ISSN 2286-976X / Online: ISSN 2539-5513

RANGSIT JOURNAL OF SOCIAL SCIENCES AND HUMANITIES

Available online at https://rjsh.rsu.ac.th

Understanding the Pali Canon through Keyword Analysis: A Comparison between Different Reference Corpora

Chirawan Sukwitthayakul and Saneh Thongrin*

Faculty of Liberal Arts, Thammasat University, Bangkok, Thailand
*Corresponding author, Email: saneh.t@arts.tu.ac.th

Received September 23, 2021 / Revised March 15, 2022 / Accepted April 18, 2022 / Publish Online November 30, 2022

Abstract

This study was to conduct keyword analyses on the English Pali Canon and compare keywords generated by four different reference corpora which varied in genre and size. The software *AntConc* 3.5.9 was employed for analyzing and generating the keyword lists. Two node corpora were compiled using samples from two English translations of the Pali Canon and the reference corpora were the node corpora themselves, a collection of other religions' canons, the Manually Annotated Sub-Corpus (MASC), and the British National Corpus (BNC). It was found that the numbers of keywords were the highest when BNC, which was the largest and more general, was used as the reference corpus. The results were compared to select the keywords that recurred at the top of most, if not all, keyword lists. It can be seen that, regardless of the reference corpora, most of the top keywords were nouns referring to people or characters in the Pali Canon, such as *the Exalted One*, *brethren*, *Gotama*, and *Ânanda* as these names and words were not frequently found in other texts. The comparison of reference corpora can help researchers find the most appropriate reference corpus and ensure the selection of keywords in the creation of a Buddhist wordlist for further research.

Keywords*: Keyword Analysis, the Pali Canon, Corpus Analysis, English for Specific Purposes, Digha Nikaya*

1. Introduction

The roles of language used in religious contexts have not received much attention by researchers, especially when compared with language use in academic and occupational contexts (Alsaawi, 2020). Within the profoundly explored area of English for specific purposes (ESP), there are only a small number of research articles on English for religious purposes (ERP). Moreover, studies in ERP usually explore language use in Christian or Islamic contexts while there are only a few studies, such as Liu's (2007) which is considered the first ERP research (Supphipat, 2017), investigating language use in a Buddhist context. Even though it seems that the 'specific purposes' of ESP are overwhelmed with academic and occupational purposes, ESP concepts and tools can also benefit the investigation of the English language for other purposes like spiritual development. In ESP exploration, one of the common tools is the use of corpus to facilitate researchers in finding the most frequent words and lexical bundles, exploring collocations, and identifying keywords of language used in a particular context. With the help of corpus tools, a corpus is analyzed and lists of words, as well as bundles, are generated as required. While researchers have been widely interested in academic and occupational vocabulary, only a small number of studies have employed corpus tools to explore words in religious contexts.

English for Buddhism has been "a completely new branch of ESP" (Liu, 2007). Buddhist language is unique on lexical, phrasal, and discoursal levels (Supphipat, & Chinokul, 2018), and the lexical level can be investigated by using ESP concepts and corpus tools. To explore words used in Buddhist texts, we conducted keyword analysis, focusing on the written language in the Pali Canon, the oldest collection of the Buddha's teachings. The results of keyword analysis on the English Pali Canon can be customized for future use, such as a development of Buddhist materials or a translation of Buddhist texts.

To conduct a keyword analysis, the studied corpus needs to be compared with another collection of texts called a reference corpus to generate a keyword list. Previous studies employing keyword analysis often use one reference to generate keywords. In this study, multiple reference corpora were used for comparing

keywords from different keyword lists and identifying the most distinct, recurring keywords with high keyness values regardless of reference corpora. The comparison of results was also carried out for the selection of an appropriate reference corpus for further study. Accordingly, the objectives of this study were to find and compare keywords in the English Pali Canon retrieved from different reference corpora.

The study used a corpus linguistic tool to bring religious language into the interest of ESP researchers as there has been an obvious lack of studies on the use of religious language from a linguistic point of view. As mentioned earlier that this paper focuses on keyword analysis of the Pali Canon, below is the review of religious teaching, the Pali Canon, and the keyword analysis method, with the inclusion of previous related studies.

1.1 Religious Teaching

The definition of religion has not been agreed upon (Greil, 2009), but its impacts on our global and local communities as well as individuals are evident. Despite their positive effects (Roberts, 2019), religious beliefs and teachings have been claimed to declare conflicts and wars (Sherwood, 2018), which are often claimed to be the results of misinterpretation of religious teachings. These inaccuracies may stem from the extraordinariness of language used in religious contexts in order to create sacredness (Sawyer, 2001) and the vocabulary used for religious purposes is unique (Dazdarevic, 2012), making religious language different from everyday language.

For many religions, the core teachings are generally collected in scriptures and other related texts. Factors, such as geography and time of origination of religion, distinguish these sacred texts (Sah & Fokoué, 2019). Religious teachings are passed on from generation to generation through these texts. For Buddhism, especially Theravada Buddhism which is one of the three main branches of the religion, it is believed that the Buddha's teachings are collected in the Pali Canon. Hence, the Pali Canon will be used as the source of this study to identify Buddhist keywords.

1.2 The Teachings in the Pali Canon

The Pali Canon (or Tipitaka in Pali and Tripitaka in Sanskrit) is believed to be the oldest collection of Buddha's teachings. The literal meaning of Tipitaka or Tripitaka is 'the Three Baskets' in which the Pali Canon is divided into three sections: Vinaya, Sutta, and Abhidhamma. The first section contains monastic rules and is primarily for monks and Buddhist nuns. The second section, Sutta, which is the focus of the study contains narrative stories and teachings on various occasions when the Buddha was alive. The third section is the special doctrines for the more experienced. Generally, Sutta is recommended for all kinds of readers to study because it is neither too advanced nor too specific for a particular group of Buddhists; therefore, this study would like to begin with the part of the Pali Canon that is readable and understandable for various groups of readers. Sutta is further divided into five sub-sections, covering over 10,000 teachings. For feasibility, this trial initially focuses on the first sub-section of Sutta called Digha Nikaya which contains 34 long discourses.

1.3 Role of Corpus

A corpus is a collection of machine-readable written texts or transcriptions (Crystal, 1992) which occur naturally (Sinclair, 1991). Results from corpus analysis, especially the generated lists of lexical items (McEnery, & Wilson, 1996) are useful for language pedagogy. For direct use of corpus, learners and teachers may perform an active investigation of the corpus by themselves while researchers and materials writers may adopt the indirect use of corpus by exploring and selecting useful information for learners (Campoy, Cubillo, Belles-Fortuno, & Gea-Valor, 2010). In ESP, one of the well-known systems of analyses of corpora is to generate a frequency list of which the units of analyses can be individual words, lexical bundles, clusters, and keywords. Responding to the objectives of this study, keyword analysis is performed to generate Buddhist keywords that are significant in the Pali Canon, but not frequently found in other texts.

Keywords are words that occur more frequently with significance in the sample text in comparison with texts in a reference corpus (Stubbs, 2010). An important concept related to keywords is keyness, 'a quality which is generally intuitively obvious' (Scott, & Tribble, 2006). Keywords mark the aboutness and style of texts in the studied corpus (Scott, & Tribble, 2006) and present "a rapid and replicable overview of

the characteristic themes" as well as the characteristic of each text (Brookes, & McEnery, 2019). Moreover, keyword analysis can benefit various fields of study, such as language teaching, content analysis, and stylistics.

Keywords can be elicited through the use of a concordancing tool, such as *AntConc* and *WordSmith*. Up until the keyword list is generated, the process is quantitative and objective. Depending on many factors, a list of keywords can be very long and unmanageable. Researchers (Scott, & Tribble, 2006; Grabowski, 2015) often get involved with the list and decide on the cut-off point as well as making some criteria to shorten and customize the lists to meet their needs and their study's purpose. These decisions affect keyword results and make the method more subjective.

To generate a keyword list, a corpus of texts that are to be investigated must be compiled. This is called a target corpus, a study corpus, or a node corpus. Then a reference corpus is required for the node corpus to be compared. There are four key factors of a reference corpus that are likely to affect keyword results and these factors are size, genre, historical period, and varietal differences (Abeed, 2017). Researchers (Scott, 2009) have acknowledged differences in results of keyword analysis when using different reference corpora, yet it seems the only requirement widely accepted for a reference corpus is that its size should be larger than the node corpus (Goh, 2011). McEnery, Xiao, and Tono (2006) did not place the size of a corpus as a significant factor for keyword analysis. This is evident in Goh's (2011) study testing the effects of these four factors and it was found that genre and diachrony of a reference corpus made the most statistically significant impacts on keyword results.

Keywords of two-node corpora from three different reference corpora with different sub-registers were compared in Geluso and Hirch's (2019) study. They reported that a reference corpus with the same sub-register of the node corpus could elicit the distinct content of the node corpus and keywords were clustered in a small number of texts in the corpus. On the other hand, using a reference corpus with a different sub-register could show shared content between the node corpora. They also stated that although the register of a reference corpus affected the results of keyword analysis, using an adequate-size reference corpus of any register could eventually generate the core keywords of the node corpus. The selection of reference corpus was also emphasized in Maiwald's (2011) study on the corpus stylistics of George MacDonald's fiction. The keywords with the highest keyness values in the fiction were proper nouns. As different reference corpora were used in his study, Maiwald (2011) stressed that the reference corpus had an effect on keywords' semantic domains.

Among previous analyses of keywords, only a small number of studies have paid attention to religious language. One of them is Lien (2022) who investigated keywords in the Buddhist corpus of 20 million words, consisting of different types of Buddhism-related texts. Overall, there were 1,244 keywords in the wordlist retrieved through *WordSmith*. Lien (2022) recommended the application of multilevel methods in performing keyword analyses to create specialized field wordlists. To do so, she used frequency, log-likelihood, which is a probability statistic indicating the confidence that a word is key, and odds ratio, which is an effect size statistic denoting the association of word frequency in the node corpus and the reference corpus. Her study presented the distinct words in general Buddhist contexts and most of the words were different from the top keywords in the Pali Canon found in this study.

2. Objectives

1) To find keywords in the samples of the English Pali Canon
2) To compare keywords of the English Pali Canon generated by four different reference corpora

3. Materials and Method

The following section explains the compilation of corpora and the process of keyword analysis for this study. Two node corpora consisting of samples from the Pali Canon were needed to find and compare Buddhist keywords from the Pali Canon and three reference corpora, varying in size and genre, were used to generate different keyword lists for comparison. Also, *AntConc* Version 3.5.9 (Anthony, 2020) with its default setting for keyword analysis was used as the concordancing software for these analyses.

3.1 The Node Corpora – Samples from the Pali Canon

As the study investigated keywords from the Pali Canon, specialized corpora containing texts from the English Pali Canon were compiled. Originally, the Pali Canon is written in the Pali language. It has been translated into many languages, including English. Two versions of the English Pali Canon were selected to be included in the corpora. The first English translation of the Pali Canon was published by The Pali Text Society and translated by various translators from the 19th century to the 20th century. The other version was translated by Maurice Walshe who was a vice president of the Buddhist Society. These two versions of the English translation are complete English translations of the Pali Canon from the Pali language. Using both translations helped in finding similarities or what was retained in the texts as the cores, regardless of influencing factors, such as time and translators as well as showing differences between the translations or elements that changed due to factors, such as language styles and translators' interpretation.

Since there were two versions of translation, each version was compiled as its own node corpus and the results of keyword analysis were later compared. The texts in the compiled node corpora were from these four books:

(1) *Dialogues of the Buddha Part I* (1899) translated by T. W. Rhys Davids

(2) *Dialogues of the Buddha Part II* (1910) translated by T. W. and C. A. F. Rhys Davids

(3) *Dialogues of the Buddha Part III* (1921) translated by T. W. and C. A. F. Rhys Davids

(4) *The Long Discourses of the Buddha: A Translation of the Digha Nikaya* (1995) translated by Maurice Walshe

The first three books were parts of the first complete English translation of the Pali Canon by the Pali Text Society and the last book was a complete English translation of teachings in Digha Nikaya of the Pali Canon. The collection *Dialogues of the Buddha* consisting of three books were compiled in the first node corpus of this study and the whole teachings in the fourth book were collected for the second node corpus. Since both versions were translated from the Pali Canon in the Pali language, the numbers of the teachings were equal: 34 discourses, but only 33 discourses were used for both corpora because one discourse merely guides readers to read a previous discourse that did not contribute much to the content of the teachings and the analysis.

Regarding the numbers of words in the corpora, the first corpus contained 210,743 words and the second corpus contained 146,964 words. In order to conduct a keyword analysis, at least one reference corpus is required. The reference corpora used in this study are described below.

3.2 The Reference Corpora

A reference corpus is needed in a keyword analysis because a node corpus must be compared with a reference in order to identify salient keywords and generate a keyword list. For this study, the node corpora were compared with four reference corpora and each node corpus was also used as a reference corpus for the other. Therefore, the details of the first reference corpora are as explained in the previous section. The other three reference corpora were for both node corpora. These reference corpora were comprised of a corpus of other religions' canons: the Bible and the Koran (totaling 3,100,000 words), MASC or Manually Annotated Sub-Corpus (500,000 words) which consisted of transcribed speech and written texts from the Open American National Corpus, and lastly, the well-known BNC (100,000,000 words). As Scott (2010) suggested that the larger the reference corpus the better, the number of words in the reference corpora, except when the second node corpus was used as a reference corpus of the first node corpus, which was all higher than the node corpora.

Apart from their size, the reference corpora were various in terms of genre. The first reference corpus of each node corpus was the texts with the same source and content. The difference between each node corpus and its first reference corpus was that they were different translations of the Pali Canon from the Pali language by different translators in different periods. Basically, they were different versions of the same texts. The second reference corpus represented the texts within the same genre; that is, it was a collection of canons of major religions: Christianity and Islam. The third and fourth reference corpora were more general. The third reference corpus contained texts in American English while the fourth was British. In terms of size, it could be said that the first reference corpus shared the same size as the node corpus. In contradiction to the norm

that the size of the reference corpus should be larger than the node corpus, the first reference corpus of the first node corpus was slightly smaller and, of course, the first reference corpus of the second node corpus was slightly larger. The second reference corpus was about 14 times larger than the node corpora. The third was approximately 2 times larger and the fourth was 475 times larger than the node corpora.

Table 1 Corpora's sizes

Corpus	Number of words	Corpus	Number of words
Node corpus #1		**Reference corpora for node corpus #1**	
Dialogues of the Buddha	210,743	The Long Discourses	146,964
		Other religions' canons	3,100,000
		MASC	500,000
		BNC	100,000,000
Node corpus #2		**Reference corpora for node corpus #2**	
The Long Discourses	146,964	Dialogues of the Buddha	210,743
		Other religions' canons	3,100,000
		MASC	500,000
		BNC	100,000,000

3.3 Keyword Analysis

After the corpora were compiled, *AntConc* 3.5.9 (Anthony, 2020) was used to generate the keyword lists with $p < 0.05$ and the default setting. Each node corpus was loaded and compared with each reference corpus. In total, there were eight keyword lists. Both content words, such as *king*, *lord*, and *world*, and function words, such as *or*, *has*, and *does*, were all included in the lists, as well as non-words. Pali words and proper nouns were also kept in the lists. This was because the keywords were later selected manually. The selection was subjective, yet the total number of keywords retrieved through the set of analyses was over 10,000 words, which were not feasible for the study. As the number of texts included in the corpus, the reference corpus, and the selection of top words could affect the keyword list (Pojanapunya & Lieungnapar, 2017), it was decided that selected keywords were those in the top 20 of at least three lists, so they were distinctly significant despite the change of reference corpus.

4. Results

Before presenting the results, the nomenclature for each keyword list should be clarified for mutual understanding. To make it convenient for this paper, the lists are mentioned by the names of the reference corpora. That is, when the reference corpus is BNC, the list is referred to as the BNC list. Hence, the Other Canons list show results from the analysis of which the reference corpus is the canons of other religions and the MASC list presents keywords retrieved when the reference corpus is MASC. When the node corpora become the reference corpora, they are called by the names of the books: Dialogues of the Buddha list and The Long Discourses list.

Overall, when comparing the first node corpus with four reference corpora, in the Long Discourses list, there were only 211 keywords, which was much fewer than 1,892 keywords in the Other Canons list, 1,895 in the MASC list, and 3,200 in the BNC list. Due to limited space, it is impossible to present all the keywords in all the lists here. Below in Table 2 are the top 20 keywords in the first node corpus as compared with each reference corpus. Although there are only 20 keywords in each list, it can be seen that some keywords appear in every list and the first three lists share many keywords.

For the second node corpus containing the other version of the English Pali Canon, there were 223 keywords when compared with the first node corpus, 1,510 in the Other Canons list, 1,491 in the MASC list, and 2,512 in the BNC list. The top 20 keywords of each list for this second corpus are presented in Table 3. Many words recur in every list, especially the first three lists. From these two tables, it can be seen that the BNC lists contain more general words, such as *the*, *and*, *in*, and *he*, than the others. This could be a result of the size of the BNC corpus which was the largest, making the BNC keyword lists the longest.

Table 2 Top 20 keywords of the 1st node corpus (Dialogues of the Buddha) VS different reference corpora

Rank	Reference corpus			
	The Long Discourses	**Other canons**	**MASC**	**BNC**
1	Exalted	Exalted	Exalted	the
2	Ânanda	Ânanda	Ânanda	and
3	brethren	one	brethren	of
4	o	Gotama	he	Exalted
5	wit	or	lord	is
6	one	such	Gotama	Ânanda
7	of	venerable	king	brethren
8	brother	has	one	one
9	Brahmans	Brahmans	thus	to
10	Brahman	Brahman	gods	Gotama
11	Brahmâ	wit	him	that
12	Tathâgata	state	venerable	he
13	thou	Brahmâ	wit	his
14	recluse	Tathâgata	Brahmans	in
15	even	Buddha	Brahman	thus
16	ye	thus	Brahmâ	or
17	soul	sir	Tathâgata	him
18	norm	does	and	this
19	Bhikkhus	so	his	so
20	doth	brethren	nor	king

Table 3 Top 20 keywords of the 2nd node corpus (The Long Discourses) VS different reference corpora

Rank	Reference corpus			
	Dialogues of the Buddha	**Other canons**	**MASC**	**BNC**
1	monks	monks	lord	and
2	lord	Dhamma	monks	the
3	Dhamma	Gotama	Dhamma	lord
4	monk	reverend	Gotama	is
5	Ānanda	Brahmins	he	of
6	reverend	monk	reverend	Gotama
7	verse	Ānanda	Brahmins	monks
8	ascetics	ascetic	Ānanda	Dhamma
9	Sutta	Buddha	ascetic	Brahmins
10	Vipassī	Tathāgata	monk	Ānanda
11	Brahmā	Brahmin	Tathāgata	he
12	Tathāgata	mind	Brahmin	his
13	ascetic	devas	devas	Tathāgata
14	Brahmins	self	Buddha	this
15	Ambaṭṭha	Sutta	mind	ascetic
16	devas	Brahmā	Sutta	reverend
17	contemplating	verse	Brahmā	are
18	s	such	blessed	to
19	Nibbāna	or	verse	Buddha
20	ca	ascetics	ascetics	Brahmin

For further application of the results, it has to be decided if the lists should be cleaned, such as by eliminating function words, or other criteria should be set, depending on a study's objective. For this paper, which was aimed at identifying Buddhist keywords and comparing the results of keyword analyses in order to choose the appropriate reference corpus for future use, both content words and function words were kept to explore the overall results of different reference corpora. Some recurring words in the lists which are Pali and proper nouns, such as *Ānanda*, should not be eliminated because these terms and names are significant and common in Buddhist contexts.

From the numbers of keywords compared above, it is noticeable that using a reference corpus that share the size and content of the node corpus yield considerably lower numbers of keywords than the others. Further use of this kind of reference corpus may result in missing significant keywords in the texts, but can be beneficial in identifying differences between different translations of the same text and highlighting the distinction of each text. For the second and third reference corpus, the number of generated keywords in the lists were almost equal for both node corpora, although the sizes and genres of the corpora were not the same. It seemed that the canons of other religions in the second reference corpus and general texts in American English in the third yielded quite similar results in terms of numbers of keywords and top keywords. The last reference corpus, which was the largest, generated the highest number of keywords; however, even though the size of the BNC corpus was about 475 times larger than the node corpora, the number of keywords was less than double the number of keywords from the second and the third reference corpus.

5. Discussion

The selection of keywords was clearly subjective but necessary since all generated keywords could not be presented here. The keywords were selected based on their recurrence in the keyword lists as well as their keyness value. Selected keywords, ten words from each node corpus, are shown in Table 4 and Table 5 with their frequency, keyness value, and rank in each list. Some words are not key in a particular list, but all of them are at the top of at least three lists.

Table 4 Selected keywords from the 1st node corpus, Dialogues of the Buddha

keywords	frequency	keyness (1st reference corpus)	keyness (2nd reference corpus)	keyness (3rd reference corpus)	keyness (4th reference corpus)
Exalted	997	1042.55 (#1)	4133.18 (#1)	8301.05 (#1)	15313.03 (#4)
Ânanda	669	708.55 (#2)	3687.49 (#2)	5700.73 (#2)	11300.63 (#6)
brethren	636	673.56 (#3)	854.41 (#20)	5262.32 (#3)	9953.99 (#7)
one	2248	295.89 (#6)	2522.66 (#3)	2639.48 (#8)	8466.6 (#8)
wit	291	296.42 (#5)	1164.28 (#11)	1857.76 (#13)	3165.65 (#29)
Brahmans	216	228.58 (#9)	1190.14 (#9)	1808.52 (#14)	3593.01 (#24)
Tathâgata	180	190.47 (#12)	991.76 (#14)	1533.42 (#17)	3040.11 (#31)
Gotama	386	-----	2127.12 (#4)	3288.7 (#6)	6494.67 (#10)
venerable	260	-----	1361.74 (#7)	1857.92 (#12)	3537.47 (#25)

Table 5 Selected keywords from the 2nd node corpus, The Long Discourses

keywords	frequency	keyness (1st reference corpus)	keyness (2nd reference corpus)	keyness (3rd reference corpus)	keyness (4th reference corpus)
monks	431	767.73 (#1)	2624.75 (#1)	3435.73 (#2)	6013.85 (#7)
lord	1383	719.43 (#2)	-	10782.01 (#1)	12835 (#3)
Dhamma	347	508.94 (#3)	2149.32 (#2)	3204.07 (#3)	5713.22 (#8)
reverend	292	388.12 (#5)	1784.79 (#4)	2542.5 (#6)	3660.92 (#16)
Brahmins	276	1337.72 (#11)	156.84 (#14)	1946.6 (#12)	4688.68 (#9)

keywords	frequency	keyness (1st reference corpus)	keyness (2nd reference corpus)	keyness (3rd reference corpus)	keyness (4th reference corpus)
Ānanda	255	416.97 (#5)	1579.31 (#7)	2354.42 (#8)	4467.65 (#10)
Sutta	183	199.41 (#9)	1133.31 (#15)	1667.5 (#16)	3200.82 (#23)
Tathāgata	234	168.26 (#12)	1449.23 (#10)	2160.49 (#11)	4121.17 (#13
Gotama	346	-	2143.13 (#3)	3194.83 (#4)	6069.33 (#6)

In Table 4, the first two keywords *Exalted* and *Ânanda* from the first node corpus, Dialogues of the Buddha, are in the same ranks (the first and the second) when compared with three reference corpora, namely, The Long Discourses, Other Canons, and MASC. Even though *Exalted* and *Ânanda* come as the fourth and the sixth in the BNC list, it is worth noting that the three keywords before them are all function words: *the*, *and*, and *of*. This makes *Exalted* and *Ânanda* the first and second content keywords in the BNC list. Apart from that, the keyness values of these words in the BNC list are the highest among the four lists.

If we look at this part of the Pali Canon as a collection of narrative stories, it is not surprising that the two keywords are at the top of the list because both words refer to the protagonists of the stories, appearing in almost every scene and they, of course, do not exist in the reference texts. As keywords point to the 'aboutness' of the texts, it can be said that this sub-section of the Pali Canon called Digha Nikaya involves people to whom these two keywords refer.

The word *Exalted* which appears 997 times in the first node corpus is actually found as *the Exalted One* for 925 times. The bundle is used to refer to the Buddha and it is not found in the other corpus which prefers the word, *lord* when referring to the Buddha. Because *Exalted* usually co-occurs with *the* and *One*, this brings *One* to one of the top 10 keywords in every list too. The 2,248 occurrences of *One* in the corpus doubles the occurrences of *Exalted* because *One* is also used as a counting number in the texts. In this context, *Ânanda* is the name of the Buddha's attendant, so his name is mentioned in the Pali Canon repeatedly. Referring to the same person, the spelling of the name is different in the two corpora: *Ânanda* and *Ānanda*. The only difference at the first letter makes them keywords even though the node corpora were used as a reference corpus for each other.

As for *brethren*, it is ranked as the seventh keyword in the BNC list, but when compared with the canons of other religions, *brethren* is not very distinct. The word is probably used frequently in other religions' texts and is common in religious contexts. In the first corpus, *brethren* is used by the Buddha to call the monks listening to his teachings. In the other node corpus, *monks* is used instead of *brethren* since the word *brethren* does not appear in the corpus which also makes *monks* the most distinct keyword of the second corpus when using Dialogues of the Buddha as the reference corpus.

Another selected keyword is *Brahmans*. In fact, its singular form, *Brahman*, is also a keyword in these lists. In the second node corpus, *Brahmins* is used instead, yet it is still among the top keywords. In this context, Brahmans or Brahmins are people in the highest ranking of the Hindu four classes. During the time the Buddha had lived and preached, Hinduism has already flourished. Hence, many narratives in the Pali Canon have involved stories about Brahmans. Some concepts and practices in both religions are similar and many Buddhist terms seem to be borrowed from existing terms in Hinduism. Other keywords in the node corpora that reflect the relationship between these two religions include *jhāna* (roughly translated as meditation), and *Nibbhāna* (nirvana) which are concepts existing in both religions.

Apart from *the Exalted One*, there are other keywords directly referring to the Buddha. The words are *Gotama* and *Tathâgata* (*Tathāgata* in the second corpus). *Gotama* is the Buddha's name and it certainly appears in both node corpora with the same spelling (386 times in Dialogues of the Buddha and 346 times in The Long discourses). Obviously, with similar frequencies, it was not a keyword when the node corpora were compared with each other, but when compared with other reference corpora, its keyness values were high. Both *Tathâgata* and *Tathāgata* are used by the Buddha as first-person pronouns in the Pali Canon. The first node corpus uses the first spelling and the second node corpus uses the other. The different spelling makes them keywords for both corpora.

Next is the keyword *wit* in the first node corpus. *Wit* which is among the top keywords generally follows the word 'to' as *to wit*, and in this context, it is used when the Buddha or a preacher would like to clarify or specify things. The examples below from the 17th discourse and the 33rd discourse of the Pali Canon show how *to wit* is used in the corpus.

> *"Suppose, now, I were to establish a perpetual grant by the banks of those Lotus-ponds—to wit, food for the hungry, drink for the thirsty, raiment for the naked, means of conveyance for those who have need of it, couches for the tired, wives for those who want wives, gold for the poor, and money for those who are in want."*
>
> *(DN-17 Mahā-Sudassana Sutta – The Great King Of Glory)*

> *"Eight wrong factors of character and conduct, to wit, wrong views, intention, speech, action, livelihood, effort, mindfulness, concentration."*
>
> (*DN-33 Sangiti Sutta*)

In the first corpus, *venerable* is usually used as an adjective to describe people. Its occurrences in the corpus are 260, but for 107 times, it is used as *Venerable Ânanda*. For another 53 times, the word is used to precede the keywords *Gotama*. *Venerable Ânanda* is also found in the second node corpus (42 times). In this second corpus, the frequency is lesser than in the first corpus as *Reverend Ânanda* is also used. However, *Reverend Ânanda* is not found in the first corpus and this makes *reverend* a keyword in the second corpus with high keyness values in every list.

From Table 5, in the second node corpus, *Sutta* is another keyword found in every keyword list. *Sutta* is the name of the second section of the Pali Canon itself and it is used to call each discourse in the section too. The word *Sutta* is also found in the first corpus, but the frequency is as low as 25 while it appears 183 times in the second corpus. Most of the time it appears in the first corpus, it is used as the name of each discourse. Likewise, in the second corpus, 33 occurrences are in the titles of the discourses, but the word is also used as a reference in the Pali Canon. For 150 times, it is used to guide readers to read other discourses. Therefore, when looking at the concordances of *Sutta* in the corpus, it does not contribute much to the content of the discourses. All 150 occurrences appear in parentheses (as Sutta 1 verse 1.9) which suggests readers go to the first Sutta (Sutta 1) and read verse 1.9 for more information or clarification. Therefore, this word has only one function which is signifying that the discourses are Suttas in the titles and it appears in two places in the corpus, that is, as a title of each discourse and as a reference.

As can be seen from the top 20 keywords of each list and selected keywords, it seems that nouns referring to people are one of the common types of keywords in both node corpora. These keywords include *brethren, Gotama, Ânanda, Ānanda, Tathâgata, Tathāgata, monks, lord, Brahmans*, and *Brahmins*. Even *Exalted*, the most distinct keyword in the three lists, is actually used with a noun as *the Exalted One* to refer to the Buddha. The prevalence of these nouns at the top of keyword lists is in accordance with Maiwald's (2011) study of George MacDonald's fiction which found that keywords with the highest keyness values were proper nouns. As the samples in both corpora are the Buddha's teachings collected in the form of narratives, the recorders of these teachings need characters who begin, continue, and end the stories like fiction. Consequently, nouns referring to people are significant and they occur more frequently and saliently in this part of the Pali Canon. Most of them even appear in the texts more frequently than familiar Buddhist terms and concepts like *Nibbāna* (nirvana) and *Mindfulness*.

Sometimes, proper nouns are excluded from a vocabulary list as they might not be significant in the studied text or the field that the list is created for. This can be true for academic wordlists. However, for the Pali Canon, we did not cut proper nouns off because the names of people and places in the Pali Canon are generally mentioned when listening to Dhamma or reading Buddhist texts. Therefore, the creation of a Buddhist wordlist should not omit proper nouns if they are significant and common in context because each name signifies its own story and these words are worth knowing for a Buddhist.

Methodologically speaking, it is suggested that for keyword analysis, the larger the reference corpus the better. A reference corpus should have a higher number of tokens than the node corpus. From Table 5, when the reference corpus is The Long Discourses which had a lower number of tokens than the node corpus (Dialogues of the Buddha), it can be seen that the top keywords are similar to the lists of other reference

corpora. The top three keywords of the list are the same as the top three keywords of the MASC list. However, the numbers of keywords when the node corpora become the reference corpus for each other are distinctly lower than those of the others' lists. Using the same text with different versions of translation as the reference corpus can highlight differences in styles of translation and publication as the results show distinct words frequently used in the node corpus, compared with those in the reference corpus. However, such a reference corpus might not provide various groups of keywords and the keywords may not have high keyness values.

On the other hand, a larger and more general reference corpus like BNC can generate a higher number of keywords and the keywords in the BNC lists are more variable than those in the others, but the generated result is not feasible and many keywords in the list are function words which might not contribute much to the main substance of the text. Then, criteria should be set to scope and select certain keywords to make the long list of keywords manageable as well as to meet a study's purpose. It is certain that a reference corpus should be larger than its node corpus in order to generate a list of significant keywords, but other factors, such as genre, also affect the results. Therefore, the trial of reference corpora is an important step before performing a keyword analysis to ensure the selection of an appropriate reference corpus that responds to the study's purpose. For example, to create a Buddhist wordlist, we may not need everyday words like *he*, but include *Exalted* if it is evident that it is significant.

Although each reference corpus yields different results from keyword analysis, the application of results is in the hands of users who apply the lists and words generated from the machine. Researchers may use keyword analysis with texts that are manually unmanageable to scope the data before focusing on specific details for further analysis. For language pedagogy, a wide range of applications of corpus and keyword analysis can be adopted by learners, teachers, and materials developers for learning language use in a particular context. ESP practitioners may especially be apt to include the results from such analysis in their ESP teaching process since these keywords can be interpreted as significant words that are not frequently found in other texts. It is applicable, for example, for materials developers who may create supplementary material for learning significant words in a Buddhist context or a glossary of Buddhist words explaining common vocabulary from the authentic sources. Also, keywords can be a starting point for materials writers for Buddhist studies to elaborate on a particular key topic evident from the analysis. A keyword list like a Buddhist wordlist might not be of extensive use in a general classroom, but it can be applied to benefit other groups of people, such as missionary monks and people interested in the religion who have to learn words in a Buddhist context for preservation and a better understanding of Buddhism.

6. Conclusion

It is known that using different reference corpora in keyword analysis yields different results, but in this study, many top keywords recurred even when a reference corpus was changed. However, the number of keywords generated from different reference corpora varied. The numbers were the smallest when the reference corpus' size was smaller or similar to the size of the node corpus and the number of generated keywords was higher when using larger reference corpora. The reference corpus that was approximately 3-20 times larger than the node corpus yielded quite the same results. Besides, by comparing the top keywords from reference corpora with different genres, it was found that content words generated from these corpora were quite the same. Choosing any reference corpus seems to serve the study's purpose if we only use the top content keywords for further analysis. The top keywords in these lists were mostly either character names or nouns referring to human beings. In this part of the Pali Canon where teachings are collected in the form of narrative stories, characters are salient as the narrators need them to develop the stories. Nevertheless, after the trial of different reference corpora, a more refined selection of keywords and the analysis of the other parts of the Pali Canon are definitely needed in order to create the complete Buddhist wordlist from the English Pali Canon to support the learning of the religion as well as to investigate the ideology, value, and theme of the scripture.

7. Acknowledgements

We would like to express our gratitude to all those who have helped with the completion of this study. Our deep gratitude goes to the English Language Studies program of the Faculty of Liberal Arts, Thammasat University, which offers us opportunities for academic exploration. We are particularly grateful

to Professor Dr. Watchara Ngamchitcharoen, Asst. Prof. Dr. Passapong Sripicharn, Assoc. Prof. Dr. Apisak Pupipat, and Asst. Prof. Dr. Lugsamee Nuamthanom Kimura, whose constructive comments were important for our journey. Our research would not have been completed without these precious hands.

8. References

Abeed, M. (2017). *News representation in times of conflict: A corpus-based critical stylistic analysis of the Libyan revolution* (Doctoral dissertation). University of Huddersfield Repository, UK.

Alsaawi, A. (2022). The use of language and religion from a sociolinguistic perspective. *Journal of Asian Pacific Communication*, *32*(2), 236-253. https://doi.org/10.1075/japc.00039.als

Anthony, L. (2020). *AntConc* (Version 3.5.9). Retrived form https://www.laurenceanthony.net/software

Brookes, G., & McEnery, A. (2019). Corpus linguistics for indexing. *The Indexer: The International Journal of Indexing*, *37*(2), 105-124. https://doi.org/10.3828/indexer.2019.16

Campoy, M. C., Cubillo, M. C. C., Belles-Fortuno, B., & Gea-Valor, M. L. (2010). *Corpus-based approaches to English language teaching*. London, UK: Continuum.

Crystal, D. (1992). *An encyclopedic dictionary of language and languages*. Oxford, UK: Blackwell.

Dazdarevic, S. (2012). English for religious purposes. *Teaching foreign languages for special purposes*, University of Foreigners of Perugia, Perugia, Italy.

Geluso, J., & Hirch, R. (2019). The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. *Computer Science*, *1*(2), 209-242. https://doi.org/10.1075/rs.18001.gel

Goh, G. Y. (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research*, *28*(1), 239-256. https://doi.org/10.17250/khisli.28.1.201104.013

Grabowski, Ł. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, *38*, 23-33. http://doi.org/10.1016/j.esp.2014.10.004

Greil, A. L. (2009). Art: Defining religion. In P. Clarke & P. Beyer (Eds.), *The world's religions* (pp. 135-149). London, UK: Routledge.

Lien, H. Y. (2022). Revisiting Keyword Analysis in a Specialized Corpus: Religious Terminology Extraction. *Journal of Quantitative Linguistics*, *29*(3), 269-282. https://doi.org/10.1080/09296174.2020.1865668

Liu, C. (2007). *A descriptive study of how English is used and learned linguistically and culturally in a Taiwanese Buddhist monastery in Los Angeles* (Doctoral dissertation), The University of Texas at Austin.US.

Maiwald, P. (2011). Exploring a Corpus of George MacDonald's Fiction. *North Wind: A Journal of George MacDonald Studies*, *30*(1), 5.

McEnery, T., &Wilson, A. (1996). *Corpus linguistics.* Edinburgh, UK: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* London, UK: Routledge.

Pojanapunya, P., & Lieungnapar, A. (2017). The quality of choices determines the quantity of key words. *Proceedings of the International Conference: DRAL 3/19th ESEA 2017.* King Mongkut's University of Technology Thonburi, Bangkok, Thailand.

Roberts, N. F. (2019). *Science says: Religion is good for your health*. Retrieved form https://www.forbes.com/sites/nicolefisher/2019/03/29/science-says-religion-is-good-for-your-health/?sh=446fd3ee3a12

Sah, P., & Fokoué, E. (2019). What do Asian religions have in common? An unsupervised text analytics exploration. *ArXiv:1912.10847.* https://doi.org/10.48550/arXiv.1912.10847

Sawyer, J. F. A. (2001). Special language uses. In J. F. A. Sawyer & J. M. Y. Simpson (Eds.), *Concise encyclopedia of language and religion* (pp. 237-238). Amsterdam, Nederland: Elsevier.

Scott, M. (2009). In search of a bad reference corpus. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 79-92). Lonon, UK: Routledge.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 43-57). Amsterdam, Nederland: John Benjamins Publishing.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language pedagogy*. Amsterdam, Nederland: John Benjamins.

Sherwood, H. (2018). Religion: Why faith is becoming more and more popular. *The guardian*, *27*(8).

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Stubbs, M. (2010). Three concepts of keywords. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 21-42). Amsterdam, Nederland: John Benjamins Publishing.

Supphipat, P. (2017). *The development of the English content-based reading materials for Buddhist student monks* (Master's thesis). Chulalongkorn University, Thailand.

Supphipat, P., & Chinokul, S. (2018). The development of the content-based reading materials for student monks: Needs analysis. *An Online Journal of Education*, *13*(2), 345-359.