A CORPUS-BASED STUDY ON THE TECHNICAL VOCABULARY OF ISLAMIC RELIGIOUS STUDIES

Srifani Simbuka[a], Fuad A. Hamied[b], Wachyu Sundayana[c], Deny A. Kwary[d]

([a]srifani.simbuka@iain-manado.ac.id)
Institut Agama Islam Negeri (IAIN) Manado
Jl. Dr. S.H. Sarundajang Ring Road I, Manado 95128, North Sulawesi, Indonesia

([b]fuadah@upi.edu; [c]swachyu@upi.edu)
Universitas Pendidikan Indonesia
Jl. Dr. Setiabudhi No. 229 Bandung 40154, West Java, Indonesia

([d]kwary@yahoo.com)
Universitas Airlangga
Jl. Dharmawangsa Dalam Selatan, Surabaya – 60111, East Java, Indonesia

Abstract: This paper charts the construction of a technical vocabulary list called the Islamic Religious Studies Textbooks Vocabulary (IRSTV) which was developed from the Corpus of Islamic religious studies textbooks (CIRST) in an Indonesian Islamic State Institute (IISI). The study is aimed to meet the need of first-year English language learners studying in Indonesian Islamic tertiary education. The IRSTV list contains selected word types extracted from five major sub-disciplines of Islamic Religious Studies (IRS), i.e. the science of Qur'an, the science of Hadiths, Islamic law and jurisprudence, Islamic philosophy and theology, and Islamic mysticism theology, taught in most Indonesian Islamic universities and colleges. The quantitative analysis of frequency, range and keyness score ranking was conducted and aided by Corpus analysis software i.e. Antwordprof and Antconc keyword tool. A final triangulated ranking of these three criteria was conducted to produce a more balanced technical vocabulary list resulting in 262 word types or 239 lemmas of English words that are needed to be learned by English as Foreign Language (EFL) students of Indonesian Islamic universities, state institutes and colleges.

Keywords: corpus-based approach, EFL, Islamic Religious Studies, technical vocabulary

DOI: http://dx.doi.org/10.15639/teflinjournal.v30i1/47-71

Acknowledging the importance of word list in English for Specific Purposes (henceforth ESP), a number of studies have been conducted and focused primarily on producing technical word lists in various disciplines. Some of the examples of those specialized vocabulary lists include a word list for finance (Kwary, 2011), pharmacy (Grabowski, 2013, 2015), nursing (Mohamad & Ng, 2013; Yang, 2015), agriculture (Muñoz, 2015), newspaper (Zhu, 2017), social sciences (Kwary & Artha, 2017), and plumbing (Coxhead & Demecheleer, 2018).

In the context of English Language Teaching (henceforth ELT) in Indonesian Islamic universities and colleges, there has been a plausible discussion on what kind of vocabulary choice is best to represent the values of Islam in English as a Foreign Language (henceforth EFL). Available choices of the versions of the Islamic-related vocabulary in English are Arabic loan words which have become English repertoires (Brown, 1996), translated Arabic-origin words, transliterated Arabic origin words (Hassan, 2016), or a combination of transliterated Arabic words accompanied by a description in English. Yet, studies in this topic suggest an ambiguous stand to the use of Islamic-oriented lexical items. For example, Erlina, Mayuni, and Akhadiah (2016, p. 51) reported that the Arabic-originated word "*surah*" was used in its original lexis instead of presenting their one-to-one counterpart or any similar vocabulary in English. Hassan (2016, p. 126) recommended transliteration for such lexical items that are highly cultural or ideological, hence they have "partial-equivalent" or even "non-equivalent" counterparts in English. On the other hand, Brown (1996, p. 2) pointed out that "It may surprise some readers that many Arabic-Islamic words are in English dictionaries." For example '*ayatollah'* appears in *Longman Dictionary of Contemporary English*, *Oxford Advanced Learners' Dictionary*, *Collins COBUILD English Dictionary* (COBUILD) and *Cambridge International Dictionary of English*.

The pedagogical usefulness of word lists developed on the basis of the occurrences for certain vocabulary items in specific texts has been related to the argument that frequency information is important in language acquisition (Davies & Gardner, 2010; Gablasova, Brezina, & McEnery, 2017; Lei & Liu, 2016). Experts argued that frequency and range of the lexical items are the key information for vocabulary selection in ELT textbooks (Flowerdew, 2012; Grammatosi & Harwood, 2014; Mukundan & Rezvani Kalajahi, 2016; Römer, 2010). For a word list to be claimed pedagogically useful, it has to be designed to address the complexity and dynamic process of language learning. This

means that "a one for all" word list is impossible and that specific type of learners required a specific 'type' of word list (Brezina & Gablasova, 2017). Given their unique needs to master English alongside Arabic (the most important language in Islamic studies), the learners of English in Indonesian Islamic universities are in need of a unique word list containing important vocabulary to meet these needs. This research, therefore, strives to find important vocabulary in the area of Islamic religious studies (henceforth IRS) using a corpus-based approach through the development of a word list related to Islamic studies. The research questions addressed in this study were:

1. What lexical items occur frequently and uniformly across the textbooks of Islamic studies taught at an Indonesian State Institute for Islamic Studies (IISI/IAIN)?
2. Based on the corpus analysis of the IRS textbooks taught at the IISI/IAIN, which word types should be included in the Islamic Religious Studies Technical Vocabulary (henceforth IRSTV) list?

Corpus, Vocabulary and Technical Vocabulary Lists: Some Key Concepts

The current study of constructing a technical vocabulary list in the discipline of Islamic Religious Studies in an IISI/IAIN was mainly built under the theory of vocabulary for language teaching (Flowerdew, 2012; Nation, 2001; Römer, 2006, 2010). This discipline-specific vocabulary list was generated from a specially designed corpus using corpus software based on the Corpus Linguistics approach (Anthony, 2013; Baker, Gabrielatos, & Mcenery, 2013; Kwary, 2011; Kwary & Jurianto, 2017).

Owing to the purpose of this study, it is worth clarifying some of the key concepts on vocabulary, as well as corpus and its related terminologies. The first important concept is 'corpus' (or its plural form 'corpora) which is defined as " … a collection of authentic language, either written or spoken, which has been compiled for a particular purpose" (Biber, 2015; Sinclair, cited in Flowerdew, 2012). The purposes of building a corpus are mainly of the interest of linguistics to describe and provide an explanation of the structure and use of language. Other experts like Stubbs (1996) and Bonelli (2001) as cited by Flowerdew (2012) pointed out that corpus investigation could be used to unpack the socio-pragmatic behavior of particular discourse communities, an interest of socio-pragmatic studies.

The next key concept of this study, which is closely related to the concept of corpus, is 'vocabulary'. A common definition is by Cambridge Advanced Learners' Dictionary (from cambridge.org accessed on November 13, 2017), which states that vocabulary is "All the words that exist in a particular language or subject". Among several classifications of vocabulary, Nation's (2001) typology of vocabulary was chosen as the main reference of this study. Nation (2001) classified vocabularies into four groups based on their frequency of occurrence in texts. The first is High Frequency Words (HFW) (Nation, 2001, p. 15) which occur the most frequent in any text in any language. These words are, very often than not, function words such as articles and prepositions as well as some content words. Nation (2001) also pointed out Michael West's (West, 1953) General Service Lists (henceforth GSL) that contained about 2000 word families as the most popular HFW (Nation, 2001). Second is Academic Vocabulary**,** which contains words that are highly frequent in an academic discourse or academic texts. Coxhead's (2000) Academic Word List/AWL is regarded as the academic vocabulary list and it contains 570 word families that do not belong to the GSL but occurs in reasonable high frequency in a corpus of academic texts across various disciplines. Next is Technical Vocabulary, which is idiosyncratic to texts in a specific subject area that their occurrence in texts outside this specific topic is rare. The items of specific technical word list varied from one discipline, interest or specialization, to another. Examples of technical words are those compiled in specialized dictionaries to meet the need of English as a second language (ESL) or EFL learners studying a specific discipline in English speaking universities or non-English speaking practitioners of a certain occupation. The last group is Low Frequency Words, which are words that occur very infrequently and relatively small in proportion compared to high frequency and academic words. Depending on the genre/text type, the words that were classified into low frequency or high frequency list varied from one particular discipline to another. Some examples of low-frequency words were archaic words that belong to an older variation or a specific dialect of a language, words that were infrequent due to some restriction based on social norms, words from foreign languages, and proper nouns (names of people and places).

This present study focuses on the third category of Nation's (Nation, 2001) vocabulary classification, aiming to filter out technical vocabulary in the area of Islamic Religious studies from a corpus of core textbooks used as references in Indonesian Islamic higher education institute. Distinguishing

technical words or technical vocabulary from other types of words is hardly a simple task. There were three fundamental principles of word list creation, namely: (i) constructing definition, (ii) identifying the target vocabulary, and (iii) identifying the purpose of the word list (Brezina & Gablasova, 2017). Complying with the first principle, the first construct that needed to be clearly defined in the context of the present study was the term 'Islamic religious studies' (IRS). Here, IRS was based on Azra's (2011) proposition that Indonesian Islamic higher education institutions acknowledge both the heritage of Islamic learning tradition originated from the Middle East brought to Indonesia by Arabic scholars and Indonesian *kyai*s since the 16th century and more recent development of Western-based/orientalist influence brought by Indonesian Moslem scholars who pursued their graduate studies in Islamic studies programs in Western universities. This definition acknowledges Islamic studies as an independent discipline with its own theories and approaches of studying Islamic studies from both academic and religious perspectives as suggested by Al-Ghazzali's definition (Al-Ghazzali & Karim, 1993).

To meet the second principle of word list creation, the target vocabulary "Islamic Religious Studies Textbook Vocabulary" (IRSTV) was defined as the vocabulary that reflects Islam-related knowledge and teaching based on a corpus of core textbooks of Islamic religious studies subject taught at an Indonesian Islamic State Institute (IISI/IAIN).

The third principle of "identifying the purpose of the word list" is met by identifying the needs of the target users of this present study in terms of the vocabulary learning aim in this specific ELT context. The needs analysis is conducted through description of most needed lexical items in the target corpus. The source of the target corpus is reference textbooks prescribed by the syllabus of IRS obligatory subjects written in English (see the Method section).

METHOD

The Construction of the Corpus of Islamic Religious Studies Textbooks-Indonesian Islamic State Institute of Islamic Studies (CIRST-IISIS)

The Target Corpus

The construction of the CRIST-IISI was the basis to create a technical vocabulary list in the discipline of Islamic religious studies. As the target

corpus, it was important that the construction of the corpus obeyed the principles of corpus development (Reppen, 2009). These principles include (a) the purpose of developing a corpus which was related to the originality of the corpus being developed, (b) clearly articulated research question(s) that guided the design of the corpus and (c) the representativeness of the corpus, which, together with the issue of practicality, defined corpus size.

The corpus of Islamic religious studies textbooks that include 18,058 word types and 305,701 tokens, was composed using the core textbooks used in the field of IRS. Five subjects of IRS topics were specifically selected from the original eight topics, based on the fact that only five subjects were taught in the research site i.e. an Indonesian Islamic state institute. These subjects were the sciences of the Qur'an or *ulum al-Qur'an*, the sciences of the hadith or *ulum al-hadith* and its methodologies, Islamic law and jurisprudence or *fiqh and/or ushul fiqh* (its methodologies and various branches), sufism/*tasawwuf* and theology, and philosophy/*Kalam*. These core subjects of IRS were taught to all first year students studying at an IISI/IAIN, specifically from the four faculties (schools/colleges) namely the Islamic law (*syari'ah*), Islamic Education (*tarbiyah*), Islamic Economy and Business, Islamic theology (*ushuluddin),* and Communication and '*Da'wah'*.

An important note was that some topics of the eight topics of IRS were not taught as individual subjects at the IISI/IAIN. It means that there were two or more topics of Islamic religious studies that were integrated into one subject of IISI/IAIN, or vice versa where a single topic of Ushul Fiqh and Fiqh was dispersed into two different subjects ("Institut Agama Islam Negeri (IAIN) Manado - Kurikulum Program studi Ahwal Al-Syakhshiyyah," 2019). For example, IRS topics of Creed and theology and 'Sufism' were overlapped in the contents of IISI/IAIN subject of *Ilmu Filsafat dan Ilmu Kalam/*'Philosophy and the Science of rational Kalaam', while at the same time some of the sub-topics of theology were combined with another IRS topic *Sufism*/'Islamic Mystic' into IISI/IAIN subject of *Akhlak Tasawwuf*. Meanwhile, Islamic law and the principles jurisprudence/*Fiqh* and *Ushul Fiqh* that were listed under one integrated topic of IRS was delivered in two different IISI/IAIN subjects, that is, *Fiqh*, a compulsory subject to be learned by all students of IISI/IAIN, and *Ushul Fiqh*, a subject taught specifically to the students of the faculty/school of Islamic Law or *Syari'ah* and the faculty of Qur'anic Science and *Da'wah*.

Complying to one of the principles of representativeness in constructing language corpora that the corpus must be representative of the language being investigated (Reppen, 2009; Sinclair, in Flowerdew, 2012), the selected textbooks were aligned with the course outlines of the aforementioned IRS subjects of IISI/IAIN. Since the syllabus rarely included references in English, the core textbooks were therefore selected on the basis of the proximity of their contents to the core references prescribed for each subject of Islamic studies in the curriculum of the investigated Indonesian Islamic state institute. Owing to the relevance of the selected IRS textbooks with the syllabus of the five IRS subjects of IISI/IAIN, the number of textbooks and chapters of the data source textbooks were purposively selected for further processing. Under the IRS sub-topic *Fiqh* and *Ushul Fiqh*, for example, there were two IRS textbooks that were employed as data sources. This is because the syllabus of the *Fiqh* and *Ushul Fiqh* subjects at IISI/IAIN covered a considerable-wide general sub-topics in this subject area, the references were also comprised of several core textbooks (mainly in Arabic or Indonesian language)

The Reference Corpora

CIRST-IISI as the target corpus of this particular study was compared to several major reference corpora. These corpora were the General Service List/GSL (West, 1953), Academic Word List/AWL (Coxhead, 2000) and the British Academic Written English (BAWE) (Heuboeck, Holmes, & Nesi, 2007).

The GSL contains approximately 2000 most frequent English headwords designed to assist ESL learners to more efficiently learn English by making use of the most frequent words listed in the GSL. The rationale for choosing the classic GSL was mainly because it has been used in a number of studies, and it contains only two word lists (or baselists) which have been noted as general words. Another consideration for using West's (1953) GSL was that our study was intended to provide vocabulary list that should be learned by EFL students. This purpose was similar to West's intention in making his general word list (Nation, 2016). Moreover, as Nation (2016) pointed out, GSL had been proven to be superior than its more recent successors in terms of its carefully decided purpose of development and its use of suitable corpus that represent its purpose. Another well-known and more recent language corpora, British National Corpus (BNC)/Corpus of Contemporary American English (COCA)

list, was not referred in this study due its big number word lists, and the absence of information on how many word lists should be included for analyzing general words. Besides, there are overlaps between the new AWL and the BNC/COCA lists (Nation, 2016).

The second reference corpus was Coxhead's (2000) AWL, which was a list of academic vocabulary containing 570 words. This list was constructed using approximately 3.5 million running words from 414 academic texts. Our study followed closely Coxhead's (2000) technique in determining the words of her list, which was simpler than Gardner and Davies' (2014) corpus-comparison technique**.** This recent study also followed Coxhead's argument that university students had already had some knowledge of English vocabulary acquired from their previous education where they presumably had been exposed to high frequency words of English. Unlike Coxhead (2000), Gardner and Davies (2014) did not use the assumed knowledge of English vocabulary when developing their NAWL list; therefore, their list contained BNC/COCA high frequency words (Nation, 2016). Moreover, the more recent 'new' AWL by Gardner and Davies (2014) has not been widely used and it leans too much towards American English.

The third reference corpus was BAWE which was developed under a project by a collaborative team of researchers from the Universities of Warwick, Reading, Oxford, Brookes (UK) and at Coventry University-towards the end of the project (Heuboeck et al., 2007). One might argue that since BAWE was constructed from students' writing product, it would not serve as an accurate match for the target corpus (CIRST-IISI) whose text source was 'reading' texts. To anticipate this criticism, it is useful to look back at the typology of corpora elaborated as follows.

Corpora were often categorized based on the modality of their (language) input. A corpus can be classified into written corpus and spoken corpus. For example, the two broad classifications of the text in the BNC are written corpus and spoken corpus. Recognizing current development of multimodality of language input, McEnery and Hardie (2012) stated that corpora include written, spoken and even audio-visual ones. In relation to the pedagogical purposes of developing corpora and word lists, Nation (2016) maintained that it was the written versus spoken mode of texts (as sources of corpora) that mostly affect the nature of corpora and the word lists derived from them. This means that there was no significant difference between a corpus from academic writing and a corpus from academic reading, since both of them are in the form of

written corpus. The target corpus of this present study indeed comprised reading texts in IRS, but they were also the products of writing process. This was the first rationale for the use of BAWE in this study.

Receptive (reading and listening) versus productive (writing and speaking) dichotomy was not brought into attention untill it comes to their units of counting, with regard to their differences in determining the nature of a language leaning course design. That is, if an EFL class is aimed to nurture receptive skills, then the unit of counting of the target corpus or word list (used as the basis of its lexical syllabus) should be in lemma or word family. Meanwhile, if the course targets on productive language skills, then the unit of counting should be 'word types' (see Nation, 2016 for more explanation). The target corpus (CIRST-IISI) and IRSTV used word types as the unit of counting, based on the argument that the word list should be used for developing both writing and reading skills of the target students. Therefore, word types were chosen as the unit of counting, in the hope that the 262 word types (meaning much lesser number of word families) of IRSTV list would be an achievable learning aim for the students of Islamic state university students as the target students. This future use of our corpus and IRSTV list served as another argument of using BAWE as one of the reference corpora, owing to its source of texts (advanced English learners' writing texts).

Data Processing

The procedure of data collection involved several steps such as scanning, digitizing and converting texts before they were ready for analysis. The process of scanning hardcopy version of the data source textbooks resulted in electronic softcopy version of the textbooks in pdf format. Digitizing the texts from pdf format into plain texts format was the next step of data processing, which was done using computer software i.e. AntFileConverter (Anthony, 2015) and Nitro10 Pro (Nitro Software Inc, 2015). The scanning and digitizing hardcopy texts were the procedures that had been commonly used in corpus approach studies (Mohamad & Ng, 2013). The next step involved corpus cleaning in which the converted texts as raw data were "cleaned" from typos and unnecessary information such as references/bibliographies. Albeit time-consuming, the cleaning stage was extremely important due to the fact that the data source texts contained inscription in Arabic that often cannot be recognized by the converter software. The results of converting the texts,

therefore, contained unidentified characters to English alphabetical system. Hence, they had to be corrected by referring to the original pdf files of the data source textbooks.

Data Analysis

The first analysis done in this study was aimed to answer the first research question on "What lexical items occur frequently and uniformly across the textbooks of Islamic studies taught at an Indonesian State Institute for Islamic Studies (IISI/IAIN)?" For this purpose, the analysis was conducted to generate a "word profile" of the CIRST-IISI based on Nation's classification of words into four categories of vocabulary (2001). Aided by the AntWordProfiler (Anthony, 2014b), a word list describing the word types and frequency of the CIRST-IISI was created. The word list was generated by uploading the raw data of CIRST-IISI in txt files into the software's word list tool. Since the texts of CIRST-IISI contained a considerable number of Arabic characters as well as transliterated Arabic ones that include apostrophe and dash such as "*al-Qur'an*", a special setting in the word list tool was set in such a way that enabled the software to accurately recognize them. The "token definition" - the formula used to define what count as a "word"- was set up to cope with the unique characteristics of the data.

The Development of Islamic Religious Studies Technical Vocabulary (IRSTV) List

Criteria for Determining IRSTV

Moving forward from developing a corpus of IRS textbooks and describing its coverage in terms of word profile, the next step was to identify lexical items that should be included in the IRSTV list. In this study, the words enlisted in the "Potential IRSTV list" were those that satisfy the criteria/measures below:

1. Word Profile: the word/types that were considered as potential IRSTV should be outside West's (1953) GSL and Coxhead's AWL (Coxhead, 2000). This measure was employed while bearing in mind that previous research had pointed out that technical vocabulary may come from these two lists (Sutarsyah, Nation, & Kennedy, 1994). For the purpose of this study, the GSL and AWL were excluded from the potential source for

IRSTV due to time constraint and other issues. This was one limitation of this study that simultaneously opened up an opportunity for further studies to consider words/types from the GSL and AWL to be integrated into a more comprehensive selection of IRSTV.

2. Range: the words/types occurred in a minimum of 3 sub-corpora of CIRST-IISI (occurred in at least 3 topics of the total 5 topics of IRS). Range was given the primary consideration over the other measures, specifically over frequency. The rationale for placing frequency measure secondary to "range" was that the occurrence of highly frequent words might be biased due to the length of the texts and topic-related words (Coxhead, 2000).
3. Frequency: the words/types of potential IRSTV should have a minimum of 9 frequent occurrences in any 3 of 5 sub-corpora (the frequency cut point is 8.7 in 305,701 words that serve as the population of this study). The frequency cut-off point used in this study referred to Coxhead's (Coxhead, 2000) way to determining the frequency cut-off point of AWL, i.e. 28.5 token per one million words. Other studies that have used this technique were Valipouri and Nassaji (2013) and Kwary and Artha (2017). The former was to determine the words that were included in their Chemistry Academic Word List (CAWL), and the latter was for their Academic Article Word List for Social Sciences.
4. Keyword: the potential IRSTV should have a "keyness value" that was ranked by chi-squared frequency difference between target corpus/CIRST-IISI versus BAWE resulted from the "keyword" analysis using the Antconc software (Anthony, 2014a). The cut off point for the "keyness value" was not based on the ranks of the keyword list. Instead, the aforementioned criteria were used as the basis for integrating the "keyword" list into the potential IRSTV list. In other words, the "keyness values" of the words/types were supplemented into the range-frequency based list of potential IRSTV. This means that only words/types that fulfilled the conditions of point a-c above were supplemented with their "keyness value" for further processing in point (e) below.
5. Triangulated Range-Frequency and "Keyness" (henceforward RFK): The triangulated RFK score is the formula created by the authors. These are based on the combination of range, frequency, and keyness (thus abbreviated as RFK). The final list of potential IRSTV should contain words/types that were included in the first 262 highest RFK triangulated

score. This value was calculated by combining the scores of range, frequency and "keyness" using the formula below:

$$Triangulated\ RFK = \frac{(Frequency + [Range \times 100] + Keyness)}{3}$$

The rationale for preferring the triangulated RFK score rather than any one of the other measures (when each of them is used individually) as the basis for determining IRSTV) was mainly due to the prevailing features of the triangulated measures over the others. This formula has been proven in this study to be an effective formula for determining the IRSTV. First, the triangulated RFK showed the prominence of the word types enlisted in the "potential IRSTV" vocabulary in terms of their distribution across all five topics of the corpus of Islamic studies textbooks (CIRST-IISI), by presenting the "keyness" score of the word types of the target corpus relative to Range-Frequency scores. Second, the triangulated RFK covered the weaknesses of the two other method/techniques by using one technique's strength to overcome the weakness of the other. For instance, the "range" criteria analysis compensated the keyword's analysis weakness for excluding "range" in its analysis.

FINDINGS AND DISCUSSION

Word profile: Mapping the Frequently Occurring Lexical Items in the Target Corpus

In answering the first research question on the description of lexical items that occur frequently and uniformly across the target corpus (data), a word profile analysis was conducted. The word profile analysis resulted in a list of the word types that uniformly and frequently occurred across the sub-corpora that built the CIRST-IISI. These were the words that were categorized based on Nation's classification of vocabulary (2001), i.e., a) most frequent words, b) academic words, c) low frequency words and d) technical vocabulary. The analysis of word profile showed that the target corpus of this study (CIRST-IISI) contained all of these categories:

A total of 234,922 tokens (or 76.85% of the total running words) and 4,520 word types (25.03 % of the total word types) belonged to the 2000 most

frequent English words as listed in GSL. The word profiler tool of AntWordProfiler had automatically categorized the words in the target corpus into words that belong to the GSL base lists (GSL first 1000 and GSL second 1000 most frequent English word family) and AWL. Bearing in mind that a word family may consist of approximately 6 word types (Nation, 2016), it is assumed that the two GSL base lists of 2000 word families would comprise roughly around 6000 word types.

A total of 20,091 tokens (or 6.57 % of the total running words**)** of word types that belonged to the category of academic words as listed in AWL. A total of 50,688 tokens (or 16.58 % of the total running words**)** belonged to the third category of the "Level 0" or "groups not found (in both GSL and AWL lists)". Table 1 showed the categories of the vocabulary of the CIRST-IISI.

Table 1. The Result of Analysis I-Word Profile (The Statistics of the categories of vocabulary that built the CIRST-IISI)

Level	File	Token	Token %	Cum Token %	Type	Type %	Cum Type %	Group	Group %	Cum Group %
1	1_gsl_1st_1000.txt	223,286	73.04	73.04	2,979	16.5	16.5	962	6.93	6.93
2	2_gsl_2nd_1000.txt	11,636	3.81	**76.85**	1,541	8.53	25.03	701	5.05	11.98
3	3_awl_570.txt	20,091	**6.57**	83.42	**1,877**	10.39	35.42	555	4	15.98
0	-	**50,688**	**16.58**	100	**11,661**	64.58	100	11,661	84.02	100
TOTAL		**305,701**			**18,058**			**13,879**		

The third category ("Group not found") contained a combination of five sub-categories (see Table 2). First, there were some potential technical words that were associated with the field of Islamic studies for which the CIRST-IISI was created. These words were termed as "potential IRSTV" and comprised of 6,989 words types and 26,601 tokens including English words and Anglicized Arabic words (verified by checking their occurrence in the data). The term "Anglicized Arabic words" was coined by one of the authors of the Islamic

religious studies used as a data source (Watt, 1985, pp. 25, 29). Some examples of this sub-category were 'caliphs', 'imams' (*'caliph', 'imam'* – Arabic nouns-attached with English plural marker), 'caliphate', 'imamate' ('*caliph'*, '*imam'* –Arabic nouns attached with -ate, English nominal marker.

The second sub-category was the low frequency words, such as proper nouns that contained Anglo-Saxon names associated with English, Anglicized Arabic mystic/*sufi* order names ('Mu'tazilites', 'Hanbalite', 'Ash'arites'), names in Arabic ('al-Ghazali', al-Ansari, al-Basri), as well as names in other languages. This sub-category comprised of 2,693 word types and 13,554 tokens. Next, there were Arabic words that occurred as many as 1,500 word types and 8,753 tokens. Some examples of this sub-category were *'hadith', 'kitab', 'shari'a' and 'sunna*'. Lastly, there was faulty data which resulted from the conversion of pdf files into plain text files. Such errors in data conversion were due to the limitations of computer aided conversion of printed textbooks into digital form, despite the use of the most recent version of the conversion software.

Table 2. Sub-Categories of the "Level 0" Group

	CATEGORY	TOKEN	TYPE
Table 4.1.1.2-A	**Potential IRSTV English-Anglicized Arabic words**	**26,601**	**6,989**
Table 4.1.1.2-B	Arabic words	8,753	1,500
Table 4.1.1.2-C	Proper Nouns (English-Anglicized Arabic words)	13,554	2,693
Table 4.1.1.2-D	Words from other language, abbreviations & numbers	1,208	263
Table 4.1.1.2-E	Faults/faulty data due to errors in the data conversion process (after 6 times cleaning processes-reduced from thousands of this kind).	572	216
	TOTAL	**50,688**	**11,661**

The Word Types of IRSTV List

Addressing the second research question on the words that were worth included in a technical vocabulary list in the discipline of the Islamic religious studies, a manual extraction of these words from the last category of word

profile analysis was conducted. The analysis resulted in IRSTV list containing 262 word types and 9005 tokens. For the purpose of ELT learning, the lemma form of the IRSTV list is preferred on a consideration that lemmas are categorized based on parts of speech, an important information for vocabulary learning (Davies & Gardner, 2010; Lei & Liu, 2016). The top 25 word types of IRSTV list based on the triangulated RFK score were presented in Table 3 below.

Table 3. The IRSTV List

Rank	Group	Range	Frequency	Keyness	R+F+K
1	Islamic	4	692	12,378.6	**4,490.2**
2	Muslim	5	528	9,329.6	**3,452.5**
3	prophet	5	401	8,354.8	**3,085.3**
4	muslims	5	253	4,522.7	**1,758.6**
5	sufis	3	205	4,469.6	**1,658.2**
6	scholars	5	285	3,035.8	**1,273.6**
7	jurists	4	144	2,777.8	**1,107.3**
8	obligatory	3	95	1,625.3	**673.4**
9	theology	3	132	1,564.9	**665.6**
10	doctrine	3	188	1,472.2	**653.4**
11	theological	4	92	1,378.2	**623.4**
12	qur'anic	4	62	1,351.8	**604.6**
13	theologians	3	75	1,425.2	**600.1**
14	mosque	5	59	1,240.3	**599.8**
15	pilgrimage	4	66	1,269.2	**578.4**
16	divorce	3	60	1,218.4	**526.1**
17	caliphs	5	49	1,024.7	**524.6**
18	divine	5	136	937.5	**524.5**
19	imams	4	49	1,068.3	**505.8**
20	god's	3	71	1,127.2	**499.4**
21	legitimacy	3	57	1,113.6	**490.2**
22	caliphate	4	44	959.3	**467.8**
23	ablution	3	46	1,002.9	**449.6**

24	fasting	5	43	794.3	**445.8**
25	prophet's	5	37	783.5	**440.2**

The result presented above highlights several important insights. **First insight**, some obvious Islam-related words such as "Islamic", "muslim" (and its plural form "muslims") ranked in the first 10 items of the vocabulary list, without rock-hard justification. Throughout the IRSTV list, the word family of these word types occurred with considerable prominence, for instance "islamists", "islamization", "non-muslims", etc., as shown below:

Example 1: IRSTV word type "Islamic" and its family of words

Rank	Group	Range	Frequency	Keyness	R+F+K
1	islamic	4	692	12,378.6	4,490.2
44	islamists	3	28	566.1	298.0
64	islamization	3	17	370.6	229.2
80	islamist	3	27	280.2	202.4
92	pre-islamic	3	12	261.6	191.2

Example 2: IRSTV word type "muslim"

Rank	Group	Range	Frequency	Keyness	R+F+K
2	muslim	5	528	9,329.6	3,452.5
4	Muslims	5	253	4,522.7	1,758.6
54	non-muslims	5	14	263.6	259.2
84	non-muslim	3	15	285.0	200.0

Second insight, as it had been mentioned earlier in the section on the word profile of the CIRST-IISI, Anglicized forms uniquely characterized the target corpus. The same feature was also apparent in the IRSTV list that contained lexical items whose base forms were from Arabic language but were adapted into English with English suffixes. Similar to the result of Anglicized Arabic proper nouns (see the section "Findings and Discussion: Word profile point no. 3), English plural marker "-s" attached to Arabic nouns such as the examples below:

Example 3: Arabic base nouns + English plural marker

Rank	Group	Range (R)	Frequency (F)	"Keyness" (K)	Triangulated R+F+K
5	sufis	3	205	4,469.6	**1,658.2**
17	caliphs	5	49	1,024.7	**524.6**
19	imams	4	49	1,068.3	**505.8**
29	madrasas	3	37	806.7	**381.2**
57	shaykhs	3	20	436.1	**252.0**
59	hadiths	3	19	414.3	**244.4**

It is important to point out that the base words of the lexical items in example #3 above, such as *sufi, caliph, imam, madrasa, shaykh* and *hadith* (or its plural form *ahadith*) were not qualified to be included in the IRSTV list, based on the fact that they belonged to the CIRST-IISI's sub-category of (pure) Arabic words. However, hybrid forms like example #3 were unavoidable and confirmed the findings of previously conducted research (Abudukeremu & Shah, 2010) that a specific vocabulary list on IRS needed to be developed either for the sake of compilation of discipline specific dictionary as well as for teaching ESP.

A special note had to be made on the word type "*hadiths*" that appeared 19 times across 3 sub-corpora in the CIRST-IISI of more than 300,000 tokens/running words, or around 60 times in one-million word corpus. The data showed that "hadiths" and its Arabic base "*hadith*" and "*ahadith*" (Arabic: plural of *hadith,* see Esposito, (2018) were less frequently used than its popular translated word "tradition" (Range 5 and frequency 103). Nonetheless, the hybrid word type 'hadiths' was included in the IRSTV list because "tradition" belonged to AWL. This implied that "hadiths" earned its position in the IRSTV list with regards to its features that meet the criteria of this particular research that words enlisted in GSL or AWL did not qualify to be part of the IRSTV. Moreover, Watt (1985) stated that *Hadith* (or *akhbar*), which has commonly been rendered as "traditions" in English, was more preferable against "tradition" because of the ambiguity of this word ("tradition"). Based on these arguments, it was strongly suggested that (Indonesian) ESP/ELT instructors in IRS promote the use of "*hadith*" or "hadiths" to their students rather than their perceived correct translation, i.e. "tradition" or "traditions". There were other alternatives, however, first to present the widely known English translation

"tradition" in any IRS text side by side with the adopted original form of "*hadith*", and second to provide an explanation of the term 'hadith' or 'hadiths' (see also Hassan, 2016).

Third insight, the list of 262 IRSTV can be further classified based on the five themes of IRS i.e. Law and Jurisprudence (*Fiqh* and *Ushul Fiqh*), the Science of Qur'an, the Science of Hadiths, Islamic Theology and Philosophy, and Islamic Mystic (Sufism). Topic-wise, the IRSTV list was comprised of 108 word types or 2,400 tokens that belonged to the Law and Jurisprudence, 24 word types or 567 tokens were in the Science of Al-Qur'an, only 4 or 46 tokens of the IRSTV list belonged to the Science of Hadith, 69 word types or 1,258 tokens were in the Islamic Philosophy and Theology, and finally there were 69 word types that belonged to the Sufism/Islamic Mystic.

Table 4. The IRSTV List by Topics/sub-corpora

Topics	**IRSTV types per topic**	**Total IRSTV types**	**% of Total type of IRSTV**
Law & Jurisprudence (*Fiqh* and *Ushul Fiqh*)	108	262	**40.6**
The Science of Qur'an	24	262	**9.0**
The Science of Hadith	4	262	**1.5**
Islamic Philosophy & Theology	61	262	**22.9**
Islamic Mystic (Sufism)	69	262	**25.9**
			100

When compared to the total tokens and types of each topic of IRS/sub-corpora, the IRSTV list showed that the IRS topic of Islamic Law and Jurisprudence contained the most number of IRSTV than other topics. That is to say that this particular topic had more word types that were recognized in English dictionaries, either due to the close correspondence to the original IRS terminologies in English or due to the fact that semantically they shared similar meaning(s) to well-known words/terminologies in legal and law topics in English. An example was the word 'jurist' (ranked- 7, range-4, Frequency-144, Keyness-2,777.8 and triangulated RFK score- 1,107.3) in the IRSTV list similarly referred to "an expert in law, especially a judge" in the Cambridge online dictionary (https://dictionary.cambridge.org/dictionary/english/jurist).

Table 5 showed the number of IRSTV in each topic or sub-corpora of the CIRST-IISI.

Table 5. IRSTV by topics/sub-corpora versus all data

Topics/ sub-corpora	IRSTV		Data		Percentage of IRSTV Vs Data	
	Token	**Type**	**Token**	**Type**	**Token**	**Type**
Islamic Law and Jurisprudence (*Fiqh* & *Ushul Fiqh*)	2,400	108	109,133	8,949	2.20	1.21
The Science of Qur'an	567	24	31,159	3741	1.82	0.64
The Science of Hadith	46	4	16,047	2,435	0.29	0.16
Islamic Philosophy & Theology	1,258	61	64,606	6,351	1.95	0.96
Islamic Mystic (Sufism)	1,139	69	84,766	9,338	1.34	0.74

Another explanation of the higher number of IRSTV word types was that the sub-disciplines of Islamic Law and Jurisprudence, Islamic Philosophy and Theology, and Sufism were influenced by the nearby discipline of Law, theology and mystic. These subsequently mentioned subjects were known to have roots in Western scholarly tradition, especially in the humanities and social studies. In the meantime, the Science of Al-Qur'an and the Science of Hadith were subjects unique to IRS (Azra, 2015; Khir, 2007). The lineage with humanities and social studies in Western tradition was important in the traffic of vocabulary, since translations of Western classic academic works into Arabic and vice versa were a common practice throughout the history of both Western and Islamic civilization.

On the contrary, the Science of Al-Qur'an and the Science of Hadith had fewer IRSTV due to the fact that many key terminologies in IRS of these sub-corpora were more in Arabic rather than English or Anglicized Arabic words (see Table 6).

As Hassan (2016) pointed out, some terminologies in IRS that come originally from Arabic language were conventionally not translated into other language(s). In addition, the Science of Arabic language, which is seen as instrumental in understanding the religious textual sources (Khir, 2007),

prescribed ways of preserving such terminologies whilst maintaining their universal recognition in the Islamic world. The percentage of Arabic words token in Qur'an and Hadith were higher than other topics. Specific terminologies in the topic of the Science of Hadith related to the word *hadiths*, such as *isnad* (Esposito, 2014) occurred in its original Arabic form. This word occurred with the frequency of 98, against the total of 305,701 tokens of the target corpus (CIRST-IISI) or 294 probable occurrences in one million word tokens in the corpus. In the data, these Arabic words were explained rather than translated into English, which confirmed Hassan's (2016) claim.

Table 6. Arabic Words by Topics/Sub-corpora

Topics/sub-corpora	Arab		Data		Percentage of Arab Vs Data	
	Token	**Type**	**Token**	**Type**	**Token**	**Type**
Islamic Law and Jurisprudence (*Fiqh* & *Ushul Fiqh*)	1960	349	109,133	8,949	1.80	3.90
The Science of Qur'an	**1,920**	**388**	31,159	3741	**6.16**	**10.37**
The Science of Hadith	**1,328**	**235**	16,047	2,435	**8.28**	**9.65**
Islamic Philosophy & Theology	1,233	382	64,606	6,351	1.91	6.01
Islamic Mystic (Sufism)	2,310	442	84,766	9,338	2.73	4.73

CONCLUSIONS

Built upon the corpus of Islamic Religious Studies Textbooks (CIRST-IISI) that include 18,058 word types and 305,701 tokens, the IRSTV list contains 262 word types (or 239 lemmas). The IRSTV list showed that the IRS topic of Islamic Law and Jurisprudence (*Fiqh* & *Ushul Fiqh*) contained the most number of IRSTV than other topics. Similar to many technical vocabulary lists created using the corpus approach, the IRST contained vocabulary that were outside the GSL and AWL list, i.e. Islam-related words such as "Islamic", "muslim" (and its plural form "muslims") ranked in the top 25 most frequent, equally distributed and highly scored words in the keyness and triangulated RFK analysis.

Both the CIRST-IISI and IRSTV list were uniquely characterized by Anglicized forms such as 'hadiths', 'caliphates', 'qur'anic' and 'mu'tazilites'. These hybrid forms of Arabic-origin bases combined with English plural markers and other suffixes, were important words in teaching specialized vocabulary, particularly on their morphological and syntactical characteristic. The nature of IRSTV that include such a variety of forms resulting from a dynamic dialogue between English and Arabic language, provided rich sources to negotiate and accommodate the specific needs of EFL learners in Indonesian Islamic higher education context.

As far as English vocabulary teaching is concerned, it is suggested that English teachers be aware of the high frequency of Anglicized Arabic words such as those that were in the data of this present study (see the findings and discussion section of this article), instead of ignoring them. ELT teachers in Indonesian Islamic universities and other Islam-based tertiary educational institutions are strongly recommended to consult specifically developed word lists, and corpus-based dictionaries and encyclopedia to check the occurrence of technical vocabulary in the naturally occurring language data, as well as their meanings. Hence, ELT in this specific context will constantly be updated with current development in the discipline of Islamic religious studies as well as research innovations in linguistics and ELT.

REFERENCES

Abudukeremu, M., & Shah, M. I. A. (2010). *A corpus-based lexical study of the frequency, coverage and distribution of academic vocabulary in Islamic academic research articles*. Kuala Lumpur: International Islamic University Malaysia.

Al-Ghazzali, A.-H. (1993). *Revival of religious learnings: Imam Ghazzali's Ihya ulumuddin Vol-1* (F. Karim, Trans.). Karachi: Darul-Ishaat.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141-161.

Anthony, L. (2014a). *AntConc*. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2014b). *AntWordProfiler*. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2015). *AntFileConverter*. Tokyo: Waseda University. Retrieved

from http://www.antlab.sci.waseda.ac.jp/

Azra, A. (2011). From IAIN to UIN: Islamic studies in Indonesia. In B.-A. Kamaruzzaman & P. Jory (Eds.), *Islamic studies and islamic education in contemporary Southeast Asia* (pp. 43–58). Kuala Lumpur: Yayasan Ilmuwan.

Azra, A. (2015). The significance of Southeast Asia (the Jawah World) for global Islamic studies: Historical and comparative perspectives. *Kyoto Bulletin of Islamic Area Studies*, *8*(March 2015), 69-87.

Baker, P., Gabrielatos, C., & Mcenery, T. (2013). Sketching muslims: A corpus driven analysis of representations around the word ' Muslim ' in the British press 1998 – 2009. *Applied Linguistics*, *34*(3), 255-278.

Biber, D. (2015). *Corpus-based and corpus-driven analyses of language variation and use*. Retrieved from https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199544004.001.0001/oxfordhb-9780199544004-e-008

Brezina, V., & Gablasova, D. (2017). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. *Applied Linguistics*, *38*(5), 764-767.

Brown, A. (1996). The treatment of religious terminology in English dictionaries. In Khan, J.U., & A.E. Hare (Eds.). *English and Islam, Creative Encounters 96: Proceedings of the International Conference*, 307-314. Kuala Lumpur: Research Centre, International Islamic University Malaysia.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes*, *51*(July 2018), 84-97.

Davies, M., & Gardner, D. (2010). *Word frequency list of American English*, *1-313*. Retrieved from https://www.wordfrequency.info/files/entries.pdf

Erlina, D., Mayuni, I., & Akhadiah, S. (2016). Whole language-based English reading materials. *International Journal of Applied Linguistics and English Literature*, *5*(3), 46-56.

Esposito, J. L. (2014). *Oxford dictionary of Islam (online version)*. Retrieved from https://www.oxfordreference.com/view/10.1093/acref/9780195125580.001.0001/acref-9780195125580

Esposito, J. L. (2018). *Hadith in the Oxford dictionary of Islamic studies*.

Oxford: Oxford University Press.
Flowerdew, L. (2012). *Corpora and language education*. London: Palgrave Macmillan.
Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(June), 155-179.
Gardner, D. E. E., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305-327.
Grabowski, Ł. (2013). Register variation across English pharmaceutical texts: A corpus-driven study of keywords, lexical bundles and phrase frames in patient information leaflets and summaries of product characteristics. *Procedia - Social and Behavioral Sciences*, *95*, 391–401.
Grabowski, Ł. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, *38*(April 2015), 23-33.
Grammatosi, F., & Harwood, N. (2014). An experienced teacher's use of the textbook on an academic English course: A case study. In N. Harwood (Ed.), *English language teaching textbooks: Content, consumption, production* (pp. 178-204). London: Palgrave Macmillan.
Hassan, S. (2016). Islamic religious terms in English–translation vs. transliteration in Ezzeddin Ibrahim and Denys Johnson-Davies' translation of An-Nawawī's Forty Ḥadīths. *Translation & Interpreting*, *8*(1), 117-132.
Heuboeck, A., Holmes, J., & Nesi, H. (2007). *The BAWE corpus manual: An investigation of genres of assessed writing in British higher education*. Retrieved from https://www.coventry.ac.uk/globalassets/media/global/05-research-section-assets/research/british-academic-written-english-corpus-bawe/microsoft-word-bawemanual-v3--bawemanual-v3.pdf.
Institut Agama Islam Negeri (IAIN) Manado - Kurikulum program studi Ahwal Al-Syakhshiyyah. (2019). Manado: IAIN Manado.
Khir, B. M. S. (2007). Islamic studies within Islam: Definition, approaches and challenges of modernity. *Journal of Beliefs and Values*, *28*(3), 257-266.
Kwary, D. A. (2011). A hybrid method for determining technical vocabulary. *System*, *39*(2), 175-185.
Kwary, D. A., & Artha, A. F. (2017). The academic article word list for social sciences 1. *Mextesol Journal*, *41*(4), 1-11.
Kwary, D. A., & Jurianto. (2017). Selecting and creating a word list for English

language teaching. *Teaching English With Technology*, *17*(1), 60-72.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*(June 2016), 42-53.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mohamad, A. F., & Ng, Y. J. (2013). Corpus-based studies on nursing textbooks. *Advances in Language and Literacy Studies*, *4*(2), 21-28.

Mukundan, J., & Rezvani Kalajahi, S. A. (2016). Developing reading materials for ESL learners. In M. Azarnoosh, M. Zerastpishe, A. Faravani, & H. R. Kargozari (Eds.), *Issues in materials development* (pp. 65–74). Leiden: Sense Publishers.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*(July 2015), 26–44.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing* (e-book). Amsterdam: John Benjamins.

Nitro Software Inc. (2015). *Nitro Pro 10*. Nitro Software, Inc. Retrieved from https://www.gonitro.com/nps/product-details/

Reppen, R. (2009). English language teaching and corpus linguistics: Lessons from the American national corpus. In P. Baker (Ed.), *Contemporary studies in linguistics: Contemporary corpus linguistics* (pp. 206-215). London: Continuum.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift Für Anglistik Und Amerikanistik*, *54*(2), 121-134.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. C. Campoy, M. C. C. Cubillo, B. Belles-Fortuno, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 18-38). London: Continuum.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus-based case study. *RELC Journal*, *25*(2), 34–50.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic*

Purposes, *12*(4), 248-263.

Watt, W. M. (1985). *Islamic philosophy and theology: An extended survey* (2nd ed.). Edinburgh: Edinburgh University Press.

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman, Green & Co.

Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*(1), 27–38.

Zhu, J. (2017). *The technical vocabulary of newspapers*. (Master's thesis, The University of Western Ontario, London, Ontario, Canada) Retrieved from https://ir.lib.uwo.ca/etd.