Domain-specific Stop Words in Malaysian Parliamentary Debates 1959 – 2018

Anis Nadiah Che Abdul Rahman [a]
p87706@siswa.ukm.edu.my
Centre for Research in Language and Linguistics,
Universiti Kebangsaan Malaysia, Malaysia

Imran Ho Abdullah
imranho@ukm.edu.my
Centre for Research in Language and Linguistics,
Universiti Kebangsaan Malaysia, Malaysia

Intan Safinaz Zainudin
intansz@ukm.edu.my
Centre for Research in Language and Linguistics,
Universiti Kebangsaan Malaysia, Malaysia

Sabrina Tiun
sabrinatiun@ukm.edu.my
Centre for Artificial Intelligence Technology,
Universiti Kebangsaan Malaysia, Malaysia

Azhar Jaludin
azharj@ukm.edu.my
Centre for Research in Language and Linguistics,
Universiti Kebangsaan Malaysia, Malaysia

ABSTRACT

Removal of stop words is essential in Natural Language Processing and text-related analysis. Existing works on Malay stop words are based on standard Malay and Quranic/Arabic translations into Malay. Thus, there is a lack of domain-specific stop word list, making it discordant for processing of Malay parliamentary discourse. In this paper, we propose a semantic approach towards identifying and removing Malay, conventional Malay spelling and English functional words in analysing a time-series corpus, namely the Malaysian Hansard Corpus (MHC), to extract a Malay specific-domain stop word list. The study utilised a combination of Z-method of most frequently occurring words, words that appear once, and the classic method. The dataset of the corpus evaluated comprised Parliament 1 (year 1959) to Parliament 13 (year 2018). The study then categorised the stop word list according to domain-specific related words. The resulting list comprised 587 stop words. New stop words that emerged from the MHC include parliamentary-related words like '*Berhormat'* (salutation to the members of the Parliament), '*Pertua'* (salutation to the Speaker of the House), '*ketawa'* (laugh) and '*tepuk'* (clap). Other than typical English stop words like 'and' and 'the', there are also words like 'hon'ble' (short for 'Honourable') and 'honourable'. The list also includes stop words in conventional Malay spelling like '*untok*' (for), '*lebeh*' (more), and '*kapada*' (to). The proposed set of stop words can be further utilised to assist natural language processing and text analysis.

Keywords: stop word removal; text filtration; Malaysian Hansard Corpus; Malay stop word; parliamentary corpus processing

[a] *Main & Corresponding author*

eISSN: 2550-2131
ISSN: 1675-8021

INTRODUCTION

Removal of stop words in Machine Learning's pre-processing is becoming increasingly exploited in Natural Language Processing (henceforth NLP) and text analysis. Primarily discovered by H.P. Luhn in 1959, the term 'stop word' is often used in computer science for NLP. According to Raulji and Saini (2017), stop words are a set of words, which normally have little semantic value, that are filtered out or excluded from a text for text processing. Generally, stop words can be identified from a collection of most frequently occurring words in a corpus. Stop words are commonly removed from a text/data set as a part of NLP or to train deep learning and to create models for machine learning. They are removed because they occur frequently in a corpus or data. According to Makrehchi and Kamel (2017), the attributes of stop words can be distinguished by looking at low discerning values, insignificant information contents, high occurrence in frequency, association with the majority of categories (in the case of labelled corpus) and association with the majority of words in the wordlist. Removal of stop words (also known as stop list in corpus linguistics field) is a common practice in Information Retrieval (IR), NLP and even corpus linguistics. Previous studies including Chong, Banchs and Chng (2012) and Sabrina, Saidah, Nor Fariza, Azhar and Anis Nadiah (2020) have employed filtration of most frequent words or stop words out of a processed document/data.

In linguistics, a stop word could also be identified as a function word. Linguistically, function words denote the words that have lesser lexical meaning or have ambiguous meaning. Function words are also regarded as non-lexical categories that mostly function as grammatical items rather than having clear semantic content. Examples of function words in English include grammatical items like determiners, pronouns, preposition, modals, auxiliary verbs, question words and qualifiers. Contrary to function words, content words are the words that have distinct meaning. Content words include grammar class like verbs, nouns, adjectives, and adverbs. In the case of stop word, function words are indicated as stop words while content words are excluded from the process of acquiring stop word list as they possess semantic meaning or value to be analysed in the field of NLP or corpus linguistics.

The literature on stop word removal has highlighted several advantages. A study by Raulji and Saini (2016) attested that by removing stop words prior to processing documents ultimately improved system performances. Similarly, Munková, Munk and Vozár (2014) found that filtering out noise or stop words from a set of data improved the feature and value of the data. Additionally, Kaur and Buttar (2018) indicated that the removal of stop words was able to reduce vector space in NLP and improved the performance by increasing the speed of the performance and calculation as well as accurateness of the result.

The need for stop word lists to assist Malay NLP has also been indicated in prior studies. Mohd Amin, Aida and Noor Azah (2017) and Chua and Nohudin (2017) utilised the use of Malay stop words to process data on Malay translation of Quranic verses. Keshavarz and Abadeh (2017) used stop words to process sentiment analysis in Malay Reviews Corpus (MRC). Previous research on Malay stop words have also leaned towards general Malay and translations of Quranic verses into Malay to process stop words, for instance in studies by Muhammad Taufik, Fatimah, Ramlan and Tengku Mohd (2005), Kwee, Tsai and Tang (2009), and Chekima and Alfred (2016). In this regard, existing lists of stop words are therefore discordant for processing of Malay parliamentary discourse.

Previous researchers including Rose, Engel, Vramer and Cowley (2010) and Choy (2012) have proposed diverse techniques to extract stop word lists for specific corpus; however, the lack of research in standardised stop words has led to the application of pre-existing stop word lists among researchers (Choy, 2012). The same problem occurs in Malay NLP because of the lack of complete stop word lists in the Malay language and this has obstructed research

in Malay NLP. Thus far, previous studies have indicated the lack of standardised Malay stop words.

Similarly, the literature on stop words has also revealed the lack of standardised stop word lists in assisting NLP. Kaur and Buttar (2018), for example, stated that a small amount of work has been conducted in languages other than English in creating stop word lists. However, research on stop words in other languages is very limited compared to the English language. In the case of Malay stop words, Chekima and Alfred (2019) remarked that many researchers have either translated stop words from the English list or manually collated them as Malay stop words because of the inadequacy of standardised Malay stop words as compared to English. This is supported by Hamood Ali, Sabrina and Nazlia (2017) who explained that research in Malay NLP has been mostly constrained as a result of inadequate resources in managing Malay Text Classification (henceforth MTC).

A specific corpus also requires a specific set of stop words; however, the lack of suitable stop words has resulted in interruption in data analysis. To date, several studies have investigated the use of stop words in domain-specific corpora. Zheng (2018) in his study on general stop word list discovered that a general stop word list is unable to precisely fit the requirements of specific corpus. This finding is supported by Hassan, Fernández, He and Harith (2014) who found that a pre-compiled list of stop words negatively affects the performance of domain-specific corpus analysis because the available stop word list is irrelevant to the domain-specific corpus like Twitter. Moreover, the application of domain-specific stop word lists in data mining tasks is able to accelerate search time and enable a more efficient finding; thus, in this regard, suitable stop word lists are required because of the benefits they could offer. According to Alshanik, Apon, Herzog, Safro and Sybrandt (2020), the removal of domain-specific stop words in a corpus minimizes the proportion of space used and enhances retrieval performance of data in text mining. Domain-specific wordlist is different from open domain wordlist as it contains words that have high frequency but possess low value in the domain-specific corpus. In the case of Malaysian parliament for example, words like '*Enche' (*Sir), '*Tuan'* (Sir) and '*Yang Berhormat'* (salutation/ Honorary) have low information on the content and are specific to the discourse of Malaysian parliamentary reports (Hansard) only. In addition, domain-specific stop words also vary from one domain or discourse to another. As an example, word like '*tepuk'* (clap) can be a stop word in the discourse of parliamentary reports but may act as a keyword in other domains like learner corpus. Therefore, the lack of suitable stop words has resulted in interruption when analysing data from the Malaysian Hansard Corpus (henceforth MHC). Accordingly, this study was conducted to address this inadequacy and to cater to the needs of researchers in processing MHC in a more straightforward manner.

As words can evolve, Makrenchi and Kamel (2017) argued that standard stop words can therefore become outdated over time. English stop words were initially published in the 1970s and over a period of time, new words had to be added into the list. A similar scenario applies in the case of Malay stop word lists as Malay words have also changed. To date, there is no stop word list made available to the public on parliamentary discourse in the Malay language, specifically the MHC. MHC therefore has the advantage in terms of producing stop words for parliamentary or political discourse. Interestingly, the MHC is a temporal corpus with data that spans a period of over 60 years. As a result of its distinctive temporal organisation according to year from 1959 to 2018 (up to the point of doing this research) compared to other sources of data utilised in previous studies, it makes it possible for stop words over different periods of time to be classified. In addition, MHC also contains English words and words from old Malay as compilation of data began in 1959. In the case of MHC, the spelling of certain words is also different. In Parliament 1 for example, the term '*Enche'* was used to denote Sir instead of the new spelling '*Encik'*. Other words include '*sa'* (conventional spelling for '*se'*, meaning *one*), '*macham'* (conventional spelling for '*macam'*, meaning *like* or *as*) and '*niscaya'*

(conventional spelling for '*nescaya'*, meaning *by all odds*). Therefore, the existence of such stop words would enable researchers to eliminate low content words from conventional and new Malay spelling.

This study aims to contribute to this growing area of research by providing a list of standardised stop words specifically for Malay parliamentary discourse for use with MHC but not restricted to MHC alone. Even though the data used is from parliament, it is also relevant to general use of the Malay language. In addition, many previous studies have utilised the MHC (Nor Fariza, Anis Nadiah, Azhar, Imran & Sabrina, 2019; Norsimah, Azhar, Anis Nadiah & Imran, 2019; Sabrina, Nor Fariza, Azhar & Anis Nadiah, 2020; Sabrina, Saidah et al., 2020). It is expected that the production of Malay stop words relating to the corpus will assist future research in terms of stop word removal and text processing in general. Therefore, the study aims to determine a set of stop word list from a domain-specific parliamentary corpus and propose an approach to determining domain-specific Malay stop word lists. The findings should make an important contribution to the field of NLP by providing a set of standardised Malay stop words that can be used to assist NLP as well as a model to extract Malay stop words for domain-specific Malay corpus.

LITERATURE REVIEW

THE CONSTRUCTION OF STOP WORD LISTS IN ENGLISH AND OTHER LANGUAGES

Studies on stop word construction have been carried out extensively across the globe and the literature on stop word construction has continued to grow. The existing literature on stop word construction is extensive where the focus has been on stop words in English as well as other languages such as Arabic, Chinese, German, Persian, Polish, Punjabi and even Sanskrit. Puri, Bedi and Goyal (2013), for instance, constructed a list of Punjabi stop words by searching the most frequent words in 10,000 news articles from the Punjabi newspaper, *Ajit*, using a statistical method. The study also proposed traits for a word to become a stop word by taking an average of 400 words per article. Khan, Bakht, Khan, Samad and Sahar (2019) extracted Urdu stop words by producing 500 highest frequency words in which a total of 358 were classified as stop words.

There are a number of stop word lists available on the internet which were extracted using different methods. According to Kaur and Buttar (2018), many of these stop words have been utilised as standard stop words in multiple research works. Lists of stop words are also available in programming tools. Wild, Kalz, Demnati, Paliwoda-Pekosz and Naili (2020), for example, developed stop word lists for *R* in English, Dutch, French, Polish and Arabic, respectively. Malay stop word lists are also available. These lists can be found in websites such as https://nlp.cs.nyu.edu/GMA_files/resources/ME.tralex and https://nlp.cs.nyu.edu/GMA_files/resources/malay.stoplist.

MALAY STOP WORDS

Several Malay stop word lists which have been proposed by previous researchers are currently available. Muhammad Taufik et al. (2005), for instance, extracted a total of 305 Malay stop words from the Malay Quranic text collection. Subsequently, by ranking the extracted words, a list of 50 stop words that occurred most frequently in the corpus was developed. Kwee et al. (2009) produced a list of 339 Malay stop words. Fatimah et al. (2011) developed the Malay Interrogative Knowledge Corpus (MalayIK-Corpus) and identified a list of stop words based on frequency of occurrence and ranking. Out of 6,479 words extracted, 35 words were selected as the most frequently occurring words in the corpus. Meanwhile Chekima and Alfred (2016)

utilised three approaches to produce 399 Malay stop words from *Dewan Bahasa dan Pustaka*'s corpus of seven million tokens. The first approach was by considering word frequencies according to rank which was orchestrated by Zipf's law. The second approach was through the use of word distribution of variance measure. The third approach used computation of Entropy measure.

Generally, Malay stop word lists consist of auxiliary verbs, adverbs, conjunctions, determinants, negatives, predicates, prepositions, pronouns, and relatives. This is based on earlier works on Malay stop word lists by Fatimah (1995), Muhammad Taufik et al. (2005) and Muhamad Taufik (2006). In a study which aimed to provide resources for text processing in the Malay language, Baldwin and Su'ad (2006) stated that Malay stop words should also consist of abbreviations of names and places, and acronyms of salutations such as *Abd., Mohd., Bhd., Inc., St., Jln., Kapt., kg., kump, LL.B., Lt., per., Pn., Pt. Tn., Tj., Y.bhg*. In addition, the same study also suggested that Malay stop words should also consider morphological aspects including affixation like prefixes (i.e., include *me-, pe-, be-, ter-, se-),* infixes of *-el-, -em- and -er-*, and suffixes like *-i, -kan, -nya,* and *-lah.*

DOMAIN-SPECIFIC STOP WORDS

A domain-specific wordlist is a wordlist that represents and is extracted from a specialised corpus of a particular discourse. Domain specific or specialised corpora are generally the corpora that are constructed for a specific purpose or have different language use compared to general corpora which represent general language use. Examples of domain specific corpora include any parliamentary corpora, corpora by Weisser (2013) like Air Traffic Control (ATC) Corpus, Business Letters Corpus, the Speech Act Annotated Dialogues (SPAADIA), and corpora by Koteyko (2014) like the corpus of presidential speeches (CPS) and the Russian press corpus (RPC). According to Liu et al. (2016), a domain-specific wordlist relies on a domain key phrase or the words or phrases that are statistically significant in a specialised corpus. Similar to general stop word lists, domain-specific wordlists should be natural, eloquent and have explicit semantic units to represent word use of a particular domain-specific or specialised corpus.

The academic literature on domain-specific stop word list construction has revealed the emergence of various stop word lists with the adaptation of different approaches. Ayral and Yavuz (2011), for example, proposed an automated approach to generate domain specific stop words in English to enhance classification of natural language content. The study also tested the stop words using Bayesian natural language classifier. The study used Zipf's law to prove that document topic coverage rank of words followed the traits of a natural language corpus. Makrehchi and Kamel (2017) proposed on automatic generation of domain-specific stop words from a large-labelled corpus.

In relation to parliamentary corpus analysis, Hofmann, Marakasova, Baumann, Neidhardt and Wissik (2020) listed criteria for parliamentary documents' stop words that include word possessing traits of numeral, name of dates, days and months and the officials' titles like 'councillor' and 'president' which were regarded to contain unimportant information. In another related study, Greene and Cross (2017) removed stop words associated with parliamentary discourse such as 'adjourn' and 'comment' as well as the name of politicians.

Meanwhile, Rani and Lobiyal (2018) in their study on Hindi language constructed a domain specific Hindi stop words by using statistical and knowledge-based method. The study proposed a new method called netting ranked performance evaluation (NRPE) to evaluate the validity of stop word lists. By using this approach, the removal of stop words is conducted by sorting the wordlist alphabetically and statistically according to frequency. By using combined

band net (CBN) performance, the researchers validated the stop word removal process. In this study, punctuation marks like ",", "!", ";", "|" were removed.

Most studies in relation to domain-specific stop words have only focused on stop word lists in other languages than Malay. Lack of available stop word list on parliamentary discourse and the Malay language has delayed Malay NLP and caused many researchers to use existing lists that might have been outdated. In addition, many researchers have also translated English stop words into Malay. However, this step has resulted in the impediment of Malay NLP. Therefore, the present study is important as it would enhance Malay NLP and analyses in studies that utilise textual data.

APPROACHES TO EXTRACTING AND REMOVING STOP WORDS

The literature on stop word list extraction and removal has highlighted several frequently applied techniques which include the Classic Method, Zipf's Law or Z-Method, the Mutual Information Method (MI) and Term Random Sampling (TBRS) (Kaur & Buttar, 2018). The Classic Method involves the removal of stop words obtained from pre-existing stop word lists. This method has been employed by numerous researchers because of its convenience and availability to assist NLP. The second method in obtaining stop word lists is by using Zipf's Law, a method proposed by Zipf in 1949. This method highlights the selection of stop words based on three additional criteria along with the classic method. The additional criteria include the selection of the most frequent words (TF-H), the words that appear once or the hapax legomanon (TF-1), and words that have inverse document frequency (IDF). IDF is also defined as a measure to describe rarity or commonness of a word in a corpus. IDF can be acquired by dividing the total amount of file or document with the number of documents that have the studied term in the corpus. Previous research that utilised this method includes Makrehchi and Kamel (2008) for a labelled corpus, and Khan et al. (2019) who extracted 358 of the highest frequently occurring words in Urdu language by using the TF-IDF method.

The third method or the MI is a method that exploits the use of computed mutual information. Low MI score indicates that the lexis is insignificant as it has low discrimination power, thus suggesting the removal of that particular lexis. Previous studies that employed MI method include Zhi (2003) who extracted stop words in Chinese and Yuan, Lo and Lawall (2014). TBRS was initially introduced by Lo, He and Ounis (2005). In this method, stop words are automatically distinguished from the documents available on the web. As its name suggests, the method utilises random selection of separate chunks of data. Subsequently, words in each chunk are ranked based on their informativeness or clarity. Kullback-Leibler's divergence measure is applied in this method to measure the clarity of the word based on the calculation (dx(t) = Px(t).log2 Px(t) P(t)). Based on this method, the final stop word list is developed by manipulating the least informative term with removal of duplications. This method has been employed by researchers like Hassan et al. (2014).

METHODOLOGY

DATA: THE MALAYSIAN HANSARD CORPUS

The MHC was initially developed by Imran, Anis Nadiah and Azhar (2017). The corpus consists of parliamentary debates of the *Dewan Rakyat* (House of Representatives) in the Malaysian parliament, ranging from Parliament 1 in 1959 to Parliament 14 in 2020. To date, there are approximately 165,061,050 word-tokens in the corpus with 1,049,053 distinctive words (type). Table 1 shows the size of each sub-corpora in the MHC.

TABLE 1. The size (token of running words) of sub-corpora (parliament) in the MHC (Imran et al., 2017)

Parliament	Tokens (running words) in text
Parliament 1	6060551
Parliament 2	9893721
Parliament 3	6264859
Parliament 4	8040934
Parliament 5	8691728
Parliament 6	9485250
Parliament 7	9106187
Parliament 8	15171864
Parliament 9	12919341
Parliament 10	14123916
Parliament 11	17047556
Parliament 12	22188820
Parliament 13	18517944
Parliament 14	7548379

The language in the MHC has several attributes. According to Imran et al. (2017), the MHC contains verbatim report of Malaysian parliamentary debates from 1959 to 2018, and thus, there are differences in language use over time. The first parliament (1959-1964) used English as its official language. However, with the permission from the Speaker, Members of Parliament (henceforth, MPs) were allowed to use Malay. The Malay language during this era still used the conventional spelling which is different from the current spelling. There are spellings like '*membahathkan'* (conventional spelling for *'membahaskan'* which means 'debating'), *chontoh* (conventional spelling for *'contoh'* which means 'example'), and *'chara'* (conventional spelling for '*cara*' which means 'way'). There are also short forms like *Hon'ble* (for Honourable).

STATISTICAL METHODS TO ACQUIRE STOP WORD LISTS

This study adapted and adopted the Z-method proposed by Zipf (1949), Hassan et al. (2014), and Hofmann et al. (2020). Based on Zipf's law, the most frequent words or Term's Frequency-High (TF-H) and the words that appear only once or with frequency of one (TF-1) were removed. The study also utilised the list of Malay stop words by Muhammad Taufik et al. (2005) as a control set of stop word list to obtain stop words from the MHC. The application of pre-existing list of stop words is also called the Classic Method. In addition to that, this study employs the framework set by Hoffman et al. (2020) and Greene and Cross (2017) where the words that indicate numbers, dates, days, months, the titles or salutation officials and Members of Parliament), parliamentary specific words like 'adjourn' and 'comments' are removed. The procedural framework can be seen in Figure 1.

eISSN: 2550-2131
ISSN: 1675-8021

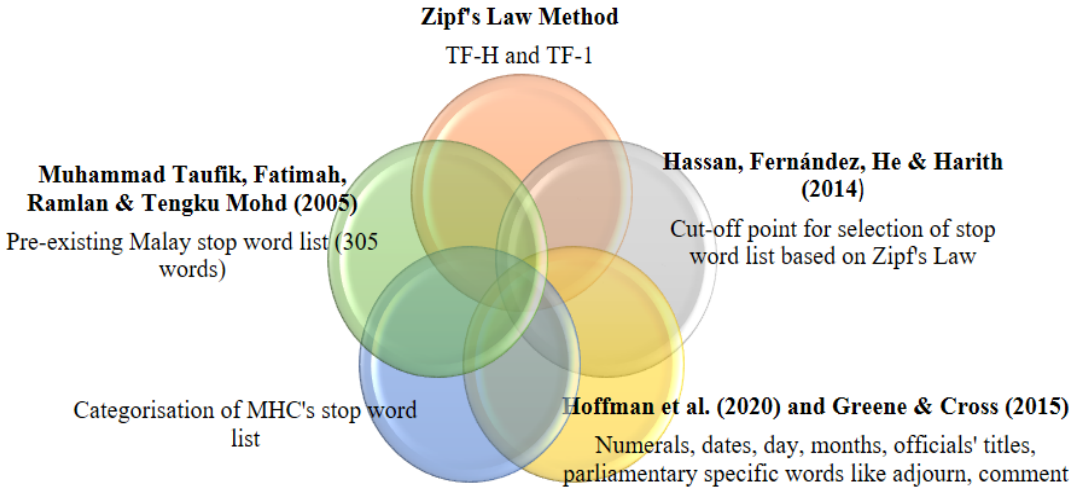

FIGURE 1. Procedural framework to obtain stop word list in MHC

Based on the procedural framework, a set of procedures was constructed to extract a stop word list from each sub-corpus from Parliament 1 to Parliament 13. These procedures are documented in Figure 2.

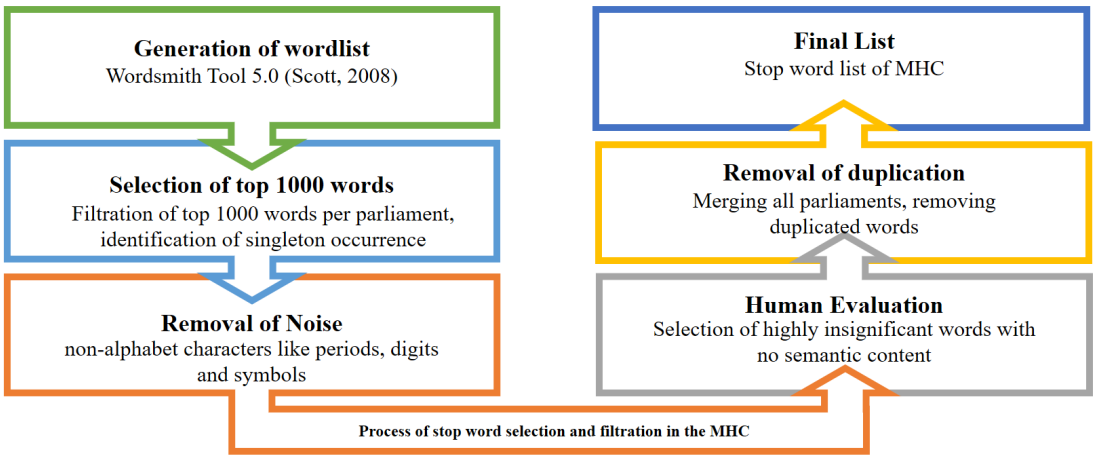

FIGURE 2. The process of stop word selection and filtration in MHC

The process began with the generation of wordlist of each parliament from Parliament 1 to Parliament 13 using WordSmith Tool 5.0 (henceforth, WST5). As a result, 13 different wordlists were generated. Two parliaments were selected based on their distinctive characteristics. Parliament 1 was chosen because of its nature where it has a different spelling than the contemporary Malay spelling; it used conventional Malay spelling and the official language medium used in Parliament 1 was English (Imran et al., 2017). Parliament 13 was selected as it is the most recent parliament with complete sessions. Even though the most current parliament in Malaysia is Parliament 14, it is still ongoing and is yet to complete. For this reason, the number of tokens in Parliament 14 is smaller (refer to Table 1) compared to the number of tokens in less recent parliaments. Therefore, by comparing Parliament 1 and 13, the shift in language use and spelling can be significantly identified. Through this methodology, distinctive stop word lists would be produced based on each parliament. Additionally, removal of noise was also conducted. Removal of noise included non-alphabet characters such as

period, digits and symbols like "#". Unnecessary words with semantic content were removed from the list too. The list was then merged, and duplicated words were removed to produce the final list.

A statistical analysis was then performed using WST5 (Scott, 2008) and Microsoft Excel. Each wordlist was saved in Microsoft Excel worksheet and filtration of top 1,000 words with highest frequency (TF-H) was executed. Words with singleton occurrence (TF-1) were also identified. After that, noise was removed manually and semi manually. Manual removal of noise included the identification of noise in two parliaments (particularly Parliaments 1 and 13).

Following the statistical analysis, human evaluation was carried out in order to select insignificant words that have low value or no semantic content. The list of TF-H and TF-1 were inspected. In addition, a comparison of wordlists was also performed to see the occurrence of selected domain-specific wordlist compared to representative use of Malay language. This approach was taken to compare the use of political discourse stop word compared to daily language use and to justify the selection of the words as parliamentary domain-specific stop words. In order to fairly compare between two corpora of different sizes, a normalisation procedure was carried out on the frequency of occurrences of the words in each corpus. The normalisation process can be easily calculated using Equation 1:

$$\text{Normalised frequency } (fn) = \frac{frequency\ of\ studied\ word}{number\ of\ token\ in\ the\ studied\ corpus} \times 1{,}000{,}000$$

EQUATION 1. Formula of normalised frequency

In order to compare stop word list from this study, a comparison was made using the controlled set of stop word list by Taufik et al. (2005) and Malay representative corpus – a corpus that represents general use of Malay language. For this study, Malay Practical Grammar Corpus (henceforth, MPGC) or also known as DBP-UKM corpus by Imran, Zaharani, Rusdi, Nor Hashimah and Idris (2004) was used to see in-context application between stop words in MHC and general Malay language corpus.

PRELIMINARY ANALYSIS TO TEST SELECTED METHODS

As a preliminary test of the method, a wordlist from Parliament 13 was generated using WST5. Based on the wordlist, there were 92,878 word-types in Parliament 13. A preliminary analysis using the Zipf's law result indicated that generating a visualisation of all frequencies was impractical because the cut-off point was way too big as a result of the big data. Thus, the researchers decided to determine the cut-off point to a specific number of stop words based on the requirement of Zipf's law (see Figure 3). For this reason, the parameter was lowered to top 1,000 words.

TF-1 of stop word list was taken from the gradual cut-off. However, it was also noted that TF-1 or Hapax legomanan (words that appear only once in a corpus) had to be manually inspected and valued because of the existence of English words in the parliament. The frequency of English words in this parliament was relatively low because the rule of the House of Representatives states that the official language used in the parliamentary debate is the Malay language. However, the MPs are allowed to use English words with the permission of the Speaker of the House. English words as TF-1 do not necessarily occur as a stop word. Words like 'Whitehouse', 'vote', and 'colonisation' have semantic meanings that may be meaningful to future studies of NLP.

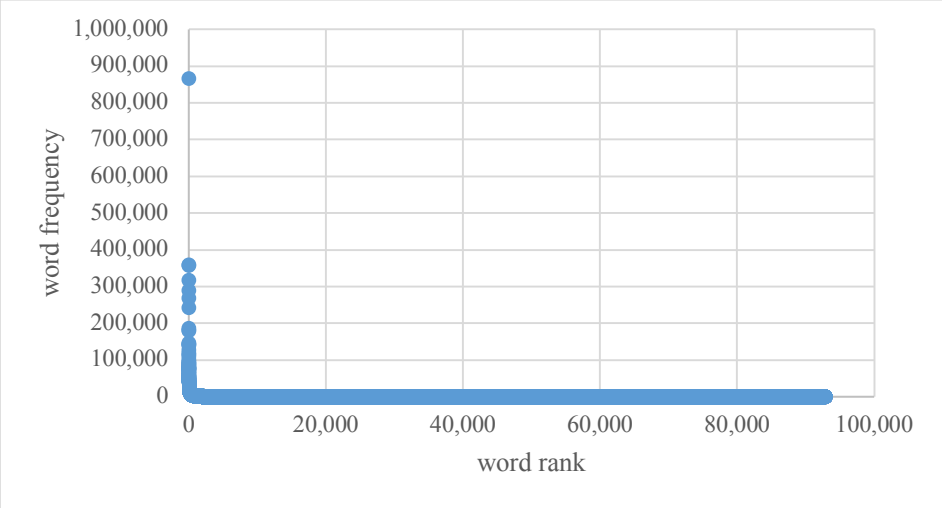

FIGURE 3. Rank-frequency distribution according to Zipf's law in MHC's Parliament 13 (sorted to top 100,000 words)

Figure 3 indicates the distribution between rank and frequency of the top 100,000 terms in MHC, Parliament 13 according to Zipf's law. The words were removed from the figure to improve clarity of visualisation on rank-frequency distribution according to Zipf's law. The Y axis represents the frequency of words while the X axis represents the ranking of word occurrences. As can be seen, the majority of words have the frequency of 400,000 and below (stopped at the word '*ini'*, meaning *this*, with frequency of 309,559). Interestingly, there is one dot plotted on the upper side of the chart, represented by the word '*yang*' (which could mean *that*, *which* or *as*) with the frequency of 865, 993.

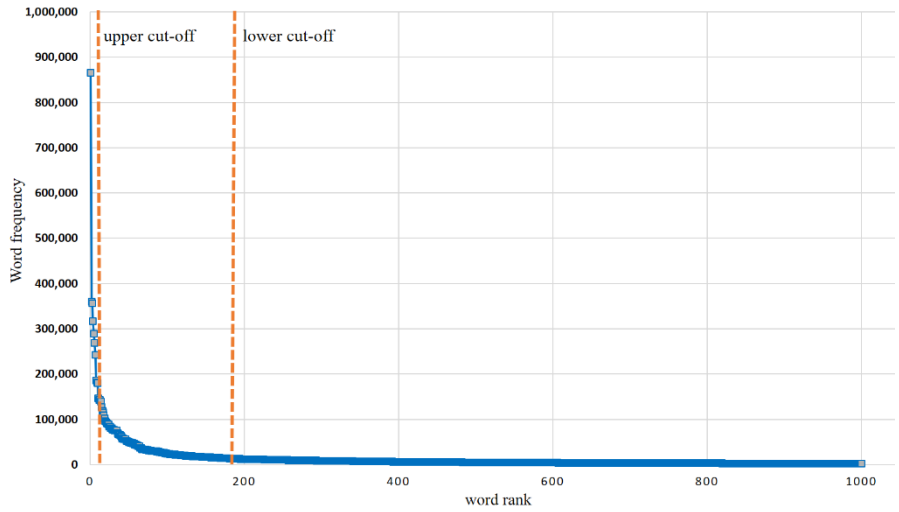

FIGURE 4. Rank-frequency distribution according to Zipf's law in Parliament 13 of the MHC (sorted to top 1,000 words)

Figure 4 shows the distribution between rank and frequency of the top 1,000 terms in MHC, Parliament 13. The words were removed from the figure to improve clarity of visualisation on rank-frequency distribution according to Zipf's law. Similar to Figure 3, the Y axis represents the frequency of words while the X axis represents the ranking of word occurrences. As seen in Figure 4, the cut-off point to determine the stop word list is at the elbow of the curve with the dotted lines. According to Hassan et al. (2014), TF-H consists of words from the beginning to the upper cut-off point while the remaining after the lower cut-off

point would be the additional stop words that fall under TF-1. The same approach was later used with Parliament 1 to determine TF-H and TF-1.

RESULTS

EXTRACTION OF TF-H AND TF-1

This section presents the result of extracted TF-H (the most frequent words) and TF-1 (words that appear once or the hapax legomanon (TF-1) in Malay and English in Parliaments 13 and 1. Overall, there were 12,747 different words in Parliament 13.

Table 2 shows the top 20 words according to frequency in Parliament 13 – TF-H and the bottom 20 words according to frequency (after removal of numerals) in Parliament 13 – TF-1. Words highlighted in yellow in TF-1 indicated that the words had semantic value (meaning); thus, they were disregarded from the stop word list's compilation. Based on Table 1, it can be seen that common Malay words like '*yang'* (English translation: *that, which, as*, frequency: 865993) and '*ini'* (English translation: *this*, frequency: 359559) were the most frequent words that occurred in the MHC with percentage of occurrences at 4.68 and 1.94%, respectively. Symbol like "#" (octothorp) also occurred frequently (575191) out of 18,517,944 words in Parliament 13, contributing 3.16% of the total words. The most striking result that emerged from the data was on domain-specific most frequent words related to the MHC like '*Berhormat'* (English translation: *Honourable*, frequency: 242042) and *'Pertua'* (English translation: *President/ Head/ to address the speaker*, frequency: 115207). These two words, namely '*Berhormat'* and '*Pertua'* contributed 1.3% and 0.62% of the total words in Parliament 13, respectively.

TABLE 2. TF-H and TF-1 in Parliament 13

TF-H					TF-1				
Rank	**Word**	**Freq.**	**English Translation**	**Word class**	**Rank**	**Word**	**Freq.**	**English Translation**	**Word class**
1	YANG	865993	that, which, as	relative phrase	62,627	ABACHA	1	Sani Abacha	noun
2	#	575191	(octothorp/ hashtag)	symbol	62,628	ABACUS	1		noun
3	INI	359559	this	adverb pronoun	62,629	ABAIKANLAH	1	disregard	verb
4	DI	356640	at	preposition	62,630	ABAN	1	No meaning (original typing error in the report) most probably from Sya-aban (Islamic month)	noun
5	DAN	317300	and	conjunction	62,631	ABANDONE	1		verb
6	SAYA	289273	I, me	personal pronoun	62,632	ABANGLAH	1	brother	noun
7	KITA	268851	we	first pronoun	62,633	ABAR	1	No meaning (original typo in the report) most probably from *surat khabar* (newspaper)	noun

8	BERHORMAT	242042	honorary	adjective	62,634	ABARAKAATUH	1	Separate form from salam	noun
9	TIDAK	186315	no	function word	62,635	ABATED	1		vrb
10	TUAN	181520	Sir	noun	62,636	ABATES	1		verb
11	UNTUK	179897	to	function word	62,637	ABBAR	1	Company's name (Abbar PJS Limited)	noun
12	ITU	145947	that/ those	preposition	62,638	ABBASIYAH	1	Abbasid (Empire)	noun
13	DENGAN	142685	with	conjunction	62,639	ABDALLAH	1		name (noun)
14	ADA	142255	available there is/ there are exist/ contain/ present/ have/ be	adjective verb verb	62,640	ABDIDIN	1		name (noun
15	DALAM	138735	in	preposition	62,641	ABDOEL	1		name (noun
16	KEPADA	127004	to	preposition	62,642	ABDUCTED	1		verb
17	BIN	119265	son of	noun	62,643	ABDULLAHNYA	1		name (noun
18	PERTUA	115207	president/ head (to address the speaker)	noun	62,645	ABET	1		verb
19	AKAN	109620	will, shall	modal verbs	62,646	ABETTED	1		verb
20	JUGA	102007	too, also	adverb	62,647	ABHOR	1		verb

A closer inspection of Table 2 reveals that words that occurred only once in the sub-corpus (TF-1) were proper nouns and verbs. The words '*Abbasiyah'* (meaning *Abbasid Empire*) and *Abdallah* (referring to a name), for example, fall under the noun lexical category. English verbs also occurred in Parliament 13 even though Malay was used as the official medium of language in Parliament 13. English words occurred in the parliamentary debates because the use of English by the MPs was allowed with the permission of the Speaker of the House. Words like *abet, abetted* and *abhor* are the English verbs listed in Table 2.

The use of *abet* and *abetted* can be seen in the following extract:

TABLE 3. Extraction of word use in context for *abet* and *abetted* in Parliament 13

Extraction	"*Kalau kita ambil 27(1), dengan izin saya baca dalam bahasa Inggeris*, on the third line, *yang dikatakan*, "...no agent provocateur shall be presumed to be unworthy of credit by reason only of his having attempted to commit or to **abet**, or having **abetted** or having been engaged in a criminal conspiracy..."
Translation	If we take 27(1), with the permission (by the Speaker), I would like to read in English. on the third line, which says, "...no agent provocateur shall be presumed to be unworthy of credit by reason only of his having attempted to commit or to **abet**, or having **abetted** or having been engaged in a criminal conspiracy..."

Source: House of Representative dated October 9, 2017 (Parliament 13)

Abet and *abetted* in Table 3 were used to present to the MPs an Act in Penal Code (Malaysian laws) that talks about preventive criminalisation. Based on the context, these two

words function as verbs. Verbs like nouns are lexical categories that have semantic meanings and thus are not suitable to be selected as stop words despite having a unique occurrence (TF-1).

Tables 4 and 5 show the top 20 words in English and Malay according to frequency in Parliament 1 – TF-H and the bottom 20 words according to frequency (after removal of numerals) in Parliament 1 – TF-1. Similar to Parliament 13, Parliament 1 also exhibited matching result for TF-H in terms of function words which emerged to be the most frequently occurring words. However, the result indicated that the top 20 most frequent words in Parliament 1 were English function words like *of* (138450), *to* (119831), *in* (75177), *is* (61722), *that* (65117) and *this* (42983). The percentages of word occurrence compared to the size of the corpus (6,060,551) were 2.23% (*of*), 1.88% (*to*), 1.24% (*in*), 1.02% (*is*), 1.07% (*that*) and 0.71% (*this*).

TABLE 4. TF-H and TF-1 in Parliament 1 - English

TF-H				TF-1			
Rank	**Word**	**Freq.**	**Word class**	**Rank**	**Word**	**Freq.**	**Word class**
1	THE	256586	determiner	30859	ABASEMENT	1	noun
2	#	222385	symbol	30861	ABATED	1	verb
3	OF	138450	preposition	30863	ABB	1	noun
4	TO	119831	preposition	30864	ABBA	1	noun
7	AND	82914	conjunction	30865	ABBERATION	1	noun
8	IN	75177	preposition	30866	ABBREVIATED	1	verb
9	THAT	65117	determiner	30871	ABDICATION	1	noun
10	IS	61722	verb	30893	ABEGGING	1	noun
13	A	58473	verb	30894	ABELL	1	noun
14	I	49137	pronoun	30895	ABERDEEN	1	noun
17	THIS	42983	determiner	30897	ABIDED	1	verb
21	FOR	37733	preposition	30898	ABIDES	1	noun
22	IT	34453	pronoun	30902	ABOLISHES	1	verb
23	BE	32857	verb	30903	ABOLISHMENT	1	noun
26	NOT	28701	adverb	30904	ABOMINABLE	1	adjective
31	HAVE	26062	verb	30905	ABORGINES	1	noun
32	WE	24624	pronoun	30907	ABORIGINEES	1	noun
33	ARE	23888	verb	30908	ABORTED	1	verb
34	AS	23279	conjunction, preposition, adverb	30909	ABORTIONS	1	noun

TABLE 5. TF-H and TF-1 in Parliament 1 - Malay

TF-H					TF-1			
Rank	**Word**	**Freq.**	**English Translation**	**Word class**	**Rank**	**Word**	**Freq.**	**English Translation**	**Word class**
5	YANG	109222	that, which, as	relative phrase	30881	ABDOMEN	1	abdomen	noun
6	DI	105993	at	preposition	30885	ABDUH	1	name	noun
11	INI	60414	this	pronoun	30980	ADAKALA	1	sometimes	adverb
12	ITU	59227	that/ those	determiner	30982	ADALAH	1	is	verb
15	DAN	45663	and	conjunction	30983	ADA'LAH	1	Other spelling for *adalah* - is	verb
16	NYA	43106	it	pronoun	30985	ADAP	1	yellow rice	noun
18	SAYA	42362	I, me	pronoun	31080	AGAM	1	mansion	noun

			lah	Spoken interjection used in Malay language at the end of sentence/ phrase to emphasize meaning			1	conventional spelling for *agihkan* - distribute	verb
19	LAH	40663	conventional Malay spelling for "se", which also means 'one" or "a"	Adverb pronoun	31083	AGEHKAN	1	aggregation	noun
20	SA	39952	in	preposition	31085	AGGERASI			
24	DALAM	32100	with	preposition	31108	AHAMAD	1	name	noun
25	DENGAN	29543	have, is	verb	31132	AIDA	1	name	noun
27	ADA	27973	we	pronoun	31134	AIDUL	1	name	noun
28	KITA	27918	no	determiner	31143	AINUL	1	ice	noun
29	TIDAK	26570	Conventional spelling for *encik* - Sir	noun	31157	AIS	1	teach	verb
30	ENCHE	26302	Sir	noun	31168	AJARI			
37	TUAN	21356	have	verb	31174	AJOK	1	imitate	verb
39	TELAH	19331	son of	noun	31175	AJOKAN	1	imitation	noun
43	BIN	18650	Conventional spelling for *kepada*, similar meaning to the word "to"	preposition	31177	AKAD	1	solemnization	noun
44	KAPADA	18493			31180	AKALKAN	1	inspire	verb

A closer inspection of TF-1 words in Table 4 indicates that all the words (sorted by alphabetical ranking) that occurred once in Parliament 1 fall under noun, verb, and adjective categories. The result of TF-1 in Parliament 1 showed that all the top words had semantic value (meaning) and thus were disregarded from the stop word list's compilation.

STOP WORD LIST

After the removal of TF-1 and TF-H words with semantic meaning, the stop word list was extracted from Parliament 1 and 13. This section presents the stop word list after the deletion of content words from TF-H and TF-1.

Table 6 shows the top 20 stop words extracted from Parliament 1 in English and Malay language respectively, sorted according to frequency of occurrence (see **Appendix A** and **Appendix B** for the full lists of English and Malay stop words). Evidently, both English and Malay stop words could be extracted from Parliament 1. However, the top 20 English stop words in Parliament 1 revealed the domination of English words compared to Malay words. Similar to the English stop words, TF-H English words in Parliament 1 also came from function words like *the, of, to, and, is,* and *that* as can be observed in Table 6. These words are standard English stop words.

TABLE 6. The top 20 Malay and English stop words in Parliament 1 sorted according to frequency of occurrence

P1 Malay				P1 English			
N	**Word**	***f***	***f relative***	**N**	**Word**	***f***	***f relative***
1	YANG	109222	1.802179	5	THE	256586	4.233707
3	DI	105993	1.7489	6	OF	138450	2.284446
4	INI	60414	0.99684	11	TO	119831	1.977229
7	ITU	59227	0.977254	12	AND	82914	1.368093
8	A	58473	0.964813	13	IN	75177	1.240432
9	I	51837	0.810768	14	THAT	65117	1.07444
10	DAN	45663	0.753446	15	IS	61722	1.018422
17	NYA	43106	0.711255	16	THIS	42983	0.709226
21	SAYA	42362	0.698979	18	FOR	37733	0.6226
22	LAH	40663	0.670946	19	IT	34453	0.56848
23	SA	39952	0.659214	20	BE	32857	0.542145
26	DALAM	32100	0.529655	24	NOT	28701	0.473571
31	DENGAN	29543	0.487464	25	HAVE	26062	0.430027
32	ADA	27973	0.461559	27	WE	24624	0.4063
33	KITA	27918	0.460651	28	ARE	23888	0.394156
34	TIDAK	26570	0.438409	29	AS	23279	0.384107
35	ENCHE	26302	0.433987	30	ON	22529	0.371732
36	TUAN	21356	0.352377	37	SIR	21905	0.361436
38	TELAH	19331	0.318964	39	WILL	20423	0.336983
40	BIN	18650	0.307728	43	BY	19326	0.318882

Following the comparison made with the available stop word list in Mohd Taufik et al. (2005), distinctive similarities and differences were found in the list extracted from the current study. Similar to the list of stop words from Mohd Taufik et al. (2005), Malay stop words which were taken from TF-H in Parliament 1 also contained similar words like '*yang'*, '*di'*, *'itu'*, *'ini'*, '*ada'*, and *'kita'*. These words are basically the function words that occurred frequently in Parliament 1.

As opposed to the controlled stop word list from Mohd Taufik et al. (2005) which is more general and represents the use of current language setting, TF-H in Parliament 1 contained words with conventional Malay spelling like '*sa'* which is now spelled as '*se'*. This word '*sa'* means *one* in English and only occurred in Parliament 1. Other than function words, there were also words that uniquely belonged to the parliamentary discourse in Parliament 1. The words include '*Enche'* (Sir), '*Tuan'* (Sir) and '*bin'* (son of). These were the words most frequently used during the parliamentary debates in the parliament of Malaysia as the setting of the parliament requires the application of formal language. As a result, all of the MPs used formal Malay language. The MPs would address other MPs with '*Yang Berhormat'* (Honourable), '*Enche'* (Mr. /Sir) or '*Tuan'* (Sir). Therefore, this explains the high occurrence of these words in Parliament 1. The use of conventional spelling like '*Enche'* occurred because Parliament 1 was set in motion in 1959 and was prolonged until 1963. During that time, the old Malaysia or Tanah Melayu still adapted the use of conventional Malay spelling introduced by Za'ba in 1949. The new spelling of Malay only started in 1972 (Muhamed Salehuddin, 2021).

Table 7 shows the top 20 stop words extracted from Parliament 13 in English and Malay respectively, sorted according to frequency of occurrence. Based on the statistics, 173 and 21 English stop words were extracted from Parliament 1 and Parliament 13, respectively (see **Appendix A** and **Appendix B** for the full lists of English and Malay stop words). In contrast, Parliament 13 has 231 Malay stop words while Parliament 1 has 135. Similar to Parliament 1, English stop words were also found in Parliament 13. It can be seen from Table 7 that English function words still dominate the list in Parliament 13. Words like *the, of, to, is* and *you* occurred as TF-H in Parliament 13.

TABLE 7. The top 20 Malay and English stop words in Parliament 13 sorted according to frequency of occurrence

P13 Malay				P13 English			
N	**Word**	***f***	***f relative***	**N**	**Word**	***f***	***f relative***
1	YANG	865993	4.676507473	26	DR	84585	0.456773
3	INI	359559	1.941678882	96	THE	26244	0.141722
4	DI	356640	1.925915718	156	OF	16435	0.088752
5	DAN	317300	1.713473201	196	TO	12572	0.067891
6	SAYA	289273	1.562122703	225	IS	11450	0.061832
7	KITA	268851	1.451840401	274	YOU	9863	0.053262
8	BERHORMAT	242042	1.307067394	303	AND	8667	0.046803
9	TIDAK	186315	1.006132245	334	SO	7921	0.042775
10	TUAN	181520	0.980238438	391	IN	6580	0.035533
11	UNTUK	179897	0.971473932	484	THAT	5279	0.028507
12	ITU	145947	0.78813827	516	WE	4937	0.026661
13	DENGAN	142685	0.770522892	523	IT	4859	0.026239
14	ADA	142255	0.768200815	575	THIS	4330	0.023383
15	DALAM	138735	0.751892238	619	NOT	4042	0.021827
16	KEPADA	127004	0.685842872	656	FOR	3785	0.02044
17	BIN	119265	0.644050956	702	BE	3451	0.018636
18	PERTUA	115207	0.622137129	712	ARE	3400	0.018361
19	AKAN	109620	0.591966391	731	ON	3315	0.017902
20	JUGA	102007	0.550854862	736	NO	3258	0.017594
21	MENTERI	95447	0.515429795	812	HAVE	2947	0.015914

The result of the Malay stop word list in Parliament 13 showed distinction not only in comparison to Parliament 1 but the controlled stop word list by Mohd Taufik et al. (2005) as well. The words occurring the most in Parliament 13 were '*yang'* and '*ini'*. These are Malay function words that have low semantic meaning, and function to complement other words. The stop word list of Parliament 13 showed that domain-specific stop words were highly presented as words like '*Berhormat'* (Honourable) and '*Pertua'* (*President/ Chief/ addressing the Speaker of the House*) frequently occurred in Parliament 13. These are domain-specific words related to parliamentary debates compared to standard stop words that only contain function words like '*ada'* (have), *'yang'* (which, that), *'saya'* (I, me) and so on.

REMOVAL OF SIMILAR WORDS/ DUPLICATION

After the list of stop words was extracted, the two lists from Parliament 1 and Parliament 13 were merged into one. Words that occurred in both parliaments were then deleted. 21 English stop words and 50 Malay stop words were removed because of duplication during comparison of lists between Parliament 1 and Parliament 13. However, there were words that existed only in Parliament 1 and words that only occurred in Parliament 13. The distribution of stop words in English and Malay were 32% and 68%, respectively. Malay stop words were found to have a higher distribution because of the use of Malay in parliamentary debates. The use of the English language in the current parliaments is much less; nonetheless, English words did occur a lot because of the official use of English as the medium of debate in Parliament 1.

CATEGORISATION OF STOP WORD LIST IN MHC

The following section discusses the categorisation of stop word list that emerged based on the top 1,000 most frequent stop words found in Parliament 1 and Parliament 13. Following the construction of a stop word list which was separated in Malay and English in Parliament 1 and Parliament 13, the list of the stop words was categorised according to its function. Other than function words including auxiliaries, determiners, modals, prepositions, pronouns, question words, qualifiers and verbs, the stop word list in MHC also consisted of other categories

including numerals and time indicators, official's titles and salutation, and parliamentary-specific words. These categories therefore offer a domain-specific stop word list rather than a more general stop word list offered in open-domain stop word. The categorisation of the Malay and English domain-specific stop word lists can be seen in Tables 8 and 9, respectively.

Table 8 shows the breakdown of categories of Malay stop word list that did not significantly emerge in the open-domain stop words used for the study. As can be seen from Table 8, it is apparent that the category Parliamentary-specific Words (column 4) has the most stop words. Based on this category, a set of domain-specific stop words like '*akta'* (act), *'dasar'* (policy), *'rang'* (bill) and '*usul'* (motion) uttered in the parliament specifically belonged to the parliamentary debates. These words are not frequently used in open-domain language. The word '*rang'*, for example, occurred only 0.2276 times in a million word in the MPGC compared to the MHC (703.33 times per million words). Similar occurrence can be seen in the other words in comparison to one million word of occurrence. Comparison of the result suggests that there are domain-specific stop words for the Malay language in the parliament. Compared to the general wordlist, the domain-specific stop word list occurred significantly higher in the parliamentary corpus compared to the general corpus. This result suggests that the stop words belong to the domain-specific stop word list compared to the general stop word list.

TABLE 8. Comparison of Malay's domain-specific stop word in MHC and general word

Official's Titles/ Salutation (domain-specific stop word)	**Occurrence in the MHC (per million words)**	**Occurrence in MPGC (per million words)**	**Parliamentary-specific Words (domain-specific stop word)**	**Occurrence in the MHC (per million words)**	**Occurrence in MPGC (per million words)**
BERHORMAT (Honourable)	8105.9563	14.9	*AHLI* (Member)	2807.8517	110.14782
DATO (a title given to a person upon being conferred with certain orders of honour)	3002.9145	236.3600	*AKTA* (Act)	807.6872	207.5360
DATUK (a title given to a person upon being conferred with certain orders of honour)	2831.7580	979.8390	*BAHAGIAN* (division)	4017.0759	654.8914
ENCHE (Mr)	296.7708	0	*DASAR (policy)*	638.9917	373.7570
HAJAH (a title for a Muslim woman who has made a pilgrimage to Mecca)	171.6617	16.1417	*HAL* (matter)	518.0102	580.7165
HAJI (a title for a Muslim man who has made a pilgrimage to Mecca)	4695.4280	342.8188	*IZIN* (permission)	669.9327	59.7627
PUAN (Madam)	431.4214	119.9097	*JAWAB* (answer)	365.5628	87.2420
PERTUA (Head, chair)	5349.2898	7.1100	*KETAWA* (laughter)	392.5533	50.5389
TUAN (Sir)	11075.5741	259.6122	*MENTERI* (Minister)	4900.4456	1208.5130
SERI (a prefix title given to a person upon being conferred with certain orders of honour)	933.10475	540.9387	*MESYUARAT* (meeting)	592.6754	243.2783
SRI	736.1260	270.3733	*PINDAAN* (amendment)	524.4871	53.2291
			PTG	241.5298	0.5765

(a prefix title given to a person upon being conferred with certain orders of honour)	(short form for *petang* which means *evening*)		
	RANG (Bill)	703.3300	0.2276
	RIUH (noisy)	108.5592	11.5298
	TANYA (ask)	325.6775	130.0943
	TEPUK (clap)	136.9015	3.6511
	USUL (motion)	287.3277	4.2276

Table 9 presents the examples of '*ketawa'* (laugh, laughter) and '*izin'* (permission) which were taken from the extracts of the MHC and the MPGC. Based on the examples, the use of '*ketawa'* in the MPGC represents the action of laughing while its use in the MHC was written in brackets. This action indicates the mood and situation in the parliament where everybody burst into laughter. The use of these words can be seen in Parliament 2 to Parliament 13 where the Malay language is applied as the official language in the parliament, and the verbatim and reporting style used involved the Malay language. For example, '*ketawa'* was repeatedly used to indicate the MPs bursting into laughter in the parliamentary proceedings. The same goes for the word '*tepuk'* (meaning *clap*) which was used repeatedly to indicate the MPs' support towards certain issues. These physical actions are normally written as '[*tepuk*]' and '[*ketawa*]'.

In the case of the word '*izin'* (see Table 9), observation of the Malay language general corpus indicates that it has two uses. The first refers to the adjective *illegal* (as in *illegal* immigrant) while the other means *permission*. Nevertheless, the use of '*izin'* in parliamentary discourse indicates the action of obtaining permission from the Speaker of the House upon using the English language.

TABLE 9. Comparison of the use of '*ketawa'* (laugh) and *'izin'* (illegal, permission) in the MHC and the MPGC

Node word	**Corpora**	**Extract from corpus**	**Translation**
ketawa (laugh, laughter)	MPGC	*Ayah saya kata kamu adalah orang asing dan berlalu sambil* ***ketawa*** Source text: Utusan	My dad said you are a stranger and she left while **laughing**
	MHC	*Dengan Yang Berhormat, saya senyumlah. Tak apa, no problem* ***[Ketawa]*** Source text: House of Representative dated March 27, 2018 (Parliament 13)	I will just smile to you. No problem [**Laughter**]
izin (illegal, permission)	MPGC	*Ini diikuti dengan langkah membenarkan pendatang asing tanpa* ***izin*** *meninggalkan negara ini.* Source text: Utusan	This is followed by the steps to allow illegal **immigrant** to leave this country
	MHC	I talk to you nicely, *dengan* ***izin*** Source text: House of Representative dated March 29, 2018 (Parliament 13)	I talk to you nicely, with **permission** (from the Speaker)

Table 10 shows the breakdown of categories of English stop word list that did not significantly emerge in the open-domain stop words used for the study. The results obtained from the analysis indicate similarity of frequency of occurrence of domain-specific stop words in the MHC compared to the open-domain wordlist. However, words like *Honourable, applause*, *mentioned* and *sitting* exclusively belonged to the parliamentary debates uttered in the parliament as they did not occur in the MPGC. In addition, words like *clause*, *constitution*,

question, *report* and *regard* also showed significant occurrence in the MHC compared to the MPGC. Similar to the Malay stop words discussed in Table 8, comparison of this result suggests that there is a domain-specific stop word for the Malay language in the parliament. Compared to the general wordlist, the domain-specific stop word list occurred more frequently in the parliamentary corpus compared to the general corpus. This result suggests that the stop words belong to the domain-specific stop word list compared to the general stop word list.

TABLE 10. Comparison of English domain-specific stop words in the MHC and general MPGC

Official's Titles/ Salutation (domain-specific stop word)	**Occurrence in MHC (per million words)**	**Occurrence in Malay language general corpus (per million words)**	**Parliamentary-Specific Words (domain-specific stop word)**	**Occurrence in MHC (per million words)**	**Occurrence in Malay language general corpus (per million words)**
HONOURABLE	251.9113	0	ACT	7.913	2.6902
MEMBER	161.7199	3.07461	AMENDMENT	16.27	0.1921
MEMBERS	107.5289	0.5765	APPLAUSE	4.61	0
MINISTER	380.9250	1.15298	ASK	9.85	8.8395
SIR	435.2997	6.1492	BEG	9.96	4.3813
SPEAKER	318.7149	4.9962	CHAIR	5.18	0.1922
			CHAIRMAN	62.41	0
			CLAUSE	13.25	0. 3843
			CONSTITUTION	11.38	0.1921
			LAUGHTER	9.53	0.3843
			MENTIONED	4.55	0
			MOTION	16.98	0.5380
			ORDER	30.46	5.95705
			ORDINANCE	11.06	0.5764
			PARTY	11.44	4.0354
			QUESTION	82.0	0.3843
			REGARD	9.83	0.1921
			REPORT	15.52	1.7294
			SECOND	15.73	2.8824
			SITTING	5.49	0

Other words like *second* and *order* have different connotations across the open-domain and domain-specific contexts. Examples are provided in Table 11.

TABLE 11. Comparison of the use of *second* and *order* in the MHC and MPGC

Node Word	**Corpora**	**Extract from Corpus**	**Translation**
second	MPGC	*Dalam menghadapi sesuatu penyakit pun, kita kerapkali mendapatkan second opinion daripada doktor yang lain.* Source text: Utusan	We would always seek for **second** opinion from another doctor if we are dealing with any sicknesses.
	MHC	I beg to second the motion. Source text: House of Representative dated January 11, 1964 (Parliament 1)	
order	MPGC	*Siapa sangka BSN yang dulunya macam Pejabat Pos, hanya ada "*money **order***" tetapi sekarang dah ada "credit card* Source text: Utusan	Who would have thought that BSN which was previously like the post office and only had money **order** now already has credit card
	MHC	Sir, on a point of order—Standing **Order** 36 (1). I think he must confine his observations to what is asked for in the Development Estimates Source text: House of Representative dated January 11, 1964 (Parliament 1)	

In the context of general Malay, *second* is used as an adjective denoting number two in a sequence while in the parliamentary context, *second* functions as a verb that refers to the action of formally endorsing a motion for further discussion or action.

The word *order* (Table 11) is also used differently in the MPGC and MHC. In the MPGC, it is used to refer to an official piece of paper with a specified amount of money printed on it which is usually issued by a post office, and you can send or give it to someone who can then exchange it for the same amount of money. It can also mean the arrangement or disposition of people with regards to a specific arrangement, pattern, or approach. However, in the parliamentary procedure, a point of order occurs when someone draws attention to a violation of rules in a meeting of a deliberative assembly. This is depicted in the MHC extract in Table 11.

DISCUSSION

This study set out to identify Malay stop words in the Malaysian parliamentary discourse through the Malaysian Hansard Corpus by utilising the corpus from Parliament 1 (year 1959-1963) and Parliament 13 (2013-2018). The study also sought to propose a framework for extracting stop words from Malay parliamentary documents that includes looking for the characteristics and statistics of the words in determining stop word traits. In addition, a methodology for extracting Malay-specific-domain stop word list was also proposed.

The most evident finding to have emerged from this study is the extraction of 587 stop words in relation to the MHC. The findings of this study provide insights for future research in Malay language data processing, and more specifically, for further analysis of MHC as well as domain-specific or specialised analysis of corpus. Interestingly, a theme that emerged from the findings is the categorisation of the stop word list based on the MHC. This categorisation indicates a domain-specific stop word list that exclusively belonged to the parliamentary discourse, specifically Malaysian parliamentary discourse. The existence of such categorisation would assist future works on domain-specific NLP. It would also provide a framework for the analysis of domain-specific stop words, especially in the Malay language in the future. Through the use of the domain-specific stop word list, words with low value but high in frequency of occurrence in parliamentary corpus would be discarded prior to text analysis. This can be performed through the use of Natural Language Toolkit (NLTK), a suite that consists of libraries and programmes to execute statistical language processing. Additionally, these words can be discarded through the use of deletion of duplicated words using Microsoft Excel (for wordlist).

The existence of this domain-specific stop word list suggests that it can be applied prior to text mining in NLP and in corpus linguistics. Removal of stop words can be performed using tools like Python. Alternatively, it can also be removed manually through the removal of duplicated words in the wordlist through the use of simple tools like Microsoft Excel.

There are many corpora, especially specialised or domain-specific corpora developed by various researchers for different purposes. In this regard, the availability of the proposed model will enable other researchers to extract stop word lists of their own corpora to produce different sets of domain-specific stop word lists.

The performance of this method can also be expanded to 5,000 to 10,000 word ranks in the wordlist. An issue that was not addressed in this study is whether the stop word list produced can be employed in general NLP for the Malay language. Nevertheless, prior to publishing this study's findings, the stop word list generated from this present study was used in a pilot study by Sabrina, Nor Fariza, et al. (2020), and its utilisation has been proven to be successful. Thus, it can be expected that the stop word list produced in this study will be able to assist NLP and corpus-based or corpus-driven analysis, particularly in the Malay language.

CONCLUSION

The aim of the present study was to propose a method for extracting Malay stop words and to produce a list of standardised stop words from a domain-specific parliamentary corpus in the Malay language. In this study, a computational approach which adapted and adopted the Z-method proposed by Zipf (1949), Hassan et al. (2014), and Hofmann et al. (2020) has been demonstrated. This method extracted stop words by selecting the most frequent words (TF-H) and the words that appeared only once or with frequency of one (TF-1). It also utilised the classic method where previous stop word list in Malay was used as a groundwork for the stop word list. The automated selection of the most frequently occurring words and the singleton words was able to speed up the search in determining stop words.

In general, this study has not only contributed to the growing literature on stop word list extraction but it has also provided a new stop word list that can be utilised to assist in NLP and data analysis in corpus linguistics. The method could potentially be employed to extract stop words in a Malay corpus, other Malay domain-specific corpus and a corpus that consists of a mixture of Malay and English.

ACKNOWLEDGEMENT

This research is supported by Universiti Kebangsaan Malaysia, research grant KRA-2018-005.

REFERENCES

Alshanik, F., Apon, A., Herzog, A., Safro, I. & Sybrandt, J. (2020). Accelerating text mining using domain-specific stop word lists. 2020 IEEE International Conference on Big Data (Big Data), 2639-2648.

Ayral, H. & Yavuz, S. (2011). An automated domain specific stop word generation method for natural language text classification. 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul.

Baldwin, T. & Su'ad Awab. (2006). Open source corpus analysis tools for Malay. Proceedings of the Fifth International Conference on Language Resources and Evaluation, Italy.

Chekima, K. & Alfred, R. (2016). An automatic construction of Malay stop words based on aggregation method. In M. Berry, Hj. Mohamed A., & B. Yap, (Eds.). *Soft computing in data science*. *Communications in Computer and Information Science*, Vol. 652. Singapore: Springer.

Chong, T.Y., Banchs, R.R. & Chng, E.S. (2012). An empirical evaluation of stop word removal in statistical machine translation. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. France: Association for Computational Linguistics.

Choy, M. (2012). Effective listings of function stop words for Twitter. *International Journal of Advanced Computer Science and Application.* 3(6), 8–11.

Chua, S. & Nohuddin, P.N.E. (2017). Relationship analysis of keyword and chapter in Malay-translated tafseer of al-Quran. *Journal of Telecommunication, Electronic and Computer Engineering*. *9*(2-10), 185-189.

Haddi, E., Liu, X. & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Comput. Sci*. *17*, 26–32.

Fatimah Dato Ahmad (1995). A Malay language document retrieval system: An experimental approach and analysis. Unpublished PhD thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Fatimah Sidi, Marzanah Abdul Jabar, Mohd Hasan Selamat, Abdul Azim Abd Ghani, Md. Nasir Sulaiman & Salmi Baharom (2011). Malay interrogative knowledge corpus. *American Journal of Economics and Business Administration*. *3*(1), 171–176.

Green, D. & Cross, J, P. (2017*). Exploring the political agenda of the European Parliament using a dynamic topic modeling approach*. Cambridge: Cambridge University Press.

Hassan Saif, Fernández, M., He, Y. & Harith, A. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. Proceeding of Ninth International Conference on Language Resources and Evaluation, Iceland. 810–817.

Hamood Ali Alshalabi, Sabrina Tiun & Nazlia Omar (2017). A comparative study of the ensemble and base classifiers performance in Malay text categorization. *Asia-Pacific Journal of Information Technology and Multimedia*. *6*(2), 53–64.

Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., & Wissik, T. (2020). Comparing lexical usage in political discourse across diachronic corpora. Proceedings of ParlaCLARIN II Workshop, 58–65.

Imran Ho-Abdullah, Zaharani Ahmad, Rusdi Abdul Ghani, Nor Hashimah & Idris Aman (2004). A practical grammar of Malay – A corpus-based approach to the description of Malay. First COLLA Regional Workshop. Malaysia: Putrajaya, June.

Imran Ho Abdullah, Anis Nadiah Che Abdul Rahman & Azhar Jaludin (2017). *The Malaysian Hansard Corpus.*

Kaur, J. & Buttar, P.K. (2018). A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*. *4*(4), 207–210.

Keshavarz, H. & Abadeh, M.S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems*. 122, 1–16.

Khan, N., Bakht, M.B., Khan, M.J., Samad, A. & Sahar, G. (2019). Spotting Urdu stop words by Zipf's statistical approach. 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). 1–5, doi: 10.1109/MACS48846.2019.9024817.

Koteyko, N. (2014). Compilation of specialised corpora. In *Language and politics in Post-Soviet Russia*: *A corpus-assisted approach* (pp. 48–64). London: Palgrave Macmillan.

Kwee, A.T., Tsai, F.S. & Tang W. (2009) Sentence-level novelty detection in English and Malay. In T. Theeramunkong, B. Kijsirikul, N. Cercone, & T.B. Ho, (Eds.). *Advances in knowledge discovery and data mining*. PAKDD 2009. Lecture Notes in Computer Science, Vol. 5476. Berlin: Springer. https://doi.org/10.1007/978-3-642-01307-2_7

Liu, J., Ren, X., Shang, J., Cassidy, T., Voss, C.R. & Han, J. (2016). Representing documents via latent keyphrase inference. *Proc Int World Wide Web Conf*. 1057–1067. doi: 10.1145/2872427.2883088.

Lo, R. T.-W., He, B. & Ounis, I. (2005). Automatically building a stopwordlist for an information retrieval system. *J. Digit. Inf. Manag. Spec. Issue.* 5th Dutch-Belgian Inf. Retr. Work. 5(2005), 17–24.

Luhn, H.P. (1960). Key word-in-context index for technical literature (KWIC Index). *American Documentation*. 11, 288–295.

Makrenchi, M. & Kamel, M.S. (2017). Extracting domain-specific stopwords for text classifiers. *Intelligent Data Analysis*. *21*(1), 39–62.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Mohd Amin Mohd Yunus, Aida Mustapha & Noor Azah Samsudin (2017). Query translation and Quran result in TreeMap. MATEC Web of Conferences 135. 1–7.

Muhammed Salehudin Aman (2021). Sinopsis sistem ejaan Bahasa Melayu. KLIKWeb DBP. Retrieved May 7th, 2021 from http://klikweb.dbp.my/?p=6003

Muhamad Taufik Abdullah (2006). Monolingual and crosslanguage information retrieval approaches for Malay and English language documents. Unpublished Ph.D thesis. Universiti Putra Malaysia, Serdang, Malaysia.

Muhamad Taufik Abdullah, Fatimah Ahmad, Ramlan Mahmod, & Tengku Mohd Tengku Sembok (2005). Improvement of Malay information retrieval using local stop words. International Advanced Technology Congress: Conference on Computer Integrated Systems. Putrajaya, Malaysia.

Munková, D., Munk, M. & Vozár, M. (2014). Influence of stop-words removal on sequence patterns identification within comparable corpora. In V. Trajkovik & A. Mishev, (Eds.). *Advances in intelligent systems and computing* (pp. 67–76). Switzerland: Springer International Publishing Switzerland.

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), 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), 66–95.

Puri, R, Bedi, R. P. S. & Goyal, V. (2013). Automated stopwords identification in Punjabi documents. *An Int. J. Eng. Sci. 8*(2013), 119–125.

Rani, R. & Lobiyal, D.K. (2018). Automatic construction of generic stop words list for Hindi text. *Procedia Computer Science*. *132*, 362-370.

Raulji, J.K & Saini, J.R. (2016). Stop-word removal algorithm and its implementation for Sanskrit language. *International Journal of Computer Applications*. *150*(2), 15–17.

Raulji, J.K. & Saini, J.R. (2017). Generating stopwordlist for Sanskrit language. 2017 IEEE 7th International Advance Computing Conference (IACC).

Rose, S., Engel, D., Cramer, N. & Cowley, W. (2010). Automatic keyword extraction from individual documents. In Berry, M.W., & Kogan, J., (Eds.). *Text mining: Applications and theory*. New Jersey: John Wiley and Sons, Ltd.

Sabrina Tiun, Nor Fariza Mohd Nor, Azhar Jalaludin & Anis Nadiah Che Abdul Rahman. (2020). Word embedding for small and domain-specific Malay corpus. In Alfred R., Lim Y., Haviluddin H., & On, C., (Eds). *Computational science and technology. Lecture notes in electrical engineering*. Singapore: Springer.

Sabrina Tiun, Saidah Saad, Nor Fariza Mohd Nor, Azhar Jalaludin & Anis Nadiah Che Abdul Rahman (2020). Quantifying semantic shift visually on a Malay domain-specific corpus using temporal word embedding approach. *Asia-Pacific Journal of Information Technology and Multimedia*. *9*(2), 1–10.

Sadeghi, M. & Vegas, J. (2014). Automatic identification of light stop words for Persian information retrieval systems. *Journal of Information Science. 40*(4).

Scott, M. (2008). WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

Weisser, M. (2103). Tools, ideas & resources for linguistics. Retrieved November 18, 2020 from http://martinweisser.org/

Wild, F., Kalz, M., Demnati, H., Paliwoda-Pekosz, G. & Naili, M. (2020). Stopwords: Stop wordlists in German, English, Dutch, French, Polish, and... in lsa: Latent Semantic Analysis. R Package Documentation. Retrieved November 4, 2020 from https://rdrr.io/cran/lsa/man/stopwords.html

Yuan, T., Lo, D., & Lawall, J. (2014). Automated construction of a software-specific word similarity database. 2014 Software Evolution Week - IEEE Conference on Software

Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE) 2014. 44–5. doi: 10.1109/CSMR-WCRE.2014.6747213.

Zheng, A. (2018). *Feature engineering for machine learning*. Sebastool, USA: O'Reilly Media, Inc.

Zhi, L.G. (2003). Using mutual information to identify new features for text documents of various domains. PACLIC 2003. 372–379.

Zipf, G.K. (1949). *Human behavior and the principle of least Effort*. Cambridge, Massachusetts: Addison-Wesley.

APPENDIX A

English Stop Word List in the MHC, Sorted in Alphabetical Order (174 words)

about	by	his	mr	should	under
accept	can	honourable	much	since	until
according	cannot	how	must	sir	up
after	certain	however	my	so	upon
again	could	if	necessary	some	us
against	day	in	never	something	various
all	days	into	next	speaker	very
already	december	is	no	still	was
also	dr	it	non	such	we
although	during	its	not	than	well
always	each	itself	nothing	that	were
amount	end	january	now	the	what
an	even	just	number	their	when
and	every	last	of	them	where
another	few	laughter	on	themselves	whether
any	first	least	one	then	which
applause	five	less	only	there	while
are	for	like	or	therefore	who
as	from	lot	other	these	whole
at	fully	many	our	they	why
away	further	may	over	this	will
back	got	me	people	those	with
be	had	member	perhaps	three	within
because	has	members	put	through	without
been	have	mentioned	rather	time	would
before	he	might	really	to	yet
being	here	months	same	too	you
between	him	more	second	towards	your
but	himself	most	shall	two	

APPENDIX B

Malay Language Stop Word List in the MHC, Sorted in Alphabetical Order (413 word)

ada	berada	engkau	katakan	menjadi
adakah	berapa	engkaukah	ke	menyebabkan
adakan	berhormat	engkaulah	kedua	menyebabkannya
adalah	berikan	engkaupun	kemudian	mereka
adanya	berikut	hai	kenapa	merekalah
adapun	berkaitan	hajah	kepada	merekapun
agak	berkenaan	haji	kerajaan	meskipun
agar	berupa	hal	kerana	mesti
akan	beserta	hampir	ketawa	misalnya
akhir	biarpun	sebagai	ketiga	mu
aku	bila	hampir-hampir	ketika	mungkin
akulah	bilakah	hanya	khusus	nak
akupun	bilamana	hanyalah	kini	namun
al	bilangan	harus	kita	nanti
alangkah	bin	hendak	ku	nescaya
allah	binti	hendaklah	kurang	niscaya
amat	bisa	hingga	lagi	nya
antara	boleh	ia	lah	okey
antaramu	bukan	iaitu	lain	olah
antaranya	bukankah	ialah	lalu	oleh
apa	bukanlah	ianya	lamanya	orang
apa-apa	che	ii	langsung	pada
apabila	chuma	ingin	lebeh	padahal
apakah	cuma	inginkah	lebih	padanya
apapun	dah	ini	lima	padamu
atas	dahulu	inikah	macam	paling
atasmu	dalam	inilah	macham	para
atasnya	dalamnya	itu	maha	pasti
atau	dan	itukah	mahu	patut
ataukah	dapat	itulah	mahukah	patutkah
ataupun	dapati	izin	mahupun	pelbagai
bagai	dapatkah	jadi	maka	per
bagaimana	dapatlah	jangan	makin	pergilah
bagaimanakah	dari	janganlah	malah	perkara
bagaimanapun	daripada	jika	mana	perkaranya
bagi	daripadaku	jikalau	manakah	perlu
bagimu	daripadamu	jua	manakala	pernah
baginya	daripadanya	juapun	manapun	pertama
bagitu	dato	juga	maseh	ptg
bahawa	datuk	jumlah	masih	puan
bahkan	demi	ka	masing	pula
bahwa	demikian	kadang	masing-masing	pun
banyak	demikianlah	kah	md	punya
banyaknya	dengan	kalangan	melainkan	ra
barangkali	dengannya	kalau	mem	ramai
barangsiapa	di	kali	memang	riuh
bawah	dia	kami	mempunyai	sa
beberapa	dialah	kamikah	men	sadikit
begitu	didapat	kamipun	mendapat	sahaja
begitupun	didapati	kamu	mendapati	saja
belaka	dimanakah	sentiasa	mendapatkan	saling
beliau	dua	kamukah	mengadakan	sama
belum	empat	kamupun	mengapa	samakah
belumkah	enam	kan	mengapakah	sama-sama
ber	enche	kapada	mengenai	sambil

sampai	sejauh	seorangpun	sudahkah	terutamanya
samping	sekali	separuh	sunggoh	tetapi
sana	sekalian	sepatutnya	sungguhpun	tiada
sangat	sekalipun	seperti	supaya	tiadakah
sangatlah	sekarang	seraya	ta	tiadalah
saperti	sekejap	seri	tadi	tiap
satu	sekian	sering	tadinya	tiap-tiap
saya	sekiranya	serta	tahu	tidak
se	sekitar	seseorang	tahukah	tidakkah
seandainya	sekurang	sesiapa	tak	tidaklah
sebab	selain	sesuatu	tanpa	tiga
sebagaimana	selalu	sesudah	Tanya	tuan
sebagainya	selama	sesudahnya	tanyakanlah	turut
sebanyak	selama-lamanya	sesungguhnya	tapi	umpama
sebarang	selepas	sesungguhnyakah	telah	untok
sebelum	seluruh	setakat	tentang	untuk
sebelummu	seluruhnya	setelah	tentu	untukmu
sebelumnya	semakin	seterusnya	tepuk	wahai
sebenarnya	semasa	setiap	terdapat	walau
sebuah	sementara	siapa	terhadap	walaupun
secara	semua	siapakah	terhadapmu	ya
sedang	semuanya	sikit	terlalu	yaini
sedangkan	semula	sini	termasuk	yaitu
sedikit	senantiasa	situ	terpaksa	yakni
sedikitpun	sendiri	situlah	tersebut	yang
segala	seolah	sri	tertentu	
sehingga	seolah-olah	suatu	terus	
sejak	seorang	sudah	terutama	

ABOUT THE AUTHORS

Anis Nadiah Che Abdul Rahman is a Ph.D candidate in Linguistics at the Centre for Language and Linguistics, Universiti Kebangsaan Malaysia. Her research interests include semantics, culturomics, corpus linguistics and parliamentary discourse. anisnadiahcar@gmail.com

Imran Ho Abdullah (Ph.D), is Professor at the Centre for Language and Linguistics, Faculty of Social Sciences and Humanities, UKM. His areas of expertise are corpus linguistics and cognitive semantics, and is currently the Deputy Vice-Chancellor (Industry & Community Partnerships) UKM.

Intan Safinaz Zainudin (Ph.D), is a senior lecturer at the Centre for Language and Linguistics, Faculty of Social Sciences and Humanities, UKM. Her areas of interest include Translation, Semantics, Bilingual Lexicography and corpus-based studies.

Sabrina Tiun (Ph.D) is a senior lecturer at the Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia. Her research interests are Natural Language Processing to Speech Processing and Information Retrieval. She is a member of the Knowledge Computing research group under the Centre for Artificial Intelligence Technology, Faculty of Science & Information Technology, UKM.

Azhar Jaludin (Ph.D) is a senior lecturer at the Centre for Language and Linguistics, Faculty of Social Sciences and Humanities, UKM and his area of expertise is corpus linguistics and web crawling.

eISSN: 2550-2131
ISSN: 1675-8021