Indonesian Journal of Electrical Engineering and Computer Science
Vol. 6, No. 1, April 2017, pp. 212 ~ 217
DOI: 10.11591/ijeecs.v6.i1.pp212-217

Comparison of the Themes of Malaysian Friday Sermons Between the Year 2010 and 2015

Muhammad 'Aasim Asyafi'ie bin Ahmad*, Mokhtar bin Harun, Puspa Inayat binti Khalid, Mohd Ibrahim Shapiai, Md. Najib bin Ibrahi, Siti Zaleha Abdul Hamid
Fakulti Kejuruteraan Elektrik, UTM Skudai, 81310, Johor Bahru, Johor, Malaysia

Abstract

One of the analyses used in the field of corpus linguistics is comparing the word occurrence from different text corpora. This technique can be used to identify how a certain discipline change over time through text analysis. In this study, the changes of the context of Malaysian Friday sermons are investigated. The text corpus was developed by taking the Friday sermons spoken in Kuala Lumpur mosques in the year 2015. A total of 52 sermons were used for the text corpus because there are a total of 52 Friday sermons in a year. The Malay text corpus was constructed by using PHP and MySQL, and only the top words spoken were inserted into the text corpus. This text corpus is then compared with a previously developed text corpus from 2010 Friday sermons. The new text corpus overlapped with the old text corpus by 82%. Analysis also shows the difference of semantic between 2010 and 2015 Friday sermons.

Keywords*: phonetics, text corpora, malay language, linguistics, word lists*

1. Introduction

1.1. Corpus Linguistics and Word Lists

According to American National Standard, intelligibility is the property that allows units of speech to be identified [1]. Intelligibility can be tested by using objective methods such as Speech Transmission Index (STI) [2]. Another method to measure intelligibility is by using subjective methods such as Phonetically Balanced (PB) word lists and modified rhyme test (MRT) [1].

Developing word list is one of the areas of study of corpus linguistics. Corpus linguistic is defined as a computer-oriented approach to text analysis [3]. Modern corpus linguistics that uses computer-based corpora began to emerge in the 1990s. This allows the researchers to work with larger data volumes compared with manual techniques while at the same time reducing researchers' bias [4]. Researchers' bias can be defined as systemic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others [5]. Nowadays, there are more researches in the area of corpus linguistics mainly due to the advent in computing power.

The information contained in different corpora is informative when the corpora are taken from different sources or different times [3]. When two corpora are compared, the words that appear more frequently in one corpus compared to another can be identified [6]. The identified keywords can indicate certain aspect of the text inspected such as what the "text" is mainly all about, the stylistic characteristics, or the text genres [3].

For example, Xiao and McEnery [7] have compared the corpus of specialized spoken professional American English with the British National Corpus to compare the genres of conversation, speech, and academic prose in American English on the basis of the most common words and the least common words used. Johnson et al. [8] have used keywords to compare three newspapers over a 5-year period against the British National Corpus in an analysis of political correctness in the British newspapers.

Other recent researches such as done by Cohen [9] have used the same techniques of measuring the most common words used by the Palestinian suicidal terrorists in their farewell letters in an effort to understand their mind.

Received December 22, 2016; Revised February 15, 2017; Accepted February 28, 2017

The development of word list for languages can be divided into general world list such as phonetically balanced word list [1], which is a word list containing 1000 words divided into 50 different list. Other example of word list includes MRT [10], General Service List [11], Academic Word List [12], Academic Word List for Clinical Case [13], Medical Academic Word List [14], Medical Academy Vocabulary List [15], and Student Engineering English Corpus [16]. All of the examples are explored further in Table 1.

Table 1. Summary of different forms of word lists

Word List	Specification	References
PB Word List	Containing a total of 1000 monosyllable English words which are further divided into 50 different lists.	[1]
Modified Rhyme Test (MRT)	Containing a total of 300 monosyllable English words and are divided into 6 different word lists.	[10]
General Service List	Containing 2000 most frequently used English words in primary schools.	[11]
Academic Vocabulary List	Containing a total of 3200 most frequently used English 500 most common words.	[12]
Academic Word List for Clinical Case	Containing a total of 241 words that are: 1. Not available from the word list compiled by West [11]. 2. The words need to overlap 50% of the time within all medical fields. 3. Need to have word frequency of at least 30.	[13]
Medical Academic Word List	Containing a total of 650 words with the exact same criteria as outlined from [13].	[14]
Medical Academy Vocabulary List	Containing a total of 819 word families which are specifically used in the medical field.	[15]
Student Engineering English Corpus (SEEC)	A list containing about 2 million running words which is further divided into 1260 most frequent word families taken from 13 different English-language textbooks compulsory for all Engineering students at Walailak University.	[16]

1.2. Tools Used in Corpus Linguistics

In the field of corpus linguistics, there exists several tools used to assists researchers in handling large amount of text data. The most common used tool in corpus linguistic is the WordSmith Tool [17]. The papers that uses this tool are [3], [9], [16]. WordSmith Tool is not a free software and a single user license cost about 50 pound sterling [17]. Other researcher may use more specialised software such as Range [18], which is geared specifically for researhcers within the medical field such as developing Medical Academic Word List [13].

Some other researcher build programmes from scratch using specific programming languages. The benefit of this approach is that the programmes can be designed to specifically suit the need of the researchers and that it does not cost much when compared with commercial tool. For example, the researchers from [15] scripted a searching programme, using the programming language Python [19], to extract words from the corpus data.

In this study, the programme used to manipulate the data within the text corpus was developed by using the programming language Personal Home Page (PHP) [20] and MySQL [21].

2. Research Method

2.1. Obtaining The 2015 Friday Sermon Text Corpus

The data used to construct the text corpus was obtained from 52 Friday sermons transcripts that were delivered in mosques within Kuala Lumpur the year 2015. Only 52 Friday sermons were selected because this is the average number of times Friday sermons are delivered in a year since there are 52 weeks in a year. Also, since the subject covered within a Friday sermon varies depending on the time of the year, a year of Friday sermons were taken to ensure that all possible subjects covered are included within the text corpus. Only the transcripts of the sermons were taken into account because the audio within the video recordings have varying qualities and can be undecipherable at some times.

The speech transcripts are available in an official website of Jabatan Kebajikan Islam (JAKIM) [22]. The sermons that are presented were read in Kuala Lumpur mosques. And since

Kuala Lumpur mosques use standard Malay that can be understood across Malaysia, this will ensure that the results are applicable across the nation. Friday sermons are divided into two parts. In this study, only the first part of the Friday sermons are taken into account. This is because only the first part of the sermon changes throughout the year whereas the second part stays the same.

The data obtained were then stored into a database. The database management system used in this study is MySQL which is based on Structured Query Language, language that is designed to manage data. In addition, MySQL does not require any form of payment for licenses for non-commercial use. To manipulate the data within the stored database, the programming language Personal Home Page (PHP) was used in this study. PHP is used to input new data into the database, and then extract the data when needed.

The transcripts obtained from [22] need to be cleaned up before any Malay texts can be extracted. Images and Arabic texts were removed. Any Arabic numerals were replaced with Malay words. After that, the text files were manipulated so that each word are separated by lines. And then, these words were inserted into the database. These raw data are stored in 52 different tables, each table representing a different Friday sermons.

Next, all the tables are combined and then additional PHP codes were used so that the same words can be grouped together and then counting how many times the same word have occurred. The final result is a table within the database which contains all the words from the 2015 Friday sermons along with the word occurrence. This final table was then extracted into an Excel format.

2.2. Comparing 2015 Text Corpus with 2010 Text Corpus

After obtaining the 2015 text corpus, the previous 2010 text corpus taken from was also extracted. The 2010 text corpus was taken from [23]. The comparison between 2015 and 2010 text corpus was done because the word frequency can only be meaningfully analysed when compared with other text corpus; furthermore, it has been shown that a text corpus may show different patterns in a timeframe as short as three years [3]. From both of these corpora, only words that have word occurrence of 52 or more are taken. This is because since there are 52 Friday sermons taken into account, for a word to appear at least once in every sermon, it needs to have occurred at least 52 times within the corpus. After that, the words that have occurred at least 52 times from both corpora are stored in different tables and then compared.

3. Results and Analysis

3.1. Basic Information of The Two Text Corpora

The summary of information of these two corpora is tabulated in Table 2. The 2010 corpus have a 7% more of running words of total of 60673, where the 2015 corpus have 56842 running words. Running words are defined as individual words in a corpus, regardless whether if it is repeated or not. Other name for running word is “token”. Even though 2010 corpus have almost 4000 more running words than 2015 corpus, the number of different words found in 2010 corpus is almost as same as 2015 corpus; 5924 different words for the former, and 5876 different words for the latter.

Table 2. Basic information of both corpora

Parameters	2010 corpus	2015 corpus
Total number of running words	60673	56842
Total number of different words	5924	5876
Total number of words with at least 52 occurrence	174	167
Highest value of word occurrence	2582	2532
The median of word occurrence of words with more than 52 occurrence	101	98

In order for a word to appear at least once in a sermon, it needs to have word occurrence of at least 52 times. This is because the total number of Friday sermons in a year is 52. The 2010 corpus has 174 words that occur more than 52 times whereas the 2015 corpus

have 167 words. In the 2010 corpus, the word that appears most often occurs 2582 times; whereas in the 2015 corpus, it occurs 2532 times. The median value of word occurrence for the 2010 corpus is 101 and 98 for the 2015 corpus. From these data, it can be seen that the 2010 corpus and the 2015 corpus are almost identical except for the number of running words.

3.2. Word Overlap

The 2015 Friday sermon corpus was compared with the Friday sermon corpus from 2010. It was found that 82% of the words overlaps, meaning that these words appear in both of the text corpora. In the 2010 text corpus, a total of 37 words does not appear in 2015 text corpus; whereas in the 2015 text corpus, 29 words does not appear in the 2010 text corpus. The non-overlapping words from both corpora are shown in Table 3 along with the value of word occurrence.

Table 3. Non-overlapping words from both corpora

2010 corpus					
No.	*Words*	*Occurrence*	*No.*	*Words*	*Occurrence*
1	rahmati	184	20	bermaksud	63
2	puluh	176	21	menjaga	63
3	dua	145	22	tanggung jawab	62
4	harta	105	23	ketakwaan	61
5	ketika	95	24	lima	61
6	alam	93	25	pendidikan	61
7	mengambil	91	26	seseorang	61
8	air	84	27	penuh	59
9	ibu	82	28	tujuh	58
10	bahasa	80	29	berpesan	57
11	malam	72	30	hanya	56
12	berada	69	31	pahala	56
13	berlaku	69	32	cara	53
14	tiga	68	33	dosa	53
15	bumi	66	34	meninggal kan	53
16	riwayat	66	35	sepuluh	53
17	amanah	65	36	tempat	52
18	banyak	64	37	tentang	52
19	bapa	64			
2015 corpus					
No.	*Words*	*Occurrence*	*No.*	*Words*	*Occurrence*
1	Malaysia	97	20	ibadah	55
2	keluarga	79	21	ketiga	55
3	nilai	75	22	beberapa	54
4	akidah	72	23	akhlak	53
5	pelbagai	71	24	jihad	53
6	wahai	68	25	mempunyai	53
7	golongan	65	26	keselamatan	52
8	jemaah	65	27	menjadikan	52
9	pertama	65	28	perpaduan	52
10	sesiapa	65	29	sunnah	52
11	bahkan	64			
12	jalan	64			
13	justeru	63			
14	kedua	63			
15	prinsip	61			
16	rahmat	60			
17	ajaran	59			
18	penting	57			
19	kaum	56			

From Table 3 above, it can be seen that most of the non-overlapping words have word occurrence that is below the median of the corpus. For example, in the 2015 text corpus, the median word occurrence is 98 and from Table 3 it can be seen that all of the non-overlapping words have word occurrence less than 98; whereas for the 2010 text corpus, only 4 of the non-

overlapping words have word occurrence more than the median. From this, it can be seen that non-overlapping words mostly have word frequency less than that of the median of the corpus.

3.2.1 Semantic Analysis

In the year 2015, it can be seen that the Friday sermons in Kuala Lumpur focuses more on the nation as the word “Malaysia” appeared 97 times, “kaum” (races) 56 times, “perpaduan” (unity), and “keselamatan” (safety) appear 52 times. Whereas these words appear less than 52 times in the 2010 Friday sermons.

In the year 2010, the Friday sermon focuses more on the following:

1. The family unit as the word “ibu” (mother) appear 82 times, “bapa” (father) appear 64 times, “tanggungjawab” (responsibility) 62 times.
2. Individual spirituality as the word “pahala” (reward) appear 56 times, “dosa” (sin) appear 53 times, pendidikan (education) 61 times, “seseorang” (individual) 61 times.

From this, it can be seen that the 2015 Friday semons made by JAKIM focuses more on the national identity whereas the 2010 Friday sermons focuses more on the family unit and individual spirituality.

3.2.2. Non-overlapping Words and Relationship with Word Occurrence

From the previous result, it can be seen that most non-overlapping words in both corpus have word occurrence less than the median. However, that does not mean that most words that have low occurrence are non-overlapping words. Table 4 shows the data of both corpora when divided based on their median and the percentage of non-overlapping words.

From Table 4, it can be seen that if a word has occurrence at the median range or above it, it is most probably a word that appear in both corpora. However, if a word has occurrence below that of the median range, the probability of it being a non-overlapping word increases to at least 35%.

Table 4. Relationship between word occurrence and non-overlapping words

Word Occurrence	Above or equal median range		Below median range	
Corpus	2010 corpus	2015 corpus	2010 corpus	2015 corpus
Total number of words	88	84	86	83
Total number of non-overlapping words	4	0	33	29
Percentage of non-overlapping words (%)	5	0	38	35

4. Conclusion

From the literature, it can be seen that the English language has made many word lists according to its discipline. However, in the Malay language, there is currently little progress made within the discipline of corpus linguistics. Furthermore, common tool used by other researchers such as WordSmith Tool can be expensive to researchers in the field of linguistics corpora.

This research has shown that there are certain words that appear consistently in Friday sermons. Although there are differences between the corpora, these differences amount to 18% of the whole corpus. Moreover, this research has shown that the word that does not appear in both corpora have word occurrence less than the median. Although, it should be noted that not all word that have low word occurrence are non-overlapping words.

Acknowledgement

This research is supported by the Fundamental Research Grant Scheme (FRGS) by Universiti Teknologi Malaysia (VOT 4F486) and Ministry of Education (MOE).

References

[1] American National Standard Institute/Acoustical Society of America. ANSI/ASA S3.2009. *Method for Measuring the Intelligibility of Speech over Communication Systems*. New York: Acoustical Society of America; 2009.

[2] International Electrotechnical Commission. IEC 60268-16. *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*. Brussels: European Committee for Electrotechnical Standardization; 2011.

[3] Pollach I. Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis. *Organizational Research Methods.* 2012; 15; 263-287.

[4] Maultner G. Corpora, Critical Discourse Analysis. In: Baker P. *Editor.* Contemporary Corpus Linguistics. Great Britain: Continuum International Publishing Group. 2009; 32-45.

[5] Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. *Plastic Reconstruction Surgery*. 2010; 126; 619-625.

[6] Baker P, Gabrielatos C, Khosravinik M, Kryzanowski M, McEnery T, Wodak R. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourse of refugees and asylum seekers in the UK press. *Discourse and Society*. 2008; 19: 273-306.

[7] Xiao Z, McEnery A. Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics*. 2005; 33: 62-82.

[8] Johnson S, Culpeper J, Bruun H. From 'politically correct councilors' to 'Blairite nonsense': Discourses of 'political correctness in three British newspapers. *Discourse and Society*. 2003; 14; 29-47.

[9] Cohen SJ. Mapping the Minds of Suicide Bombers using Linguistic Methods: The Corpus of Palestinian Suicide Bombers' Farewell Letters (CoPSBFL). *Studies in Conflict & Terrorism*. 2016; 39; 749-780.

[10] House AS, Williams CE, Heckers MHL, Kryter KD. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *The Journal of the Acoustical Society of America*. 1965; 37; 158-166.

[11] West M. A General Service List of English Words. London, Longman. 1953.

[12] Campion ME, Elley WB. An Academic Vocabulary List. Wellington: New Zealand Council for Educational Research. 1971.

[13] Mungra P, Canziani T. Lexicographic studies in medicine: Academic Word List for clinical case histories. *Iberica*. 2013; 25; 39-62.

[14] Wang J, Liang S, Ge G. Establishment of a Medical Word List. *English for Specific Purposes.* 2008; 27; 442-458.

[15] Lei L, Liu D. A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes.* 2016; 22; 42-53.

[16] Mudraya O. Engineering English: A lexical frequency instructional model. *English for Specific Purposes*. 2006; 25; 235-256.

[17] Scott M. *WordSmith Tools version 7* [Computer Software]. (2016 September 9). Liverpool, Great Britain, Stroud: Lexical Analysis Software.

[18] Heatley A, Nation ISP, Coxhead A. *Vocabulary Analysis Programs* [Online]. (2016 September 9). Available: www.vuw.ac.nz/lals/staff/Paul_Nation/.

[19] Python. *Python Programming Language* [Computer Software]. (2016 September 4) Available: www.python.org.

[20] The PHP Group. *PHP programming language* [Computer Software]. (2016 September 1). Available: php.net.

[21] MySQL AB. *MySQL Database* [Computer Software]. (2016 September 9). Available: www.mysql.com.

[22] Jabatan Kemajuan Islam Malaysia (JAKIM). *Senarai Khutbah Jumaat* [Online]. (2016 September 7). Available: e-muamalat.gov.my/khutbah-online

[23] Mokhtar H, Muhammad AAA, Siti ZAH, Fareha AR, Puspa IK. Determination of Bahasa Melayu word list from Friday sermon transcripts using PHP and MySQL. *Jurnal Teknologi*. 2013; 64; 1-6.