International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 6 No. 2; March 2017
Australian International Academic Centre, Australia
Flourishing Creativity & Literacy

Compiling an OPEC Word List: A Corpus-Informed Lexical Analysis

Ebtisam Saleh Aluthman
Department of Applied Linguistics, College of Languages, Princess Nourah bint Abdulrahman University,PO box 84428, Riyadh, Saudi Arabia
E-mail: esaluthman@pnu.edu.sa

Received: 07-09-2016 Accepted: 15-11-2016 Advance Access Published: January 2017
Published: 01-03-2017 doi:10.7575/aiac.ijalel.v.6n.2p.78 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.6n.2p.78

Abstract

The present study is conducted within the borders of lexicographic research, where corpora have increasingly become all-pervasive. The overall goal of this study is to compile an open-source OPEC[1] Word List (OWL) that is available for lexicographic research and vocabulary learning related to English language learning for the purpose of oil marketing and oil industries. To achieve this goal, an OPEC Monthly Reports Corpus (OMRC) comprising of 1,004,542 words was compiled. The OMRC consists of 40 OPEC monthly reports released between 2003 and 2015. Consideration was given to both range and frequency criteria when compiling the OWL which consists of 255 word types. Along with this basic goal, this study aims to investigate the coverage of the most well-recognised word lists, the General Service List of English Words (GSL) (West ,1953) and the Academic Word List (AWL) (Coxhead, 2000) in the OMRC corpus. The 255 word types included in the OWL are not overlapping with either the AWL or the GSL. Results suggest the necessity of making this discipline-specific word list for ESL students of oil marketing industries. The availability of the OWL has significant pedagogical contributions to curriculum design, learning activities and the overall process of vocabulary learning in the context of teaching English for specific purposes (ESP).

Keywords: Vocabulary Profiling- Vocabulary Learning- Word List- OPEC- ESP

1. Introduction

The linguistic corpus, which is described by Sinclair (1991, p.171) as 'a collection of naturally occurring language text, chosen to characterise a state or variety of a language', had been previously used in a variety of lexicographic research studies. Further, the compilation of different dictionaries and lexical word lists has provoked a growing interest within this field. Coxhead's (2000) AWL endeavour revolutionised corpus-informed lexical research and initiated the field of discipline-specific word lists research. Attention had begun to be paid to the fact that different disciplines that displays different registers require the availability of discipline-specific word lists (Hyland & Tse, 2007). Stressing the need for the compilation of such word lists, many studies have been carried out to examine the coverage of the most well-known word lists, namely the AWL and the GSL, in specific corpora. This body of investigation has yielded significant results indicating the low coverage of the AWL in specific corpus and, thus, the need for specialisation-specific word lists (see for example, Chen & Ge, 2007; Vongpumivitch et al., 2009; Martinez, 2009).

Over the past two decades, scholarly investigation has begun to take this need into account. Many word lists have been compiled in the fields of medicine (e.g. Chen and Ge, 2007; Wang et al., 2008), pharmacology (Fraser, 2007), nursing (Yang, 2014) and engineering (Ward, 2009). Similarly, a number of word lists have been compiled to achieve specific purposes. Jin et al. (2012) compiled the most frequent 100 technical words in the TOFEL course books using WordSmith Version 4.0 RANGE software with the aim of assisting students and instructors in familiarising themselves with the TOFEL exam registers.

As far as ESP is concerned, a word list that assists students in oil industry and marketing majors could potentially have great significance. This is not only because of the lack of such word lists, but also because oil marketing is gradually gaining importance on both the global and local levels. According to recent OPEC reports, The Kingdom of Saudi Arabia (KSA) owns 18% of the entire world's oil reserves, and it has been ranked as the main exporter of oil. This importance is reflected in the local educational context by the establishment of the King Fahd University of Petroleum and Minerals (KFUPM) in 1963, with the mission of making a change within the KSA in the fields of sciences, engineering and business. Arabian-American Oil Company (Aramco) is the largest government-owned oil and Petroleum Company in the KSA, and it provides numerous opportunities for training and education.

Apparently, oil marketing and oil education are increasing in visibility also at different global contexts. Oil marketing reports and reviews constitute a permanent section in most, if not all, global newspapers, as well as in broadcast news. English is the lingua franca of both the oil industry and the oil marketing. In fact, many institutions related to the field of oil industry have designed their English language classes with the aim of improving their ESL students' English

[1] OPEC stands for Organisation of Petroleum Exporting Countries.

proficiency so that the students are able to manage and market oil products, analyse oil marketing reports and conduct oil financial marketing analyses. A variety of English textbooks for students of oil and gas majors have been made available by ESP publication houses in order to provide assistance to the educational institutions serving the oil marketing industries.

The present study aims to compile a list of the words most commonly used in the oil marketing industries. To this end, a corpus of OPEC monthly reports consisting of 1,004,542 words was compiled. The Lexical Frequency Profile was created by using the RANGE computer programme developed by Laufer & Nation (1995). Both the OMRC and the generated OWL will be investigated with regard to the three main concerns identified in the word list research: the coverage of the AWL and in GWL in this particular corpus, the coverage of the OWL in the OMRC, and the pedagogical implications and suggestions of the availability of such word lists for educators and language instructors. The results obtained from the current study will benefit multi-disciplinary areas of teaching and research, including vocabulary learning and teaching, curriculum design and lexicographic research. The generated OWL and the results concerning the coverage of both the GSL and the AWL in the OMRC will have a significant impact on the field of English learning as a foreign language, as it will direct a prioritised language instruction approach, compile learners' dictionaries and develop pedagogical materials.

2. Preliminary Review on Academic Word Lists

The significance of vocabulary learning in all phases of second language learning has been well recognised in second language acquisition (SLA) fields. This awareness of vocabulary learning's significance in the diversity of linguistic disciplines is stated early by Wilkins (1972) in his *Linguistics in Language Teaching* by emphasising the essential role of vocabulary learning in all forms of ESL communication (p. 110-11). The centrality of vocabulary to all learning tasks is reflected by the huge body of scholarly research investigating the mechanisms of vocabulary learning and teaching from a variety of perspectives (see Nation, 2000a, 2000b, 2001a and 2001b for an overview of the issue). Nation (1990) indicated that there are 54,000 English word families and he categorised these word families into four main categorises (2001a): the high-frequency words that are the most common words in English, the academic words that are frequent in the academic context, technical words that are different from a discipline to another, and the low frequent words. The first category refers to the general service vocabulary occurring frequently across different English texts. Academic vocabulary is most frequently found in academic registers, but not in any specific knowledge discipline, and they have semi-technical implications. On the other hand, technical vocabulary is context-bound and most frequent in a specific knowledge discipline; consequently, technical vocabulary differs from one discipline to another. The final category of vocabulary includes words that are used infrequently and are the least significant for learning.

Identifying the vocabulary that ESL learners need for the purpose of making a particular word list is one of the most crucial areas of vocabulary learning research. The main thrust of word lists research, however, has been towards the compilation of word lists that serve real pedagogical purposes. Without a doubt, the availability of different corpus analysis tools has significantly contributed to the field of lexicographic research in general, as well as to the compilation of different types of word lists in particular (Hartmann, 2001).This interest in compiling word lists dates back to the 1950s, when West (1953) created the GSL, the first English word list, which includes high-frequency words across 2,000 word families and typically make up 80% of the English texts (Nation, 2001a). Thus, the GSL has been a learning priority in ESL learning contexts. Coxhead's (2000) AWL is considered to be the most representative academic word list, against which the relevance and coverage of many of the recent academic discipline-specific word lists are measured. The AWL, which consists of 570 word families, was created from a corpus of balanced academic texts in science, arts, commerce and law, totalling 3,5000,000 words. The AWL list covers 10% of academic texts and is claimed to be the second learning priority after obtaining the first 2,000 top words in the GSL list in academic contexts (Coxhead, 2000; Nation, 2006a, 2006b).

Calls emphasising the significance of learning Coxhead's AWL list for academic goals have been reflected in the appearance of numerous ESL textbooks and English learning and teaching websites based on the AWL list (see Paribakht & Webb, 2016 for examples). However, though the compilation of the AWL has significantly contributed to ESL learning and has inspired numerous word list studies, including the present one, it has been subjected to criticism regarding its relevance to discipline-specific needs. The AWL's coverage across different specific disciplines has been investigated and found to account for only 10.07% of medical papers (Chen & Ge, 2007). Similarly, Vongpumivitch et al. (2009) reported that the AWL constitutes only 11.17% of their 1.5 million word corpus of research papers in the discipline of applied linguistics. In the agriculture discipline, Martinez (2009) found that only 92 words from the AWL were considered frequent in his corpus.

Hyland and Tse (2007) insisted that different disciplines or knowledge fields display noticeable variations regarding their most frequently used words in their context-bound register. This goes in line with what Nation (2001a) stated that 'when learners have mastered the 2,000-3,000 words of general usefulness in English, it is wise to direct vocabulary learning to more specialised areas' (p. 187). Such scepticism of the AWL's relevance to specific discipline has resulted in the recent compilation of different discipline-specific word lists. Prior to these compilations, specific corpora have been compiled to serve the corpus-lexical analysis. The usefulness of specific corpus in the ESP context in general, as well as specific disciplines in particular, has been documented in corpus research (see Sinclair, 2012). McEnery and Wilson (2001) stated that a specialised corpus can be utilised to assess discipline-specific language learning by providing 'quantitative accounts of vocabulary and usage which addresses the specific needs of students in a particular domain more directly than those taken from more general language corpora' (p. 121). The current study is based on a

number of criteria regarding the compilation of both the corpus and the generated word lists (An overview of these criteria will been given in the methodology section below.)

A number of academic discipline-specific word lists have been conducted in the field of medicine. Based on medical research articles, Chen and Ge (2007) utilised a corpus-based analysis in order to examine the distribution and the coverage of the AWL word families in medical texts. They found that the AWL word list is different from the top frequent words that appeared in their medical corpus. The Medical Academic Word List (MAWL) was the first representative medical word list, compiled by Wang et al. (2008) from a corpus of medical research articles (1.09 million words). MAWL comprises 623 words and contains 342 words from the AWL list (Wang et al., 2008). Hsu (2013) created another medical word list (MWL) using the RANGE software developed by Nation and Heatley (2005). The MWL constitutes 595 frequent words that make up 10.72% of the running words of Hsu's entire medical corpus.

Using a similar methodological approach, Fraser (2007) compiled the Pharmacology Word List (PWL), which is a corpus of pharmacy research articles. The PWL consists of 601 words and covers, along with the GSL and the AWL, 88% of Fraser's corpus. In 2009, Fraser excluded the GSL and the AWL and developed another PWL consisting of only core pharmacological words (2,000 words); he found it to have a higher lexical coverage in his corpus (89%) than the first PWL. Yang (2015) created a specific word list for the nursing field. He compiled the Nursing Research Articles Corpus and aimed to create an academic word list in nursing using the RANGE software. Yang's Nursing Academic Word List (NAWL) consists of 676 word families and are found to be covering 13.64% of the texts in the nursing corpus (1,006,934 words). Yang also compared the NAWL to the MAWL and the AWL and found out that 378 word families in his compiled nursing wordlist, the NAWL, overlap with the AWL, and that these 378 word families make up 8.93% of the entire corpus. He also found out that only 192 (33.69%) of AWL word families appear in his corpus, which is considered to be quite low in terms of frequency. On the other hand, 429 (63.46%) word families were found in both the NAWL and the MAWL, indicating great lexical similarity between the two disciplines.

In the field of engineering, Tigchelaar (2015) triangulated rating criteria corpus tools with instructors and compiled the Engineering Academic Formula List (EAFL). He created a corpus of engineering research articles (1,000,000 words) and extracted the most frequent 765 formulistic lists. To decide the pedagogical functionality of these formulistic lists, a rating feedback on the relevance of these words to the engineering discourse was obtained from 12 engineering teaching assistants.

The well-documented studies summarized in this section have (a) an average corpus size (1,000,000 words) and (b) a Lexical Frequency Profile (LFP) approach. However, none of the available word lists have targeted the discipline of oil marketing industries. Following these studies and filling the literature gap, this study aims to compile a word list for the field of oil marketing industries with an accessible format. In the context of developing a knowledge society, which is increasingly recognised as a source of global competitiveness and economic well-being, this project will make a contribution by enhancing both oil-based language analysis and language learning. Adopting the well-known criteria of corpus design, best documented in the literature (Sinclair, 2005), and by relying on LFP tools, this study will result in a generated word list as its primary goal. To the best of this researcher’s knowledge, this is the first attempt at compiling a discipline-specific word list in this particular discipline.

3. Research Questions

This study’s investigations are guided by the following two questions:

1- How much of both the GSL and the AWL are found in the OMRC? What is the coverage of the first 1000 words in the GSL, the second 1000 words in the GSL, and the AWL in the OMRC?

2- What are the most frequently used words (apart from those in the GSL and the AWL) that are generated from the corpus-based on frequency and range criteria?

4. Methodology

This section is devoted to discussing the present study’s methodological framework. To best answer the research questions posed in the current study, a discipline-specific corpus was compiled on which an LFP had to be conducted. Below is an explanation of these two phases, including creating, annotating and utilising corpus tools and certain criteria in compiling the main deliverable product of this study, the corpus-based OPEC word list.

4.1. The OMRC Design

4.1.1 The OMRC Design Criteria

Due to the heterogeneous nature of the available and well-recognised English corpus, it is particularly significant to compile a discipline-specific corpus design that best meets the scope of corpus-based lexicographic research (Flowerdew, 2012). Forty monthly OPEC reports, released between 2003 and 2015, have been collected in order to create the OMRC (1,004,542 words). The word count of these reports is between 11,591 and 122,710, and all of these reports are published in the OPEC main website at http://www.opec.org/opec_web/en/.

Basically, the OMRC has been designed based on an extensive reading of the well-documented practices and criteria of corpus-based lexicographic research, with Sinclair (2004, 2005) and Flowerdew (2012) serving as the main references. Below is a summary of the main criteria followed in creating the OMRC:

1-The selected monthly and annual OPEC reports have been selected without regard to the type of language, but rather to the communicative functions they achieve in a particular community,

i.e. experts, analysts, merchants and politicians in the field of oil marketing and oil industry.

2-In order to achieve the representative criteria, samples of the collected reports have been read and rated in terms of their relevance to the field by two instructors in the field of teaching ESP.

3-The corpus design criteria determining the structure of the OMRC clearly delineate the corpus under investigation from other corpus. Table 1 illustrates the corpus design criteria proposed for compiling the OMRC in this study.

Table 1. The Corpus Design Criteria Proposed for the OMRC

OMRC Design Criteria	Attributes
Mode	Writing (written-to-be-read)
Genre	Reports
Domain	Oil marketing and industry
The language of the corpus	English
Participants	Analysts, experts, specialists and OPEC professionals
Setting	Business
Function	Informative, reflective
Technicality	Technical, Semi-technical

4-All metadata illustrative information regarding the collected texts is stored separately from the OMRC plain texts.

5-The process of designing, annotating and compiling the OMRC is documented clearly in the metadata information document.

6-While targeting an adequate and convenient size, the OMRC is aimed primarily at homogeneity of the collected texts.

4.1.2 The size of the OMRC

Sinclair (2005, p. 1-21) maintains that corpus builders have to target the homogeneity in the collected texts while still keeping a sufficient coverage. As far as corpus size is considered, Sinclair (2005, p. 1-21) argues that the minimum corpus size is dependent on a number of issues, including the purpose of the investigation and the kind of methodology used in studying the corpus. Pravec (2002) argues for considering the amount of time consumed adequately in compiling a representative learner corpus. He agrees with Sinclair (2005) in that the size of the corpus size relies heavily on the purpose and needs of its builders. Many of the corpora recorded in the Centre for English Corpus Linguistics, UCL, as well as those used in lexicographic research, are huge in size, with approximately 1,000,000 words. Taking the purposes and the duration of the present project in mind and following most corpus-based dictionary studies, the OMRC is designed to be consisting of 1,004,542 words.

4.1.3 Collecting, Converting and Preparing the OMRC.

All the collected monthly reports were published in PDF format. All PDF files have been converted into Microsoft Word documents for the purpose of cleaning up and normalising the data (Weisser, 2016). Cleaning up the data was conducted via 'Cleaning Written Data' (cleanup.html) from the text editor. All irrelevant figures, appendices and charts have been removed. In order to compile the OMRC, the collected files have been converted into plain text format using the extension '.txt'. The data of the OMRC is available for download in different formats (PDF, TXT and XLM with metadata documentation) and can be obtained by contacting the researcher. It is within the intention of the researcher that the OMRC is utilized for a variety of corpus-informed studies.

4.2 Data Processing: Lexical Profiling

In order to best answer the research questions of the present study, the OMRC have been processed through a lexical profiling analysis that produces statistical summaries. This lexical profiling was conducted by the RANGE and FREQUENCY software, which can be downloaded as a zip file from http://www.vuw.ac.nz/lals/staff/paul_Nation (Healy, Nation & Coxhead, 2002).

Processing the OMRC into this software generates a compilation of the most recurrent words, based on frequency and range criteria. It also generates a statistical summary about the coverage of the AWL and GSL in the OMRC. The RANGE software also generated statistical descriptive results regarding the overlapping between the processed corpus and the three base lists available in the program. These base lists are as follows: (1) BASEWARD1.txt is the first 1,000 words in the GSL; (2) BASEWRD2.txt is the second 1,000 words in the GSL; and (3) BASWARD3.txt is the AWL word lists. The function words were excluded from the processing LFP analysis by choosing the 'Use Stop' list function and clicking the function.txt file available in the software.

5. Results

Processing the OMRC corpus into the RANGE and FREQUENCY software yielded a number of lexical profiling results. Statistical lexical profiling provides lexical analysis in the terms word tokens, word types and word families. According to Nation (2001, p. 7-8), token refers to every word form in a text, regardless of whether the same word is repeated. Type is to the actual words in a text. A word family contains a headword and all of its derived inflectional related words.

The OMRC files were analysed through the RANGE and FREQUNCY program, and an initial summary was given on the number of lines and words in each file, the total number of words, the size of nodes, the number of words in each file, the memory used and the time taken to finish the analysis (four seconds). Below is a summary of the results related to the two research questions posed in this study.

5.1 Overlapping between the GSL and the AWL and the OMRC

The RANGE software provides an analysis showing how much coverage each of the three base word lists has in the OMRC. Lexical profiling is generated in terms of four main vocabulary frequency zones (the four word lists provided in the programme): (1) the GSL first 1,000 words, (2) the GSL second 1,000 words, (3) the AWL and (4) words that are not in all of the above lists.

Table 2. The coverage each of the three base word lists has in the OMRC.

WORD LIST	TOKENS%/	TYPES%/	FAMILIES
One	308,024/48.22	2,224/11.71	767
Two	37,144/ 5.82	1,157/ 6.09	532
Three	90,365/14.15	1,689/ 8.90	524
Not on the lists	203,219/31.82	13,917/73.30	

Table 2 shows that 308,024 word tokens in the OMRC belong to the first 1,000 words in the GSL, and those words account for 48.22% of all of the running words in the OMRC. Regarding word types, 2,224 word types in the processed data belong to the GSL's first 1,000 words, making up 11.71% of the total word types in the OMRC. Along the same line, 767 word families in the data are in the same word list of the GSL's first 1,000 words. These results suggest that the GSL's first 1,000 words cover a great amount of the OMRC.

However, analysis shows that the GSL's second 1000 words have very low coverage. Table 2 reveals that 37,144 word tokens in the OMRC are in the GSL's second list, with those word tokens making up only 5.82% of the total running word tokens. Similarly, only 1,157 word types in the OMRC belong to the GSL's second 1,000 words, which constitute only 6.09% of the overall word types.

Regarding the third basewrd3.txt referring to the AWL, Table 2 reveals that 90,365 word tokens in the text are in the AWL and make up a moderate coverage of the whole OMRC (14.15%). Along the same line, 1,689 word types of the OMRC fall under the AWL and make up only 8.90% of the total word types in the OMRC.

The aforementioned results give an idea about the lexical coverage of the OMRC in terms of the three well-recognised wordlists. Lexical coverage is 'the percentage of running words in the text known by the reader' (Nation, 2006, p. 61). It has been claimed that the higher lexical coverage the reader has, the better his or her reading comprehension is (Hu & Nation, 2000; Nation, 2001, 2006; Laufer, 1989). The smallest lexical percentage that ensures successful reading comprehension is 95% (Laufer, 1989). Acquiring the GSL will ensure a high percentage of lexical coverage (54.04%) of the OMRC, with consideration of the huge lexical coverage of the first GSL 1,000 words, as compared to the next 1,000 words. On the other hand, mastering both the GSL and the AWL will ensure 68.19% coverage of the text.

The remaining 31.81% coverage of the text includes words that do not belong to three main base words. As illustrated in Table 2, 203,219 word tokens occurring in the OMRC do not belong to either the GSL or the AWL (13,917 word types). The second phase of the present study is concerned with compiling the OWL using these 203,219 word tokens. It is worth mentioning that acquiring the GSL (the most common 2,000 words in English) and the generated OWL will only ensure 85.86% lexical coverage of the OMRC texts. Below is an explanation of how the OWL was compiled.

5.2 The Compilation of the OWL

Word types that are not found in any of the first three base words were displayed in terms of their range, frequency and number of occurrence in each file in the OMRC. Those word types were investigated, and the OWL was compiled. Following Coxhead (2000), two main selection criteria were set for a word to be included. First, it had to have a frequency indication of no less than 20. Second, it had to have a range indication of no less than 10. Given the fact that 40 file reports constitute the OMRC, the range of 10 accounts for a quarter of the total number of files.

Based on these two selection criteria, 255 words were included in the OWL. Appendix A shows these 255 word types into alphabetical order, along with their frequency and range. A noticeable variance was observed when considering both frequency and range. The word 'crude' occurs 6,319 times with a range of 40. On the other hand, the word 'disclaimer' occurs only 20 times with a range of 15. Table 3 shows a list of the 60 most frequently used words in the

OWL, which have a frequency rate between 203 and 6,319 and a range rate between 17 and 40.

Table 3. The 60 most frequent words in the OWL in terms of frequency and range.

Word types	Range	Frequency	Word types	Range	Frequency
CRUDE	40	6319	INFLATION	38	940
FUEL	39	3082	JET	38	872
OECD	39	3058	AMID	38	861
GASOLINE	39	2679	INVENTORIES	37	821
IMPORTS	39	2314	PETROLEUM	40	780
GRAPH	36	2143	DOWNWARD	38	762
REFINERY	39	2093	REFERENCE	39	628
FORECAST	39	1720	DISTILLATE	38	625
FREIGHT	39	1358	SENTIMENT	36	602
BRENT	38	1126	SECRETARIAT	39	583
TANKER	38	567	CONSECUTIVE	35	364
VERSUS	38	553	SURPLUS	37	318
SEASONAL	38	516	MOMENTUM	39	317
REFINERIES	38	477	FIXTURES	35	305
FUELS	36	457	ARBITRAGE	36	303
RESIDUAL	38	416	REFINERS	38	302
ANALYTICS	16	409	LONG-TERM	22	295
HAVER	16	409	TONNAGE	36	292
PREMIUM	39	408	SUEZMAX	36	289
PACE	39	396	PORT	29	288
ROUNDING	36	281	CARGOES	37	245
BEARISH	37	278	FISCAL	37	198
RETAIL	38	276	PIPELINE	36	241
DISTILLATES	37	267	DEFICIT	39	209
AFRAMAX	36	262	BIOFUELS	25	206
SLOWDOWN	37	261	PEAK	38	206
BULLISH	36	249	CFTC	36	205
CONTRACTION	34	249	RECESSION	30	204
MONETARY	39	249	REFORM	17	203
DISCOUNT	38	246	UNLEADED	37	203

6. Conclusion

This study was a response to the lack of discipline-specific word lists in a workplace domain of oil marketing industries that serves ESP context. The first phase of this study's investigation was guided by the first research question, which aimed to identify the GSL's and AWL's coverage of the OMRC. The second phase of the investigation resulted in the compilation of an OWL.

Results concerning the GSL's and AWL's coverage in the OMRC significantly inform what to focus on after acquiring the first 1,000 words of the GSL (the most common words in English) in the context of learning English for the purpose of the oil marketing industries. The first 1000 words of the GSL were found to account for 48.22% of all of the running words in the OMRC. On the other hand, the second 1000 GSL words were found to account only for 5.82% of the running words in the OMRC. This implies that the second 1000 GSL words have a low occurrence in the OMRC and there is very few probabilities of being occurred and encountered in the OPEC related written texts. The total coverage of the GSL in the OMRC is only 54.04% of the OMRC running words. As the GSL accounts for 80% coverage of the English academic texts (Nation, 2001), attention should be paid to the coverage of technical words that are specific to the kind of discipline investigated in this study.

The focus of the coverage of technical words in the OMRC and the inevitability of compiling a specific-discipline word

list is also emphasised by the low coverage of the AWL in the OMRC. The AWL has been found to account for only 14.15% of the OMRC. The results in this regard are consistent with earlier studies, which indicated that the AWL's coverage across different disciplines is between 10% and 12% (Chen & Ge, 2007; Vongpumivitch et al., 2009; Martinez, 2009) and also agrees with the early claims that have necessitated the making of specific-discipline word lists (Nation, 2001a; Hyland & Tse, 2007; Chen & Ge, 2007; Wang et al., 2008; Martinez, 2009; Hsu, 2013). Though both the GSL and AWL are the most cited word lists in the existing literature, they do not account sufficiently to the OPEC related written texts. The OWL is a specific- discipline world list relates to the field of oil marketing industries. The availability of such a word list, according to Hyland & Tse, (2007), will assist in preparing ESL students in their academic studies by familiarizing them with the kind of oil marketing industries discourse. This is due to the fact that different disciplines have different norms of explaining knowledge (Hyland & Tse, 2007). The generated OWL accounts for 31.81% coverage of the OMRC and has a range of (10 to 40). This suggests that the OWL is a representative of the OPEC written discourse.

The Range and Frequency analysis yielded 255 words which have been chosen to compile the lexical repertoire needed by ESL students after acquiring the GSL. Following Coxhead, (2000) when compiling the AWL, range criteria was prioritised over frequency. Only words that occurred within a range of 10 to 40 times were included. According to Coxhead (2000), depending on the frequency criteria alone would result in a biased decision towards topic-related words. An important result to be taken into account is that acquiring the GSL (the most common 2,000 words in English) and the generated OWL ensure a high percentage of lexical coverage (85.86%) of the OMRC texts. On the other hand, acquiring the GSL and the AWL will ensure 68.19% coverage of the OMRC. The lexical coverage percentage has been claimed to be determining the reading comprehension adequacy (Nation, 2006). The lexical coverage that insures an optimal adequate reading comprehension is found to be 98% (Nation, 2001) and 95% for a satisfactory comprehension ability (Laufer, 95%). These results makes the OWL of a more priority to learn and acquire than the AWL for ESL students starting to be enrolled in educational courses in the field of oil marketing industries.

7. Pedagogical Implications and Future Research

The significance of the investigation conducted in the present study is related to the significance of ESL vocabulary learning in in a number of ways. Identifying the AWL's and the GSL's coverage and generating the OWL from the statistical lexical profiling process have pedagogical implications in the field of learning and teaching English vocabulary. Specifically speaking, they contribute to the context of learning English for the purposes of oil marketing industries in a variety of ways, including both direct and indirect pedagogical applications.

First, the compilation of the OWL contributes to what Sinclair (1998) and Willis (1990) called 'the lexical syllabus' in the ESL context. The 'lexical syllabus' is that which is designed according to the most frequent lexical words, patterns and phrases, rather than grammatical patterns and structures. Willis (1990) emphasised the fact that grammatical structures are not neglected when designing the lexical syllabus. Rather, he claimed that the most frequent lexical patterns are used to exemplify the main grammatical forms. With the exception of McCarthy et al.'s (2005) Touchstone Series, most of the educational materials in the field of English learning and teaching are not designed using a corpus-based approach. The very few educational textbooks available for teaching English for the purposes of oil marketing industries are no exception. Moreover, the compilation of the OWL does contribute to vocabulary earning activities related to learning and teaching English for the purposes of oil marketing industries. The OWL assists language instructors in deciding what vocabulary they should focus on in concordance activities and cloze exercises. Within this context, the instructor's role is to guide EsL students in their learning by providing the input that they really need.

Second, the integration of the OWL into the context of teaching, learning and using English for the purpose of oil marketing industries facilitates learners' access to the academic and professional discourse related to that genre. By being familiar with the most frequent technical words in this field, learners can perform a variety of tasks, including negotiating, analysing, debating, arguing and writing about this discipline-related topics.

Third, the compilation of the OWL can have further significance for translation activities related to the field of oil marketing and industry. It is the researchers' intention that the generated OWL be presented in a bilingual English/Arabic thesaurus. Such a thesaurus can be used to complement translation activities by providing translators with the following: (a) what they need to define the genre in terms of the most frequent lexical patterns and (2) acquiring discipline-specific terminology and phraseology, which has been argued to be essential to translation activities (Pearson, 1998).

An additional significant aspect of the present study is the availability of the OMRC (1,004,542 words) in plain .txt format. The OMRC is quite comprehensive in corpus size. O'Keeffe et al. (2007) indicated that, in order to make a reliable generalisation, a corpus should not be less than 1,000,000 words. The OMRC will significantly contribute to any topic-related investigation in the field of lexicography, discourse or register.

The corpus-informed lexical investigation conducted in this study yielded significant results in relation to oil marketing industries discipline. However, it would be advantageous to conduct future studies on the collocational patterns of the most common words of the OWL. Further research along this line of enquiry would be valuable if it duplicates the range and frequency statistical analysis utilized in the present study with rating criteria conducted with teachers and experts in the related field. It is also recommended that other corpus-informed studies are to be conducted to determine the most common lexical bundles and formulistic lists in the existing OMRC utilizing a variety of corpus analysis tool.

References

Chen, Q. & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes, 26*. 502-514.

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

Heatly, A., Nation, I. S. P. & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved from: http://www.vuw.ac.nz/lals/staff/paul_Nation

Hu, M. & Nation, I.S.P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. London, United Kingdom: Pearson.

Hyland, K. &Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Hsu, W. (2013). Bridging vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research, 17*(4), 454-484.

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education, 10*(Issue Number), 127-143.

Fraser, S. (2009). Breaking down the divisions between general, academic and technical vocabulary: The establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education, 12*(Issue Number), 151-176.

Jin, N., Tong, C., Nor, M., Tarmizi, M. & Mahmad, A. (2012). Corpus based analysis of the TOFEL course book: What are the words we should teach our students? *International Review of Social Sciences and Humanities, 3*(2), 152-160.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), Special language: From human thinking to thinking mechanics (p. 316-323). Clevedon, United Kingdom: Multilingual Matters.

McCarthy, M., McCarten, J. & Sandiford, H. (2005). *Touchstone*. Cambridge, United Kingdom: Cambridge University Press.

McEnery, A.M. & Wilson, A. (2001). *Corpus Linguistics (Second Edition)*, Edinburgh, United Kingdom: Edinburgh University Press.

Mudraya, O. (2006). Engineering English: A lexical frequency instructed model. *English for Specific Purposes*, *25*(2), 235-256.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York, New York: Newbury House.

Nation, I.S.P. (2000a). Review of what's in a word? Vocabulary development in multilingual classrooms by N. McWilliam. *Studies in Second Language Acquisition, 22*(1), 126-127.

Nation, I.S.P. (2000b). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal 9*(2), 6-10.

Nation, I.S. P. (2001a). *Learning Vocabulary in Another Language*. Cambridge, United Kingdom: Cambridge University Press.

Nation, I.S.P. (2001b). *Managing Vocabulary Learning*. Singapore, Singapore: RELC.

Nation, I.S.P. (2006a). Language education - vocabulary. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (pp.494-499), 2nd Ed., Oxford, United Kingdom: Elsevier.

Nation, I.S.P. (2006b). Second language vocabulary. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (pp.448-454), 2nd Ed. Oxford: Elsevier.

O'Keeffe, A., McCarthy, M. & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge, United Kingdom: Cambridge University Press.

Paribakht, T. & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes, 21*(Issue Number), 121-132. Available at: http://dx.doi.org/10.1016/j.jeap.2015.05-009

Pearson, J. (1998). Terms in Context. Amsterdam, Netherlands: John Benjamins.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, United Kingdom: Oxford University Press.

Sinclair, J. (2004*). How to Use Corpus in Language Teaching*. Amsterdam, Netherlands: John Benjamins.

Sinclair, J. (1998). Large corpus research and foreign language teaching. In R. de Beaugrande, M. Grosmn & B. Seidlhofer (ed), *Language Policy and Language Education in Emerging Nations* (pp.79-86). London, United Kingdom: Albex.

Sinclair, J. (2005). Corpus and text-basic principles. In M. Wynne (ed.), *Developing Linguistic Corpus: A Guide to Good Practice*, (pp. 1-12). Oxford, United Kingdom: Oxford Text Archive. Available at: http://ahds.ac.uk/linguistic-corpora/

Sinclair, J. (2012). *Corpora and Language Education.* New York, New York: Palgrave Macmillan.

Tigchelaar, J. (2015). Creating an engineering academic formulas list. *The Journal of Teaching English for Specific and Academic Purposes*, *3*(2), 295-304.

Vongpumivitch, V. Huang, J. & Chang, Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33-41.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170-182.

Wang, J., Liang, S. & Ge, G. (2008). Establishment of a medical word list. *English for Specific Purposes*, *27*(4), 442-458.

West, M. (1953). *A General Service List of English Words*. London, United Kingdom: Longman.

Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. West Sussex, United Kingdom: John Wiley & Sons Inc.

Wilkins, D. A. (1972). *Linguistics in Language Teaching*. London, United Kingdom: Arnold.

Willis, D. (1990). *The Lexical Syllabus: A New Approach to Language Teaching*. London, United Kingdom: HarperCollins.

Yang, M. (2015). A nursing academic word list. *English for Specific Purposes*, *37*(Issue Number), 27-38.

Appendix A: OPEC Word List (OWL)

TYPE	FREQ	RANGE
ACCELERATE	38	18
ACCELERATED	94	31
ACCELERATING	37	23
ACCELERATION	34	20
ADVERSELY	25	19
AFRAMAX	262	36
AMID	861	38
AMPLE	115	33
ANALYTICS	409	16
ANNOUNCED	158	35
ANNOUNCEMENT	32	23
ARBITRAGE	303	36
ASSETS	44	14
AVIATION	66	14
BACKWARDATION	63	24
BARGES	47	36
BEARISH	278	37
BEARISHNESS	52	19
BENCHMARK	177	37
BIOFUEL	78	22
BIOFUELS	206	25
BITUMEN	71	10
BLEND	50	26
BOOST	91	33
BOOSTED	119	35
BOOSTING	49	22
BRASILEIRO	29	14
BREAKDOWN	49	18
BRENT	1126	38
BUDGET	111	27
BULLISH	249	36
BULLISHNESS	52	20
BUNKER	72	29
BURDEN	29	13
BUREAU	114	30
BUZZARD	44	19
CAPEX	65	15
CARBON	80	10
CARGO	40	19
CARGOES	245	37
CFTC	205	36
CHARTERERS	50	20
CHARTERING	107	36
CIRCULATION	30	30
CLIMATE	95	17
COMPETITIVE	44	17
CONDENSATE	152	20
CONSECUTIVE	364	35
CONTANGO	174	35
CONTINUATION	98	33

CONTRACTION	249	34
COPYRIGHT	48	16
CRACKER	36	23
CRISIS	143	31
CRUDE	6319	40
CUMULATIVE	102	20
CURB	35	17
CUSHING	69	29
DECELERATED	30	16
DECELERATION	85	26
DEFERRED	20	10
DEFICIT	209	39
DEFLATION	28	17
DEPLETING	32	15
DEPLETION	20	12
DEPRECIATED	39	22
DEPRECIATION	59	21
DESTINATION	24	13
DESTINATIONS	67	28
DESULPHURIZATION	82	10
DETERIORATING	32	15
DETERIORATION	57	22
DIFFERENTIAL	32	17
DIFFERENTIALS	166	39
DISCLAIMER	20	15
DISCOUNT	246	38
DISRUPTION	36	17
DISRUPTIONS	94	32
DISTILLATE	625	38
DISTILLATES	267	37
DOWNSIDE	176	32
DOWNSTREAM	64	17
DOWNTURN	46	20
DOWNWARD	762	38
DRILLING	139	18
ECONOM	49	15
ELEVATED	23	11
EMISSIONS	185	13
EQUATORIAL	57	37
EQUITY	106	26
EXCL	63	16
EXERTED	58	27
EXERTING	30	19
EXPANSIONS	59	16
EXPENDITURE	93	30
EXPENDITURES	66	23
FEED	30	13
FEEDSTOCK	77	22
FIRMED	57	29
FIRMER	55	24
FISCAL	198	37
FIXTURES	305	35

FLIPPED	36	19
FORECAST	1720	39
FORESEEN	57	22
FOSTER	21	11
FRAGILE	20	16
FREIGHT	1358	39
FUEL	3082	39
FUELS	457	36
GASOLINE	2679	39
GEOPOLITICAL	142	34
GRAPH	2143	36
GROSS	81	27
HAVER	409	16
HEADLINE	30	19
HEFTY	100	31
HEMISPHERE	40	19
HUGE	44	21
HUGHES	107	35
HURRICANE	53	12
HURRICANES	64	14
IMBALANCE	26	13
IMPORT	156	36
IMPORTS	2314	39
INDICES	105	28
INFLATED	38	16
INFLATION	940	38
INFLATIONARY	68	28
INFLOW	57	24
INFLOWS	93	29
INVENTORIES	821	37
INVENTORY	107	39
JET	872	38
LIQUEFIED	55	23
LIQUIDITY	62	26
LIVESTOCKS	33	19
LONG-TERM	295	22
LUBRICATING	20	18
MACROECONOMIC	52	21
MAGNITUDE	43	20
MASSIVE	55	27
MERCHANDISE	52	36
MISCELLANEOUS	73	36
MOMENTUM	317	39
MONETARY	249	39
MORTGAGE	53	19
NON-COMMERCIALS	98	21
NON-CONVENTIONAL	210	37
ONLAND	161	36
ONSHORE	43	14
ONSTREAM	82	18
OUTAGES	157	31
OUTRIGHT	38	18

PACE	396	39
PARAFFIN	34	19
PAYROLLS	40	24
PEAK	206	38
PEAKED	34	18
PEAKING	55	25
PEMEX	43	14
PERMIAN	59	15
PETROBRAS	62	17
PETROLEUM	780	40
PIPELINE	241	36
PIPELINES	69	25
PLATFORM	71	20
PLATTS	62	18
PLUNGED	75	30
PORT	288	29
PORTS	106	27
PREM	101	33
PREMIUM	408	39
PROPANE	103	21
PROPYLENE	71	21
RAMPING	32	16
REBOUND	100	36
REBOUNDED	64	26
RECESSION	204	30
REFERENCE	628	39
REFINERIES	477	38
REFINERS	302	38
REFINERY	2093	39
REFORM	203	17
REFORMS	92	17
RESIDUAL	416	38
RESUMPTION	20	16
RETAIL	276	38
REVIVED	69	21
RISEN	66	30
ROTTERDAM	149	38
ROUNDING	281	36
RUPEE	25	14
RURAL	43	19
SEASONAL	516	38
SEASONALITY	21	15
SEASONALLY	93	31
SECRETARIAT	583	39
SENTIMENT	602	36
SHALE	115	19
SHORT-TERM	86	27
SHRANK	54	19
SHRINK	31	20
SHRINKING	39	22
SKEWED	54	16
SLOWDOWN	261	37

SLUGGISH	88	33
SOARED	30	18
SOYBEAN	98	27
SPECULATION	29	11
SPECULATIVE	189	33
STAGNANT	47	21
STARTUPS	30	15
STERLING	84	36
STIMULUS	119	25
STOCK-DRAWS	48	17
STORAGE	149	30
SUBDUED	27	16
SUEZMAX	289	36
SULPHUR	197	17
SURGE	99	31
SURGING	25	15
SURPLUS	318	37
SWAP	48	15
SWITCH	64	24
TANKER	567	38
TAXATION	32	14
TERRITORY	79	25
THRESHOLD	42	25
THROUGHPUT	149	38
THROUGHPUTS	149	36
TONNAGE	292	36
TONNES	100	28
TRANSATLANTIC	71	30
TRILLION	83	22
TRINIDAD	47	37
TURMOIL	44	15
TURNAROUND	33	22
UNDERPINNED	23	18
UNLEADED	203	37
UNPLANNED	72	27
UPCOMING	52	24
UPSTREAM	91	27
UPTICK	39	14
USAGE	60	23
VERSUS	553	38
VOLATILITY	161	38
WARMER	41	22
WEAKENING	189	36
WHATSOEVER	31	16
WHOLESALE	52	26
WORLDSCALE	156	36
WORLDWIDE	73	32
WORSENED	21	15
WORSENING	47	18