AN INVESTIGATION OF ACADEMIC WORD LIST (AWL) AND GENERAL SERVICE LIST (GSL) IN HARD SCIENCES’ RESEARCH ARTICLES: MEETING THE RISING DEMANDS IN ESP

Ali Jahangard¹, Amin Rajabi Kondlaji², Karim Khalaji³, Ali Cholmaghani⁴

¹Sharif University of Technology, Iran, E-Mail: ghasemi.aliasghar@gmail.com
²Sharif University of Technology, Iran, E-Mail: amin2007_rajabi@yahoo.com
³Iran, E-Mail: karim_khalaji@yahoo.com
⁴Sharif University of Technology, Iran, E-Mail: ali.cholmaghani@gmail.com

Abstract. Taking the dynamic nature of English language in Academic Context into account, there is a consistent need in ESP and EAP for specific corpora, to which the students can refer to check their language productions and several other uses by the professionals and instructors. The present study aims to have an evaluative view over the two well-known group of words namely GSL and AWL. To this end, a Hard Sciences corpus of about 4 million words was developed from research article texts in six main fields of Energy Engineering, Electrical Engineering, Mechanic Engineering, Computer Sciences, Chemistry and Physics. Then, it was further analyzed by a web application which was also designed by the researchers. The Coverage of Coxhead’s (2000) AWL and also West (1952) were both analyzed and the results of the analysis showed that the developed corpus is a valid one and can be used to develop a new academic Word List for application in the respective fields.

Key words: Hard sciences corpus, GSL and AWL coverage, web application

1. INTRODUCTION

Recently there has been an increasing need to master and accept English as an international language. The field of science had not been an exception as there has been a growing interest on the part of university professors and students to access or to write and to publish scientific articles in English. This imposes a great demand on English for Specific Purpose (ESP) and English for Academic Purpose (EAP) teachers to develop materials that are in line with the specific needs of different people in different fields or various branches of science.

Considering the importance of vocabulary in language learning and especially in an academic setting, a very critical point to decide in advance is which vocabularies are worth teaching in the limited hours of class time. (Cohen, Glasman, Rosenbaum- Cohen, Ferrara, & Fine, 1988) state that academic vocabulary may trigger serious difficulties and problems for language learners since they are not familiar with academic words as they are with the technical vocabularies of their own field of study, nor as with general service vocabulary items which have high frequency of occurrence (see also Worthington & Nation, 1996; Xue & Nation, 1984).
While the corpus of previous academic wordlist was a combination of texts from different majors, the sub-corpora were not of a considerable amount. On the other hand, the previous corpus used large bodies of text, i.e. books that may in some way represent certain editors or writers’ preferred way of writing which in some ways may lead to some bias in the corpus. In the present paper the researchers tried to include more words in the corpus for the fields of science and to include short materials in corpus in contrast to corpus of previous studies.

The present study aims to answer the following questions:

1. What percentage of the words in the Hard Sciences Corpus does the AWL cover?
2. What percentage of the words in the Hard Sciences Corpus does the GSL cover?

2. REVIEW OF LITERATURE

Renouf (1987) defines a corpus as “a collection of texts, of the written or spoken word, which is stored and processed on computer for the purpose of linguistic research” (p. 1). Language analysts and teachers may benefit from corpora through gaining new visions into features of their linguistic structures and language use (Renouf, 1987). It can also help language learners succeed in language learning by providing materials that are really beneficial and useful for them. Mudraya (2006) states that with the arrival of corpora, the roles of teachers and students are somehow changed, contrary to their previous roles, teachers are now the administrators of research and learners are now researchers responsible for their own learning activities, utilizing corpus data as their major source of language learning.

The General Service List (GSL) (West, 1953) which was abstracted from a corpus of 5 million words to meet the needs of ESL and EFL learners, includes the most common 2000 word families in English language. The factors that West adopted to select the 2000 most frequent words were, frequency, stylistic level, coverage of useful concepts, and ease of learning.

There have been several criticisms against GSL. Engels (1968) states that one of the shortcomings of GSL was the size of its corpus, since it could be larger for a claim of that sheer size. Richards (1974) mentions that the age of GSL corpus is old, it cites the age matter as one of the short-comings of the GSL corpus. Some other studies such as Hwang (1989) call for the need for a revision in the GSL due to the dynamic nature of English as any other language. In spite of the mentioned criticisms, the GSL makes up to 90% of fiction texts as Hirsh (1993) mentions it. And also 75% of nonfiction texts based on Hwang (1989), and up to 76% of the Academic Corpus (Coxhead, 1998).

Academic Word List (AWL) (Coxhead, 2000) was the second popular list developed to cover the shortcomings of GSL especially with regard to the needs of the academics. One of those inadequacies was that GSL did not account for the needs of students in different Academic Fields. The corpus for AWL included 3.5 million words which was pooled from several sources including textbooks, and some texts form other corpora such as Wellington Corpus of Written English (Bauer, 1993), Learned and Scientific section of the Brown Corpus (Francis & Kucera, 1982), Learned and Scientific section of the Lancaster-Oslo/Bergen (LOB) Corpus (Johansson, 1978), and a few other ready corpora.

Based on Coxhead (2012), the main criteria for the word selection in AWL were as follows: the exclusion of 2000 word families in GSL in the word count, since it was
assumed that AWL was aimed at Academic uses, not for General ones. Also the words selected for AWL were to possess the three conditions of frequency (at least 100 times), range (at least in 15 subject areas) and uniformity (at least four times in the Disciplines).

A critical question is about validity of GSL and the AWL, i.e. how their coverage is in different texts and fields. Using some previous studies, Nation and Waring (1997) stated that, on average, GSL provided a coverage of 82%. The range of coverage however varied from 78% to 92% in different types of written texts.

According to Coxhead (2000), the AWL on the other hand provides an average coverage of 10% in the Academic Corpus. The coverage of AWL was also found to be different on different subcorpora. For example, Coxhead (2002) found that AWL coverage in the corpora of Commerce is about 12%, while it ranges from 9.0% to 9.4% for other three academic areas of Arts, Law, and Science. So Coxhead (2002) states, the AWL profits the commerce students more than the others. The actual potential of the lists, however, can be viewed when these two are used together. Combined with AWL, GSL provide a coverage of 88.8% for the commerce subcorpus in the Academic Corpus, which is the highest in all four academic subjects (Coxhead, 2002). When assessed on the whole Academic Corpus, i.e. including all the sub-corpora (3.5 million words), the two lists give 86.1% coverage (Coxhead, 2000).

In the past few years there have been attempts to prepare wordlists that are more subject specific and are made for the needs of the learners in their academic area. This trend has been in line with Hyland and Tse (2007) who claimed that some of the word families in AWL have diverse meanings in different academic fields. Of the early attempts to provide a field-specific academic word list in the last decade was done by Coxhead and Hirsh (2007) who tried to make a list of the most frequent words in pilot sciences. A year later (in 2008) Wang, Liang and Ge did an study on Medical sciences Research articles, using a corpus of about 1,093,011 running words, and made a list entitled ‘A Medical Academic Word list (AMAWL)’, which included the frequent words in medical sciences. Another study was done by Ward (2009), who prepared a list of the vocabulary for foundation students. He considered three important criteria in his study, first: the list presupposed little or no grammatical knowledge; second, the list could be used by learners with low levels of proficiency, and finally it was applicable in all the engineering disciplines. Furthermore, Chung (2009) did a study using a corpus of 579,849 running words form the newspapers and their primary purpose was to come up with a list of the words that were most applicable in reading newspapers. Their study rendered 588 word families that were most applicable in reading newspapers.

As there have been different explanations and claims about GSL and AWL, the present study aims to investigate the amount of AWL and GSL coverage in the realm of Hard Sciences.

3. Method

3.1. Corpus

The main corpus of the present study was comprised of 600 articles from 6 main engineering and science fields, namely, Energy Engineering, Electrical Engineering, Mechanic Engineering, Computer Sciences, Chemistry and Physics. The articles were also selected from technical journals of each field so as to include different standards and
styles of writing as a means to increase reliability of the corpus and improve the generalizability of findings. List of journals from which the articles were selected and used as the corpus of the study are listed in Appendix A. A total number of 4 million words made the whole corpus. The general service list and (GSL) and academic word list (AWL) was also needed to make the required research questions possible to answer.

3.2. The data analysis

To provide the required information about the degree and percent of Coverage of academic word list and general service list, a web application was developed by the researchers (available at http://takvaj.ir/main/aw_analyzer.php). This software calculates the degree of overlapping between the selected corpus and AWL or GSL. The procedure was run for the whole corpus of the study to achieve the percentage of AWL and GSL coverage in the whole corpus that include the data from the total selected fields of science as well as for each fields of science to see the coverage of AWL and GSL in each of them separately and also to make comparisons on the use of AWL and GSL among fields of study.

4. DATA ANALYSIS

The Hard Sciences Corpus developed in this study was analyzed using the web application developed by the researchers available at (http://takvaj.ir/main/aw_analyzer.php) to find the coverage of AWL and GSL. The results of the analysis are presented below:

Table 1 The Percentage of AWL and GSL Coverage in the Corpus and Sub-corpora, with Corpus Statistics

<table>
<thead>
<tr>
<th>Field of science</th>
<th>Corpus words</th>
<th>AWL frequency</th>
<th>GSL frequency</th>
<th>Percentage of AWL use</th>
<th>Percentage of GSL use</th>
</tr>
</thead>
<tbody>
<tr>
<td>energy engineering</td>
<td>931231</td>
<td>15273</td>
<td>730857</td>
<td>1.6</td>
<td>78.4</td>
</tr>
<tr>
<td>electrical engineering</td>
<td>636090</td>
<td>8774</td>
<td>536330</td>
<td>1.3</td>
<td>84.3</td>
</tr>
<tr>
<td>mechanic engineering</td>
<td>843946</td>
<td>11066</td>
<td>673227</td>
<td>1.3</td>
<td>79.7</td>
</tr>
<tr>
<td>computer sciences chemistry</td>
<td>698641</td>
<td>12018</td>
<td>555352</td>
<td>1.7</td>
<td>79.4</td>
</tr>
<tr>
<td>physics</td>
<td>493544</td>
<td>7116</td>
<td>382663</td>
<td>1.4</td>
<td>77.5</td>
</tr>
<tr>
<td>Total</td>
<td>4053313</td>
<td>59457</td>
<td>3223705</td>
<td>1.4</td>
<td>79.5</td>
</tr>
</tbody>
</table>
As is evident from Table 1 and Figure 1, GSL covers a wide range of corpus of the study. On average 79.5% of the whole content words in the corpus of the study turned out to be included in the GSL. But there are also slight variations in the percent of GSL coverage in the corpus of each selected field of study. Papers in electronic engineering, physics, mechanical engineering, computer sciences, energy engineering and chemistry had respectively the highest use of GSL in proportion to their corpus, selected for the purpose of the present study.

Fig. 1 Percent of GSL coverage in different fields

Fig. 2 Percent of AWL coverage in the corpus and subcorpora
As it is presented in Table 1 and graphically in Figure 2, on average 1.4 percent of the whole corpus was comprised of academic word list (AWL). At first glance this percent might seem to be low, but high level of coverage of words in the corpus (nearly 80) in one hand and variation in the technical vocabulary used in papers and articles of different fields of sciences on the other hand might account for it. Regarding GSL we can also see variations of coverage in the corpus at different fields by AWL, though very limited. Computer sciences, energy engineering, chemistry, electronic engineering, mechanical engineering and physics papers were respectively found to have more coverage of AWL in corpus relative to the proportion of the words in the corpus.

5. DISCUSSION

The present study was an attempt to compensate for the shortcomings in the previous Academic corpora (e.g. inadequate corpus size), and also the famous wordlists (i.e. AWL and GSL) which were derived from those corpora. The most influential academic word list to date, is AWL developed by Coxhead (2000). The list was the inspiration for many textbook developers, instructors and also education programmers throughout the world. Considering the fact that the corpus from which a certain list is drawn affects the validity and applicability of the list, and whether the list can be employed in ESP education for a certain academic genre. Coxhead’s AWL was made using the 3.5-million-word-corpus developed at the time. The corpus for AWL was made using different pieces of texts from several academic fields including fields from both hard and soft sciences.

About fifty years before the AWL, with the advent of computer technology and the ability to process large bodies of information, West (1953) compiled a 5-million-Corpus from which he made GSL. GSL was not considered an academic list due to its corpus, which did not contain academic texts.

Assuming the above argument one can propose that the more specific a corpus is, the more applicable the list drawn from that corpus can be. Some field-specific vocabulary lists have been developed in the past few years using more specialized corpora, to make for the needs of particular learners (Wang, Liang, & Ge, 2008; Coxhead & Hirsh, 2007; Ward, 2009; Chung, 2009).

Unlike the corpus developed by Coxhead (2000), the researchers used smaller pieces of texts from a variety of fields, since the longer texts, e.g. books, may have biased the corpus in some way. The 4-million-word-corpus was made from 600 research articles from six main fields under the hard sciences category, namely Energy Engineering, Electrical Engineering, Mechanic Engineering, Computer Sciences, Chemistry and Physics. Development of Hard Sciences corpus through this study exploring the coverage of AWL and GSL word families in them was the core purpose of this study.

The first question that was pose was, what percentage of the words in the Hard Sciences Corpus the AWL covers. After the data analysis it was revealed that AWL covered 16 percent of the whole words in the corpus. This is a higher percentage compared to the study done by Hyland & Tse (2007), in which they measured 10 percent coverage of AWL in a 3.292-million-words-corpus made from the writings of sciences, engineering, and social sciences students and professionals.
Chen & Ge (2007) made a 190425-word corpus of Medical Sciences Research articles, and found that AWL covered 10.073 percent of the words in their corpus. This study was the only attempt to create a corpus only with Research Articles.

There have been several studies exploring the coverage of AWL in their respective corpora, which have reported various percentages of coverage ranging from 1.4 percent to 11.51 percent (see Coxhead 2000a & 2000b; Cobb & Horst, 2004; Konstantakis, 2007; Coxhead & Hirsh 2007; Ward, 2009; Martínez, Beck, & Panza 2009; Vongpumivitch, Huang, & Chang, 2009; Li & Qian, 2010; Coxhead, Stevens, & Tinkle, 2010).

The study that has reported the widest coverage of AWL in its corpus (about 11.51 percent), was Konstantakis (2007), who compiled a corpus of about one million words form Business Research Articles. The present corpus for this study, having 16 percent coverage of AWL is one of the most valid corpora up to now.

The second question of high importance for this study was related to GSL. The researchers tried to find the coverage of GSL in the Hard Sciences corpus. After the analyses it was revealed that GSL covered an average of about 80 percent of the sub-corpora, and about 78.4 percent of the whole hard sciences corpus. 80 percent coverage of GSL in hard sciences Corpus is slightly higher than the amount proposed by Coxhead (1998) (cited in Coxhead, 2000), that was 76 percent coverage for GSL in academic corpus. This also confirms that the Hard Sciences Corpus developed in this study is a valid corpus and is in line with other corpora developed earlier. On the other hand considering the higher coverage of AWL in Hard Sciences corpus (i.e. 16 percent), the corpus seems to be more useful than the previous similar corpora for academic purposes.

6. CONCLUSION

The present study was an attempt to develop, evaluate, and validate a corpus made from six main hard sciences fields, namely Energy Engineering, Electrical Engineering, Mechanic Engineering, Computer Sciences, Chemistry and Physics. To analyze the data, a web application was developed by the researchers.

The results of the analyses showed that the Hard Sciences Corpus developed by the researchers of the present study has the highest coverage of AWL and GSL among all the previous corpora, which makes it a more valid corpus. Another distinctive feature of the Hard Sciences Corpus was that it was made up of only Research Articles. The corpus will have the following applications: The researchers will develop the web-application so that it can calculate the concordance of any given piece of text with any of the sub-corpora or the total corpus. The concordance calculation can be very helpful in ESP writing courses from two points: first, the concordance system can be used by the students to evaluate their own writing and second, teachers can use the system as a grading tool for their students’ compositions. Furthermore, the Corpus can be used by the journal editors to check the received articles for publication for their proximity to the academic writing styles of the field.

The researchers suggest the following topics for further research: a. What Non-AWL content words are frequent in the corpus? The exploration of the above question can lead to a hard sciences word list which contains the word families other than the ones in AWL and also GSL. The study can be replicated for soft sciences and the results can be compared. However, there were certain limitations to the study which need to be considered
in the use of the findings. The researchers included only six of the majors in hard sciences, and for future further studies the number of the sub-corpora can be extended to include more fields in hard sciences.

REFERENCES


**APPENDIX A**

**List of Journals from which the articles were extracted from**

**Energy engineering**

1) *Progress in Energy and Combustion Science*
2) *IEEE Transactions on Energy Conversion*
3) *Applied Energy*
4) *Energy*
5) *Electric Power Systems Research*
6) *International Journal of Electrical Power and Energy Systems*
7) *Energy Conversion and Management*
8) *Applied Thermal Engineering*
9) *Energy and Fuels*
10) *Progress in Nuclear Energy*
11) Journal of Natural Gas Science and Engineering
12) Power System Technology
13) Heat Transfer Engineering
14) Journal of Natural Gas Chemistry
15) Oil and Gas Science and Technology
16) Journal of Petroleum Geology

**Electrical engineering**

1) Systems and Control Letters
2) International Journal of Robust and Nonlinear Control Pattern Recognition
3) IEEE Network
4) IEEE Microwave and Wireless Components Letters
5) Journal of Electronic Materials
6) IEEE Transactions on Consumer Electronics
7) Journal of Micromechanics and Microengineering
8) Optics and Lasers in Engineering
9) Solid-State Electronics
10) Iranian Journal of Electrical and Electronic Engineering
11) Sensors
12) IET Communications
13) Journal of Electrostatics
14) Journal of Power Electronics
15) Iranian Journal of Electrical and Computer Engineering
16) European Transactions on Electrical Power
17) Radioengineering
18) AEU - International Journal of Electronics and Communications
19) Electrical Engineering
20) Applied Artificial Intelligence

**Mechanic engineering**

1) International Journal of Plasticity
2) Journal of the Mechanics and Physics of Solids\ International Journal of Thermal Sciences
4) International Journal of Solids and Structures
5) Journal of Sound and Vibration
6) Engineering Fracture Mechanics
7) Mechanism and Machine Theory
8) Tribology Letters
9) International Journal of Mechanical Sciences
10) Journal of Hydraulic Engineering
11) Journal of Tribology
12) Open Mechanics Journal
13) Tribology - Materials, Surfaces and Interfaces
14) Annals of Solid and Structural Mechanics
15) Advances in Mechanical Engineering
16) Industrial Lubrication and Tribology
17) Journal of Computational and Nonlinear Dynamics
18) Fluid Dynamics Research
19) International Journal of Automation technology

**Computer sciences**

1) Journal of Scientific Computing
2) IEEE Communications Letters
3) Computer Networks
4) World Wide Web
5) Mobile Networks and Applications
6) Iranian Journal of Information Processing Management
7) Iranian Journal of Fuzzy Systems
8) Journal of Informetrics
9) Information Sciences
10) Computer Methods in Applied Mechanics and Engineering
11) International Journal of Sensor networks
12) Journal of Web Librananship
13) Parallel Computing
14) Journal of Combinatorial Optimization
15) International Journal of Automation and Computing
16) Networks and Heterogeneous Media
17) Foundations and Trends in Human-Computer Interaction
18) Information Fusion
19) Medical Image Analysis
20) Natural Computing

**Chemistry**

1) Chemical Reviews
2) Accounts of Chemical Research
3) Nature Chemistry
4) Chemical Science
5) Topics in Current Chemistry
6) Chemical Communications
7) Chemistry - A European Journal
8) Green Chemistry
9) Chemical Record
10) Journal of Chemistry and Chemical Engineering
11) Molecular Diversity
12) Biointerphases
13) Cereal Chemistry
14) Chemistry Letters
15) Environmental Chemistry Letters
16) Journal of the Iranian Chemical Society
17) Journal of Chemical Sciences
18) Chimia
19) Symmetry
Physics

1) Reviews of Modern Physics
2) Nature Physics
3) Physics Reports
4) Reports on Progress in Physics
5) Progress in Optics
6) New Journal of Physics
7) Laser Physics Letters
8) Annals of Physics
9) Applied Physics Letters
10) Journal of Applied Physics
11) Journal of Low Temperature Physics
12) Quantum Information and Computation
13) Chaos
14) Laser and Particle Beams
15) Foundations of Physics
16) Review of Scientific Instruments
17) Laser Physics
18) European Physical Journal D