista-Navarro (2018), large quantity of data, distinctive discourses and use of jargons can impede manual interpretation by non-experts. Due to that, many researchers have developed parliamentary corpora of various countries to further use the data more efficiently. For example, Hansard Corpus (Davies, 2015), The Icelandic Parliamentary Corpus (Steinþór, Helgadóttir, Rögnvaldsson, Barkarson and Guðnason, 2018) and Polish Parliamentary Corpus (Maciej, 2018). There are also parallel corpora of various languages that have been developed. A parallel corpus comprises of a collection of translated texts from the original language. The example of parallel corpora includes Hansard French/English (Roukos, Graff and Melamed, 1995), European Parliament Proceedings Parallel Corpus 1996-2011 (Philipp, 2005). Previous studies that make use of Hansard have provided a range of literature in various fields. e.g. in discourse (Berracheche, 2020), sociolinguistic (Offersgaard, 2018) and Natural Language Processing (NLP) (Rheault, Beelen, Cochrane and Hirst (2016).

A corpus or its plural corpora, is a collection of natural texts compiled from written or spoken speech or writings, constructed for specific purposes, and is stored in digital format. A corpus can be analysed using software tools including WordSmith Tools (Scott, 1996 – 2012) and AntcConc (Anthony, 2004 – 2012). To date, there is no parliamentary corpus in Malaysia. Therefore, there is a need to develop a new parliamentary corpus based on the Malaysian Hansard by producing a corpus

comprising all digitally available parliamentary documents written in Malaysian setting and discourse.

The aim of this paper is to present the development of the Malaysian Hansard Corpus (henceforth, MHC) in term of extraction, organisation and structure and propose a model for constructing a domain-specific parliamentary corpus. The key goal of this corpus development is to provide a publicly accessible portal for the Malaysian Parliamentary reports to be reachable in its original and official version (PDF format) and in plain text files (txt). The corpus was initially developed to be beneficial as a data set in cross-multidisciplinary fields. It can further be utilised for research in corpus linguistic, sociolinguistics, discourse analysis, political science and any other possible scientific disciplines including Culturomics or the study of cultural trends and human behaviour that can be done in quantitative analysis of digitized text. This corpus provides the history of Malaysia, the revolutions of language use in Malaysia and the issues covering 6 decades of time since Malaysia's independence in 1957. This corpus was developed according to the criteria by Sinclair (2004), Wynne (2005) with close reference to DutchParl (Marx & Schuth, 2010).

2.0 Related Work

2.1 Development of Parliamentary Corpus

A considerable amount of literature has been published on the development of corpora in Parliamentary discourse. Davies (2015) developed Hansard Corpus from British Parliament (1803-2005). It consists of approximately 1.6 billion words and it is housed in https://www.english-corpora.org/hansard/. The corpus offers the end-user to see frequency of words, phrases and grammatical construction over time, compare words, phrases or collocates (nearby words) between decades, compare the collocates, see the Key Word in Context (KWIC) or the concordance lines or the examples from the context, see semantic tags and create personal virtual corpora by personalising texts/ opinion related to a researcher's interest.

2.2 Available Malay Corpora

In Malaysian context, there are several corpora available. Imran, Zaharani, Rusli, Nor Hashimah and Idris (2004) developed Malay Practical Grammar Corpus (MPGC) from Dewan Bahasa dan Pustaka (DBP) database, consisting of newspaper, magazines and books. This general corpus or a corpus that represents general language use of one particular language has 5 million words. In addition to that, Zuraidah (2010) developed MALEX (Malay Lexicon) consisting of 30,000 different orthographic words. The data was taken from a set of novels (80,00 words), newspaper corpus (5 million words), a corpus of speeches from former Prime Minister Tun Dr. Mahathir Mohamad (1.3 million words) and academic texts (20,000 words).

In addition to that, there are other Malays specialised corpora that have been constructed by various researchers. A specialised corpus contains texts from specific genre or index, or has a specific time or context. Suhaimi and Normaziah (2004) developed an English-Malay bilingual parallel corpus consisting of 250,000 words. This specific and parallel corpus covers the domain of agriculture and health. Lee and Low (2011) constructed a corpus based on Malay textbooks for primary schools for teaching and learning corpus. The Online Malay Language Word Corpus for Primary Schools or MyBaca contains 4035 words.

There are also parallel corpora in Malay language. Tengku Sepora, Vaezian, Akbari, Nor Aini and Cheng (2009) for example, built English-Malay Parallel Corpus of Legal Texts consisting of 210,000 words and overall with translation of 500,000 words. Norsimah, Intan Safinaz and Imran (2013) developed a parallel corpus from businesses' annual reports. The corpus is known as UKM Parallel Annual Reports Translation Corpus or UKM PART Corpus and consists of 158,620 words in English and 154,148 words in Malay. Based on previous literature, there is no evidence that there is a corpus on parliamentary discourse yet.

2.3 The Application of Parliamentary Corpora

A large and growing body of literature has employed the use of parliamentary corpora over the times in various disciplines. The use of parliamentary corpora. Rheault, Beelen, Cochrane and Hirst (2016) used parliamentary corpus of the United Kingdom and analysed the data using natural language processing methodology to measure emotion or parliamentarians during parliamentary debates. The study suggests that politicians' mood has become more positive over times and variations of emotional polarity can be seen to predict the state of national economy. Offersgaard (2018) studied gender differences in Danish Parliament from 2009 to 2017. The analysis indicates a statistically differences of speech duration according to gender and age. Bachman (2011) utilised keyword analysis to study the language use in UK Parliament regarding the approval of civil partnership by revealing same-sex relationships. Hiltunen, Räikkönen and Tyrkkö (2020) used a pattern-driven approach and emphasize on n-gram frequencies to study sociocultural changes in British Hansard from 1870–1930. Berracheche (2020) employed corpus-driven study to evaluate discourse revolving asylum seekers in Australian parliamentary debates to see how the Australian MPs expressed unfavourable attitudes towards the issue.

3.0 Methodology

This section presents the terminology and technical descriptions used in this study, the procedures to extracting data, the compilation of data, methodology to combine and organise files.

3.1 Terminology and Technical Descriptions

This section will further describe and summarised on the technical terms used in this study.

Terminology	**Description**
Corpus	a large collection of written texts collected for specific purpose. E.g. general corpus to represent general use of language, specialised corpus of parliament to represent the use of parliamentary discourse.
Token	The number of word in a corpus
Type	Number of different words used in a corpus
Portable Document Format (PDF)	A file format developed by Adobe. It can be used and transferred crossing multi platforms
txt	Plain text file format - a file extension for a text file
Noise	Text that are different from the source after conversion/ processing. A noisy text derived from the difficulty of a software to decipher the exact words/ code. E.g. the conversion of earliest documents converted the word "*kerja*" (work) into "*keija*" or "*ketja*"
Markup	A system of accustomed codes. Markup is added into a document to give particular information about the word/ sentences or section in the corpus. E.g. "<MD… /MD>" id added into the corpus to note the main page (*Muka Depan*) in the corpus. Markups will be useful for further processing
Annotation	An additional system of interpretative linguistics information. Annotation includes the insertion of tags to indicate word class or semantic labels.
Parliament	The Parliament is the supreme Legislative Authority of the nation consisting of His Majesty The Yang di-Pertuan Agong, the Senate and the House of Representatives. Parliament also refers to general parliamentary proceedings from its opening time until it is prorogued to next general election
Session (*Penggal*)	A Parliamentary Session begins from the opening of new Parliament until it is prorogued
Meeting	Meeting refers to the meeting. Each session has three to four meeting and a special meeting. A meeting is normally conducted in a span of a month.
Sitting	Sitting refers to the meeting in each day or parliamentary proceeding

3.2 Data Source

MHC utilised the data (Hansard) provided in the Official Portal of Parliament Malaysia. The data is in Portable Document Format (henceforth PDF). Hansard Malaysia is categorised as an open data that can be publicly used. A total number of 3,684 PDF files was downloaded. Before turning into the corpus, the data retrieved from the Official Portal of Parliamentary Debates Malaysia at https://www.parlimen.gov.my/hansard-dewan-rakyat.html?uweb=dr& was stored individually in the portal. The retention of the retrieved data can be depicted in the figure 1:

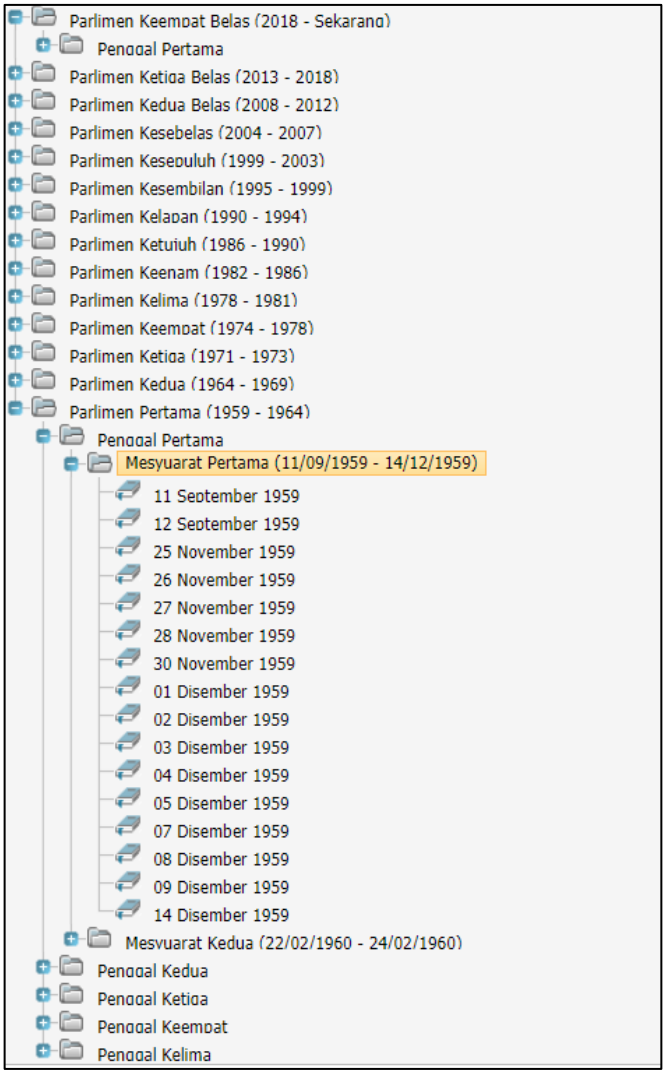
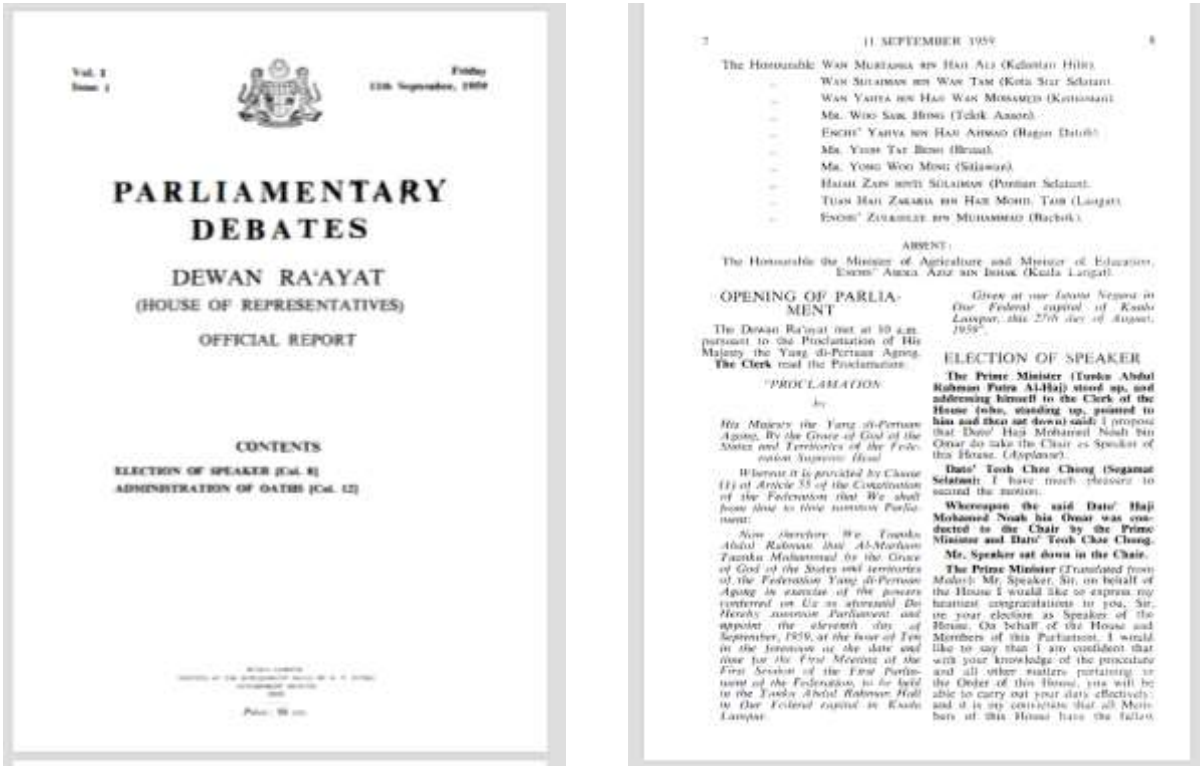


Figure 1: The storage of parliamentary reports in the Official Portal of Parliamentary Debates Malaysia

Figure 1 shows the storage of parliamentary reports in the Official Portal of Parliamentary Debates Malaysia. Each of the parliament is tagged with the date of parliamentary proceedings. Each PDF file is labelled according to the date of the sitting.

Original format of Malaysian Hansard in PDF



PARLIAMENTARY DEBATES

DEWAN RA'AYAT

(HOUSE OF REPRESENTATIVES)

OFFICIAL REPORT

CONTENTS

ELECTION OF SPEAKER [Col. 8]

ADMINISTRATION OF OATHS [Col. 12]

11 SEPTEMBER 1959

OPENING OF PARLIAMENT

ELECTION OF SPEAKER

Figure 2: Malaysian Hansard in PDF file, (Parliament 1, Session 1, Meeting 1, Sitting 1)

Figure 2 shows the extract of the original file of Malaysian Hansard in PDF file. The main page generally contains the volume and issue number, the day and date, the Coat of Arms of Malaysia, the title, the contents and its pages and the publisher. After the main page is normally the attendance list followed by the content. The new parliament will begin after every general election and the opening of parliament will be done by His Majesty the Yang di-Pertuan Agong. The structure of parliamentary reports in general will be clarified in the following section.

3.3 Structure of the parliamentary reports

The reporting structure differs according to years. Typically, Malaysian Parliamentary reports comprises of the following structure:

(i) Front page (Title, content)
(ii) Attendees
(iii) Prayers Written as "prayers" only
(iv) Announcement by Mr. Speaker
(v) Speech by the His Majesty the Yang di-Pertuan Agong (for opening only)
(vi) Oral answers to questions
(vii) Draft Standing Orders (*Dewan Rakyat*)
(viii) Bills presented
(ix) Motions
(x) Adjournment

3.4 The procedure:

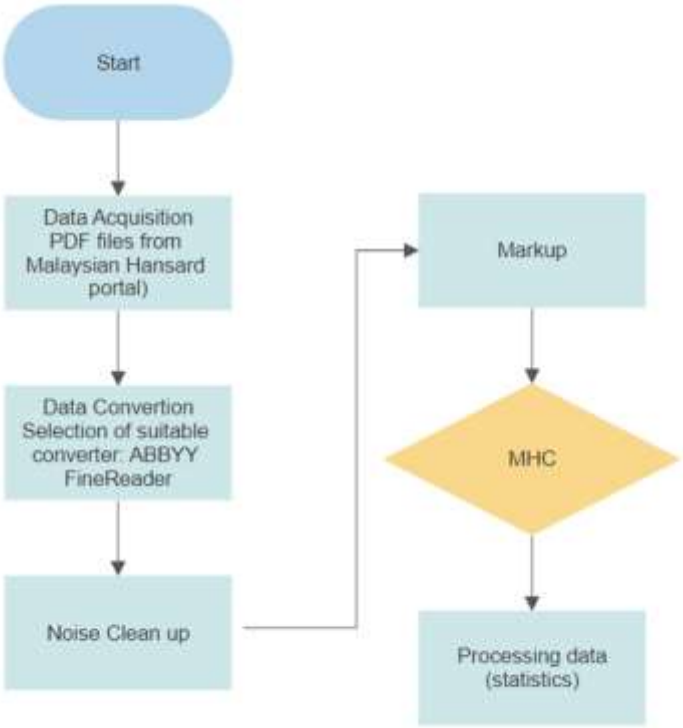


Figure 3. The procedure to Develop MHC

According to Figure 3, MHC begins with data acquisition. The data was retrieved from the Official Portal of Parliamentary Debates Malaysia at https://www.parlimen.gov.my/hansard-dewan-rakyat.html?uweb=dr& . Retrieved data was in PDF format hence the conversion had to be done to convert the file in plain text file. The conversion was done to further process the files and to add further tagging or annotations. Before the conversion, the selection of suitable converter was done as the conversion of files are in high volume. ABBYY FineReader tool was selected to convert all retrieved files due to its suitability to convert multiple PDF files in high volume with higher reliability compared to other tested converters (Anis Nadiah, Imran, Intan Safinaz and Azhar, 2019).

Following the conversion, the files underwent noise clean up. This step is important to check the preciseness of the conversion and to ensure that the converted data are able to be processed. After the clean-up, simple markup was added to each of the files in the corpus. The mark ups include <MK…/MK> for *Muka Depan* (Front Page) and <SK…/SK> for *Senarai Kehadiran* (Attendance list). After the markup process, MHC is ready to be utilised. The corpus was then analysed using WordSmith Tool 5.0 (Scott, 2008) to generate its statistical analysis.

4.0 Findings

4.1 Coverage and Size of MHC

The storage of MHC is generally stored according to sub-corpora in different folders. Coverage of corpus includes the parliaments, sessions, meetings and sittings covered in MHC. In each type, there are 14 sub-corpora, based on the numbers of Parliamentary sessions that have been conducted in Malaysia from 1959-2016. In each sub-corpora, there are smaller sub-corpora which are divided into the sessions and meetings. The present version of MHC contains 3,630 files from Parliament 1 to Parliament 14 (last measured August 2020).

Table 1: The Divisions of House of Representative according to Parliaments and Sessions

Par-liament	**Session**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**Grand Total**
1	**M1**	10	46	43	29	42	50	43	13	13	4	17	18	19	20	
	M2	2	4		34	66	59	20	95	13		10	16	8	30	
	M3							39		40		41	45	27		
	M4									4			1			
2	**M1**	40	45	39	70	59	66	65	65	30	39	23	22	20	20	
	M2	6	5	11	2				31	14	2	52	12	8	12	
	M3		10							40	36	2	36	28	36	
	MK													1		
3	**M1**	24	14	45	57	64	60	64	41	32	27	34	24	20	1	
	M2	22	42		4					10	16	14	24	46	25	
	M3									44	34	43	35	25	29	
	MK													2		
4	**M1**	34	31		71		21	24	67	26	17	26	19	20		
	M2	4								16	8	17	12	8		
	M3									30	36	42	32	25		
5	**M1**	32	14							21	8		22	20		
	M2	8	30							10	59		12	12		
	M3									15			36	25		
6	**M1**													20		
TOTAL		182	241	138	267	231	256	255	276	358	286	321	366	334	**173**	**3,684**

*M denotes meeting, MK denotes Special Meeting

Table 1 shows the Divisions of House of Representative according to Parliaments and Sessions. There are 14 parliaments altogether. Each parliament has 5-6 sessions. In each session, there are 1 to 3 meetings. In an atypical occasion, there are an additional of Special Meeting in each session. Each meeting has a number of sittings which ranges from 2 to 66 sittings. Parliament 3 has the least number of sittings (138) while Parliament 12 has the most number of sittings (366). A sitting would normally be conducted in a day. The total number of sittings that have been conducted in Malaysian Parliament has exceeded 3,630 from Parliament 1 to Parliament 14.

Table 2: Division of Basic Statistical Information according to Parliaments

Parliament	tokens (running words) in text	types (distinct words)	type/token ratio (TTR)	standardised TTR	mean word length (in characters)	sentences	mean sentence length (in words)
P1	6060551	50203	0.8599	35.4964	4.7687	296093	19.7191
P2	9893721	66175	0.7034	35.1139	4.8546	498285	18.8777
P3	6264859	59224	0.9832	37.4615	5.1567	366898	16.4181
P4	8040934	79824	1.0274	38.2063	5.3888	501449	15.4942
P5	8691728	72942	0.8651	37.2059	5.6203	570632	14.7744
P6	9485250	67406	0.7350	37.1455	5.6888	642935	14.2637
P7	9106187	95112	1.0813	37.8431	5.6341	656096	13.4056
P8	15171864	127432	0.8678	38.0783	5.5949	1089354	13.4794
P9	12919341	95795	0.7597	37.2275	5.5674	1013445	12.4454
P10	14123916	72954	0.5289	36.5835	5.5432	1074310	12.8340
P11	17047556	70272	0.4235	36.2619	5.5285	1169646	14.1940
P12	22188820	74818	0.3472	36.4894	5.5294	1834140	11.7436
P13	18517944	72415	0.4036	36.5704	5.5453	1329088	13.4991
P14	10000368	49947	0.5168	36.7907	5.5800	744794	12.9754
Total	**165061050**						

(Generated using WordSmith 5.0)

Table 2 shows the Division of Basic Statistical Information according to Parliaments. MHC in general covers a number of 167,513,039 of token (generated using WordSmith 5). Token is the number of words in a corpus. The tabulation of token can be visualised in figure 4.

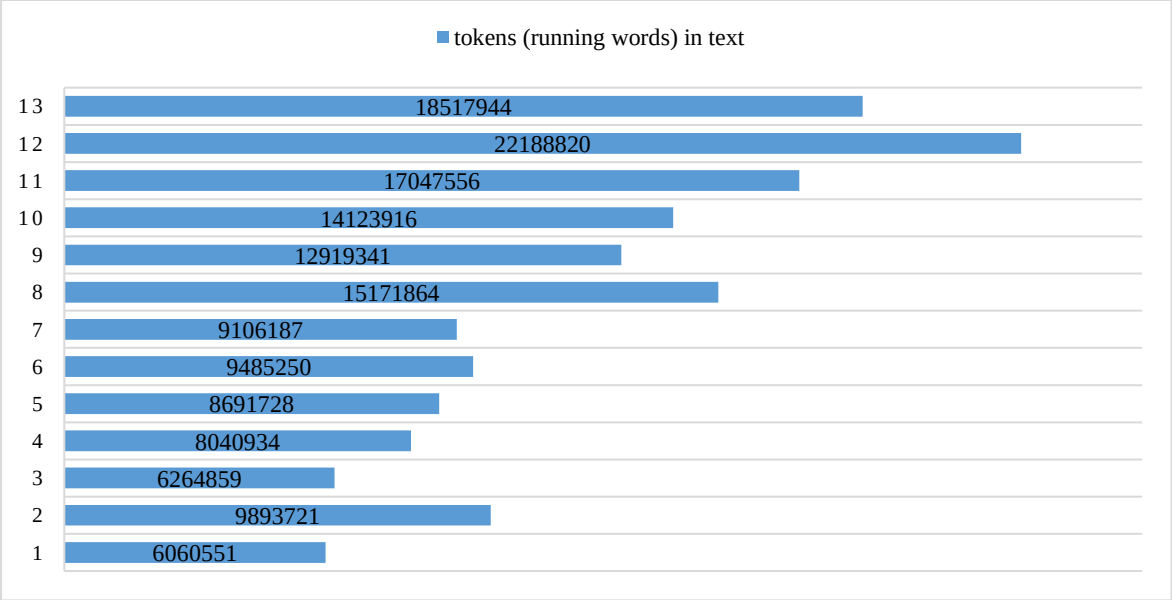


Figure 4: The visualisation of tokens according to Parliaments in MHC

Table 3 shows the span of each parliamentary proceeding. Parliament of Malaysia was augmented on 11th September 1959.

Table 3: Span of Time for each Parliamentary Proceeding

Parliament	**Date/ Year**
Parliament 1	11th September 1959 – 11th January 1964
Parliament 2	18th May 1964 – 14th February 1969
Parliament 3	20th February 1971 – 21st December 1973
Parliament 4	11th April 1974 – 7th April 1978
Parliament 5	31st July 1978 – 21st December 1981
Parliament 6	9th March 1982 – 4th August 1986
Parliament 7	6th October 1989 – 22nd June 1990
Parliament 8	12th March 1990 – 22 December 1994
Parliament 9	7th June 1995 – 10th November 1999
Parliament 10	20th December 1999 – 11th November 2003
Parliament 11	17th May 2004 – 19th December 2007
Parliament 12	28th April 2008 – 29th November 2012
Parliament 13	24th June 2013 – 5th April 2018
Parliament 14	16 July 2018 (Still proceeding)

4.2 According to Parliaments, sessions, meeting, and sittings (txt file)

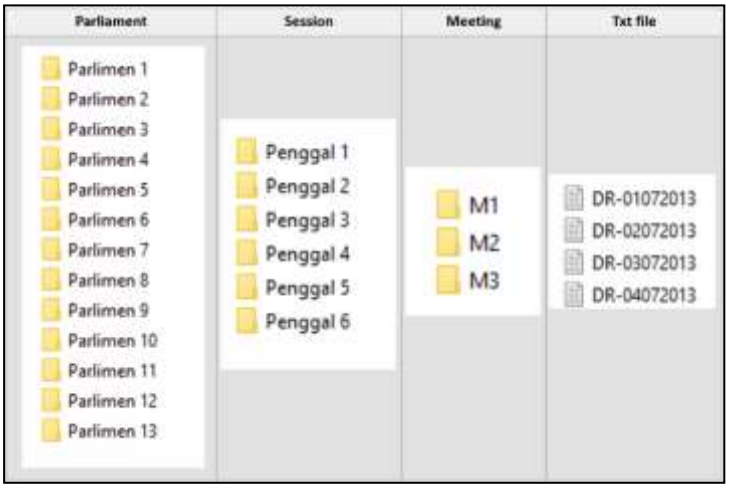


Figure 5: The manual storage of MHC according to Parliaments, sessions, and meeting

Figure 5 shows the manual storage of MHC according to Parliaments, sessions, and meeting. Each of individual txt file is stored in different subfolder according to sub-corpora. This storage would benefit anyone who would like to study the corpus on certain dates or according to different set of years.

4.3 Corpus Typology

A corpus typology presents the characteristics of the corpus. MHC is presented in plain text files (txt) formats.

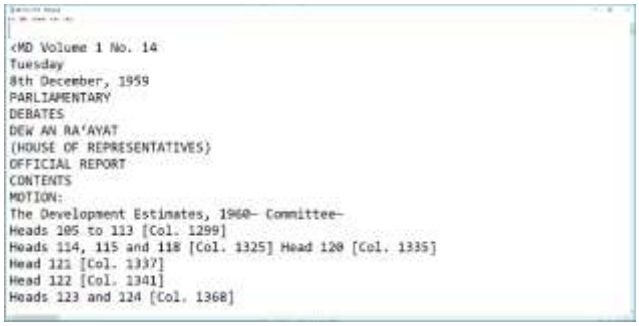


Figure 6: The front page of the plain text file
Source Parliament 1, Session 1, Meeting 1, Sitting 1

Figure 6 displays the front page of the corpus. Basically the front page displays the title, dates and attendance list. Each txt file has a simple mark-up that is done manually. The mark ups include <MK…/MK> for *Muka Depan* (Front Page) and <SK…/SK> for *Senarai Kehadiran* (Attendance list)

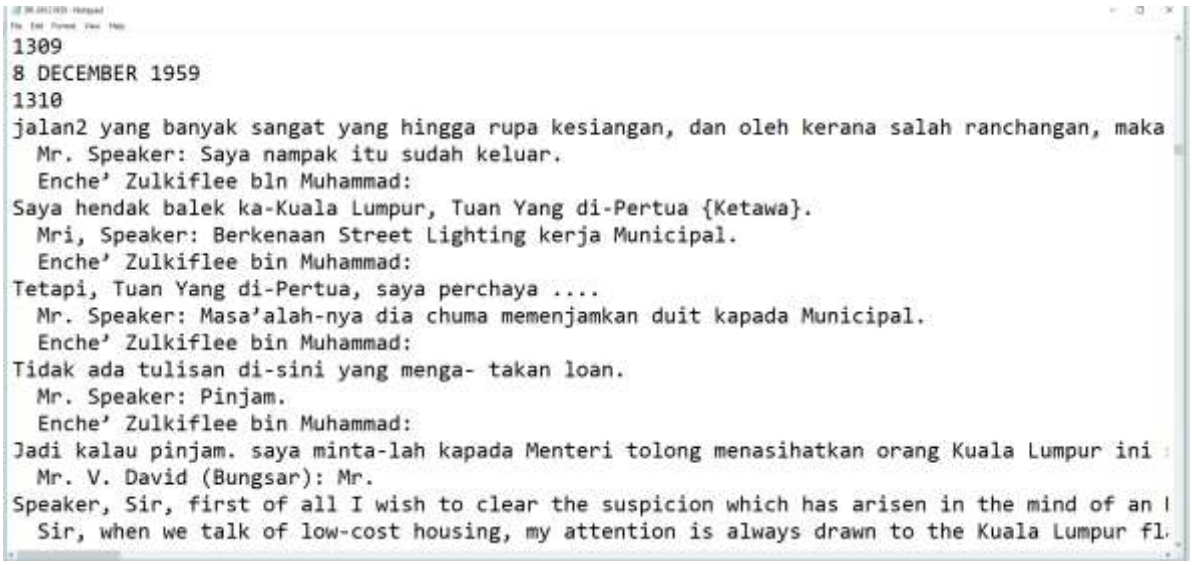


Figure 7: The middle page of the plain text file
Source Parliament 1, Session1, Meeting 1, Sitting 1

MHC has the following distinctiveness to distinguish it from other corpora. The characteristic of a corpus is important to distinguish the corpus for further analysis. The characteristic of this corpus is as follows:

(i) monolingual - includes only one language; MHC is in Malay language but in certains cases, there are is use of English language. i.e. Parliament 1 (fully English with some Malay debates)
(ii) dichronic
(iii) specialized (political discourse corpus)
(iv) monitor corpus
(v) raw format – no annotation.

5.0 Limitation of studies/ Issues during corpus development

There are several limitations in developing this corpus. Firstly, the original PDF data retrieved from the Official Portal of Parliamentary Debates Malaysia has different structures of reporting. The reporting style are also different from one parliament to another. Several issues have been addressed during the development of this diachronic corpus. The first issue includes the technical difficulty in converting old PDF files to txt format. The first PDF files have two columns which made it difficult to be converted. The other issues include (i) the use of dash in old Malay like *sa-bagai*, *hanya-lah ke-hari*, (ii) the use of old spelling convention of Malay language like *membahathkan. chontoh, chara*, (iii) the use of symbol like *ke'rana, undang*$_2$, the use of short forms like *Hon'ble* (for Honorable), and (iv) the use of diacritics in spelling like the word *mêminta* with acute *ê*. These technical aspects have affected the precision of conversion from PDF to txt files.

Converting a large volume of data requires time, and format of unconverted documents may affect the process of conversion. To solve the conversion issue, we have taken the measure to choose the most suitable Optimal Character Recognition (OCR) tool to convert PDF files into txt. There are various of OCR tools in the market. However, ABBYY FineReader was selected to convert the whole files from Malaysian Parliamentary Reports into txt due to its low global rate, zero rejection rate, and high in recognition and reliability rate compared to other OCR tools (Anis Nadiah et.al, 2019).

6.0 Conclusions and Future Work

We have proposed a replicable methodology for designing a diachronic corpus in Malay. This corpus will further be extended from time to time, following the assembly of the parliamentary debates in the Parliament of Malaysia. This designing framework is hoped to provide an outline for corpus development, specifically in Malay in the future. The corpus can be exploited for cross-multidisciplinary fields including corpus linguistic, sociolinguistics, discourse and political science research including Culturomics – the study of cultural trends and human behaviour through quantitative analysis of digitised texts.

For future work, the corpus will be housed online with the tagging of name of MPs, Political Party and the Parliament that the MP is representing like the British Hansard. Future work will also include tagging and annotation of the corpus. In the future, the corpus aims to complete its collection in two parts: The House of Representative (Dewan Rakyat) and Senate (Dewan Negara) from Parliament 1 (1959) to Parliament 14 (2019) in editable text file type (*txt*).

Retrieving The Corpus

This raw corpus (corpus with no annotation or mark-up) is presently available for retrieval at the Centre of Linguistics and Language Studies, the Faculty of Social Science and Humanities, Universiti Kebangsaan Malaysia (UKM). The corpus is available online at URL: https://drive.google.com/drive/folders/1lQ0sE9QOPbNngMIZxEUG836DYaIGtFk2?usp=sharing

Previous Works utilizing MHC

The availability of MHC has assisted in different disciplines. To date, there are a number of researches that have utilised the MHC for example Azianura and Marlyna (2017), Nor Fariza, Anis Nadiah, Azhar, Imran and Sabrina (2019), Norsimah, Azhar, Anis Nadiah and Imran (2019), Jamaluddin (2019a, 2019b) and Imran (2020).

Acknowledgements

We acknowledge the financial support from the National University of Malaysia (UKM) for the research grant AP-2014-018 and KRA-2018-005. Thank you to the friends who have also helped in the clean-up process and also the members of Digital Humanities Group UKM. We also thank the government of Malaysia through the Official Portal of Parliament of Malaysia for granting an open-access to the public to use the Hansard data.

References

Abercrombie, G., & Batista-Navarro, R. T. (2018). A Sentiment-labelled Corpus of Hansard Parliamentary Debate Speeches. In *ParlaCLARIN* http://lrec-conf.org/workshops/lrec2018/W2/pdf/book_of_proceedings.pdf

Alonso, J et al. 2009. Improving access to govern-ment through better use of the web. W3C Interest Group Note 12 May 2009 http://www.w3.org/ TR/egov-improving/, May.

Anis Nadiah Che Abdul Rahman, Imran Ho Abdullah, Intan Safinaz Zainudin & Azhar Jaludin. (2019). The Comparisons of OCR Tools: A Conversion Case in the Malaysian Hansard Corpus. *Malaysian Journal of Computing*.

Anthony, L. (2012). AntConc (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/ *(PDF)*

Azianura Hani Shaari, Marlyna Maros & Imran Ho Abdullah. (2017). Tracking Trends: The Corpus Analysis of Lexical Item Jenayah in the Malaysian Hansard. In Prosiding Kolokium Penerokaan Dataraya Kulturomik Malaysia 2017, pp 40 – 45.

Bachmann, I. (2011). Civil partnership – "gay marriage in all but name": a corpus-driven analysis of discourses of same-sex relationships in the UK Parliament. *Corpora*. 6(1), pp. 77-105

Bennet, D. & Harvey, A. (2009). Publishing open gov-ernment data (W3C Working Draft 8 September 2009). http://www.w3.org/TR/gov-data/.

Berracheche , B. (2020). Appraisal and Party Positioning in Parliamentary Debates: A Usage-

Based Critical Discourse Analysis. *International Journal of English Linguistics*. 10(6).
Berners-Lee, T. (2006). Linked data. http://www. w3.org/DesignIssues/LinkedData.html.
Domingos and M. Pazzani. 1996. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *International Conference on Ma-chine Learning*, pages 105–112.
Blätte, A., & Blessing, A. (2018). The GermaParl Corpus of Parliamentary Protocols. *LREC*.
Davies, Mark. (2015) *Hansard Corpus*. Part of the SAMUELS project. Available online at https://www.hansard-corpus.org/.
Hartig, O. (2009). Provenance Information in the Web of Data. In *Proc. of the Linked Data on the Web Workshop at WWW*.
Hiltunen, T., Räikkönen, J. & Tyrkkö, J. (2020). Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century. *Language Science*. 79
Imran Ho Abdullah, Zaharani Ahmad, Rusli Abdul Ghani, Nor Hashimah Jalaludin & Idris Aman. (2004). A Practical Grammar of Malay – a corpus based approach to the description of Malay: extending the possibilities for endless and lifelong language learning. *First COLLA Regional Workshop*, Putrajaya, Malaysia, 28-29 June, 2004.
Imran Ho Abdullah. (2020). Virus in Parliament: A corpus analysis of keywords related to public health in the Malaysian Hansard in *COVID-19 INSIGHTS AND LINGUISTIC METHODS*. *3L: The Southeast Asian Journal of English Language Studies*. 26(2). 1 – 23
Jamaluddin Aziz. (2019a). Exploring Gender Issues Associated With wanita/woman and perempuan/woman in Malaysian Parliamentary Debates: A Culturomic Approach. *GEMA Online® Journal of Language Studies*. 19(4). Pp 278-303
Jamaluddin Aziz. (2019b). Retrieving trends and issues of penyiaran/broadcasting in Malaysian Parliamentary Debates from 1957 to 2018: a culturomics approach. *Jurnal Komunikasi ; Malaysian Journal of Communication*, 35 (4). pp. 172-192.
Koehn. P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. MT summit*, volume 5.
Lee, L.W. & Low, H.M. (2011). Developing an online Malay Language Word Corpus for primary schools. *International Journal of Education and Development using Information and Communication Technology (IJEDICT)*, 7(3). pp. 96-101.
Maciej, O. (2018). *Polish Parliamentary Corpus*, CLARIN-PL digital repository, http://hdl.handle.net/11321/467.
Marx, M. & Schuth, A. (2010) Maarten Marx and Anne Schuth. DutchParl: The Parliamentary Documents in Dutch.
Norsimah, Mat Awal, Azhar, Jaludin, Anis Nadiah Che Abdul Rahman. & Imran, Ho Abdullah. (2019). "Is Selangor in Deep Water?": A Corpus-driven Account of Air/water in the Malaysian Hansard Corpus (MHC). *GEMA Online® Journal of Language Studies*. 19(2). Pp 99-120
Nor Fariza Mohd Nor, Anis Nadiah Che Abdul Rahman, Azhar Jaludin, Imran Ho Abdullah & Sabrina Tiun. (2019). A Corpus Driven Analysis of Representations Around the Word 'ekonomi' in Malaysian Hansard Corpus. *GEMA Online® Journal of Language Studies*. 19(4). Pp 66-95
Offersgaard, L (2018). A Pilot Gender Study of the Danish Parliament Corpus. ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora, Miyazaki 2018
Philipp, K. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit 2005
Rheault, L., Beelen, K., Cochrane, C. & Hirst, G (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLoS ONE 11(12): e0168843. https://doi.org/10.1371/journal.pone.0168843*
Roukos, S., Graff, D. & Melamed, D. (1995). Hansard French/English LDC95T20. CD. Philadelphia: Linguistic Data Consortium.
Scott, M., (2008). WordSmith Tools version 5, Liverpool: Lexical Analysis Software.
Scott, M. (2012). WordSmith Tools (Version 5.0) [Computer Software]. Available from http://www.lexically.net/software/index.htm
Steinþór, S., Helgadóttir, S., Rögnvaldsson, E. Barkarson, S. & Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. *Proceedings of LREC 2018*, pp. 4361-4366. Myazaki, Japan.
Suhaimi abdul Rahman & Normaziah Abd Aziz. (2004). Improving word alignment in an English–Malay parallel corpus for machine translation. *Pre-LREC 2004 Workhop on Amazing Utility of Parallel and Comparable Corpora*
Tengku Sepora Tengku Mahadi., Vaezian, H., Akbari, M., Nor Aini Ali. & Cheng, C.S. (2009). Building a Legal TM and Glossary from an English-Malay Parallel Corpus. https://www.translationdirectory.com/articles/article2069.php
The Official Portal of Parliament Malaysia. Hansard. 2017. Retrieved from http://www.parlimen.gov.my/hansard-dewan-rakyat.html?uweb=dr
Trotman, A. & Sigurbjornsson,B. (2005). Narrowed Ex-tended XPath I (NEXI). In *Advances*

in XML Informa-tion Retrieval, pages 16–40.
W3C/IETF URI Planning Interest Group. (2001). URIs, URLs, and URNs: Clarifications and Recommendations 1.0. W3C Note 21 September 2001. http://www. w3.org/TR/uri-clarification.
Wynne.M. (2005). Archiving, distribution and preservation. In M. Wynne, editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 71–78. Oxford: Oxbow Books. Available online from http://ahds.ac. uk/linguistic-corpora
Zuraidah Mohd Don. (2010). Processing Natural Malay Texts: A Data-Driven Approach. TRAMES. 14(64/59), pp 90–103

About the Authors

Imran Ho Abdullah (PhD) is the Professor of Linguistics in the Faculty of Social Science and Humanities, Universiti Kebangsaan Malaysia. His area of specialty is in cognitive linguistics, corpus linguistics and semantics. He is presently the Deputy Vice Chancellor in Universiti Kebangsaan Malaysia.

Anis Nadiah Che Abdul Rahman is the PhD candidate is Centre of Linguistics and Languages Studies, Faculty of Social Science and Humanities, Universiti Kebangsaan Malaysia. Her area of interest is corpus linguistics and semantics.

Azhar Jaludin is a lecturer in the Faculty of Social Science and Humanities, Universiti Kebangsaan Malaysia. He specialises in web crawling and corpus linguistics.