6-15-2021 10:00 AM

Ontario High School Science Word List (OHSWL)

Mohamed Mahfouz, *The University of Western Ontario*

Supervisor: Faez, Farahnaz, *The University of Western Ontario*
Co-Supervisor: Boers, Frank, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Arts degree in Education

Abstract

This research aims to explain the development of an Ontario High School Science Corpus and subsequently an Ontario High School Science Word List (OHSWL). The OHSWL is a list of the most frequent technical words in the Ontario high school science curriculum. The science corpus was compiled from Ontario science textbooks and public written lecture material. A total of 803 lemmas were identified as part of the OHSWL. The coverage of the OHSWL in the science corpus vs non-science corpus is 7.79% and 1.52% respectively. The high frequency vocabulary (top 3,000 words) of the Corpus of Contemporary American English (COCA) and OHSWL had a coverage of 85.44% and 75.67% in the science corpus compared to the non-science corpus. With an approximately 10% difference in coverage, the OHSWL proves to be a significant source of vocabulary for an Ontario science learner. While coverage of the first and second 1,000 words of the COCA were greater in the science corpus compared to the OHSWL, coverage of the third 1,000 words was only marginally greater. Therefore, past the top 3,000 words of the COCA, the greatest value for someone learning the Ontario science curriculum is achieved by knowing the OHSWL. This corpus-based study has the potential of helping students in Ontario, regardless of whether they speak English as their first language or not.

Summary for Lay Audience

This research aims to explain the development of an Ontario High School Science Corpus and subsequently an Ontario High School Science Word List (OHSWL). A Corpus is "a collection of texts that is designed to be representative of some aspect of language" (Webb & Nation, 2017). The OHSWL is a list of the most frequent technical words in the Ontario high school science curriculum. The science corpus was compiled from Ontario science textbooks and public written lecture material. A total of 803 lemmas were identified as part of the OHSWL. A lemma is made up of the headword and its inflection. For example, the headword "add" would have its inflections as "adds", "adding" and "added". (Webb & Nation, 2017). The coverage of the OHSWL in the science corpus vs non-science corpus is 7.79% and 1.52% respectively. The high frequency vocabulary (top 3,000 words) of the Corpus of Contemporary American English (COCA) and OHSWL had a coverage of 85.44% and 75.67% in the science corpus compared to the non-science corpus. With an approximately 10% difference in coverage, the OHSWL proves to be a significant source of vocabulary for an Ontario science learner. While coverage of the first and second 1,000 words of the COCA (1 to 1,000 and 1,001 to 2,000) were greater in the science corpus compared to the OHSWL, coverage of the third 1,000 words was only marginally greater. Therefore, past the top 3,000 words of the COCA, the greatest value for someone learning the Ontario science curriculum is achieved by knowing the OHSWL. This corpus-based study has the potential of helping students in Ontario, regardless of whether they speak English as their first language or not. By teachers implementing the use of the OHSWL in their classrooms, beginning with students in grade 7 up to grade 12, understanding the scientific jargon will no longer be as difficult. Students will be able to focus on applying their knowledge rather than memorizing terminology.

Acknowledgments

I would like to start off by thanking my wonderful supervisor, Dr. Faez, for always being so kind and amazing. I did plenty of research when choosing my supervisor, and Dr. Faez was the one individual that stood out to me because everyone spoke very highly of her. Whenever we had meetings, there would always be a smile on her face. Because of Dr Faez's positive nature, she has always been so helpful to me, motivating me to work to the best of my abilities. I have the utmost respect for her, and I feel very privileged to have been her student.

I would also like to thank my co-supervisor, Dr. Boers. From the first meeting until the last, I can confidently say he has helped immensely. Dr. Boers questioned my thought process and research constantly to ensure that I understood the purpose of my work. He pushed me to work my hardest to ensure perfection, and I am grateful for that.

To every single professor that has been a part of my masters journey, thank you. Thank you for helping me grow as a human being. Thank you for taking the time and effort to support me and help reach my goal in my research.

Finally, I am extremely thankful for the support of my family to embark on this journey. Without them, I would not have been constantly reminded that I have come this far, and I can go even further in my educational journey. There are no words to describe my appreciation for all that they have done for me, and I know that I am fortunate to have such a wonderful family pushing me to be better.

Table of Contents

List of Appendices

Chapter 1 – Introduction

1.1 Introduction

Since the early 2000s, an important area of research has been the development of word lists. A word list is a collection of the most frequent words in a specified field or area. Examples of word lists would be the most frequent English words, the most frequent academic words, etc. It needs to be understood that the word list is the target vocabulary, and it needs to be taught to the learner rather than given with the expectation that they will memorize it. Since word lists are lacking in context, the educational experience is driven by the educator. Whether the goal is to help the learner, guide the course designer, provide a resource tool for educators, or all the above, mastery of the target vocabulary is an indispensable component of academic success (Nagy & Townsend, 2012).

As an ESL student, learning English was undoubtedly hard, but learning science was even more difficult. Having to understand English well enough to understand the science teacher, as well as learn the science specific terms was a challenge. Speaking from experience, an ESL student who does not understand the scientific jargon will opt to memorize everything. They will find the task too challenging due to the amount of vocabulary and lack of understanding on which words are important and which ones are not. While the students may choose to read the book to understand the material, reading an entire textbook, translating it, and memorizing the unknown vocabulary is unrealistic. From an ESL student's perspective, learning science feels like learning a third language on top of English. ESL learners are still at a stage where they are learning the most common English words because the value it has is much greater than learning words pertaining to a specific field (Schmitt & Schmitt, 2014). Science teachers have an expectation from

the students to learn scientific jargon, which tends to not be as commonly used in our daily lives (Rolls & Rodgers, 2017). For the student to be expected to start learning words that are not commonly used in the English language, it becomes more difficult and they do not benefit as much (Cobb, 2007). This is not to say that only ESL students will benefit from a science specific word list. The scientific jargon itself may be novel to both ESL students as well as native English speakers.

Having gone through ESL science and now having become a high school science teacher, it is still difficult to determine which words are essential and which ones are not. Although some teachers only use the end of unit questions in the textbook to provide the students with extra practice problems, some teachers completely neglect the use of any textbook. This leads to a lack of standardization in terminology across the different educators. While teachers might think that they know which words are necessary and need to be prioritized, their ranking is simply based on intuition and not facts (Alderson, 2007). Although learners currently have access to Coxhead's (2000) Academic Word List and Gardner and Davies's (2014) Academic Vocabulary List, these word lists are general and have been questioned on their effectiveness in a specific field (e.g., medical field, science field, nursing field, etc.). Due to lexical differences, there have been arguments explaining the complexities of a discipline and the need for field specific word lists (Hyland, 2002, 2006).

The goal of this study is to create an Ontario high school science corpus and through that, create an Ontario High School Science Word List (OHSWL). By no means does this word list aim to be a study guide to get 100% in science. Similar to other word lists, it is meant to be used as a guide for learners, teachers, and course designers. It is meant to be used as a tool to help the

educator with decision making on the curriculum and help the learner prioritize their time on the more high-frequent science words.

1.2 Research Questions

The questions to be answered by this research are:

1) Which words (lemmas) occur most frequently across a wide range of the Ontario high school science material, but are not among the 3,000 most frequent words in the COCA?
2) What percentage of the words (lemmas) in the Science corpus does the OHSWL cover?
3) What percentage of the words (lemmas) in the non-science corpus does the OHSWL cover?
4) What is the coverage of the top 3,000 words of the COCA in the science corpus?
5) What is the coverage of the top 3,000 words of the COCA in the non-science material?

Chapter 2 – Literature

"Different words have different values for learners… it is much more useful to know 'find', 'flower', and 'food' than it is to know 'fluctuate', 'foam', and 'fragrant'" (Webb & Nation, 2017). The value of the word is dependent on its frequency. A high frequency word tends to be used more often in communication compared to a lower frequency word, and therefore provides greater value to the learner (Webb & Nation, 2017). Having said that, there is no universal list where an

individual can study and become an expert in English regardless of the situation. The value of a word is different depending on the context and fields of study.

2.1 Background Information

Starting off, there are some necessary terms that must be defined to have a complete understanding of this study. The words to be explained are lemmas, word family, running words, and corpus. A lemma is made up of the headword and its inflection. For example, the headword "add" would have its inflections as "adds", "adding" and "added". (Webb & Nation, 2017). A word family consists of its headword, inflection, and derivation, such as assume, assumes/assumed, and unassuming, respectively (Webb & Nation, 2017). Unlike word families and lemmas, running words are the collection of every word in a text. For example, the sentence "assuming that she assumed…" contains 3-word families but 4 running words. One thing to note is that lemmas, word families, and running words are units of counting words. While researchers may define the words in their word lists as lemmas or word families, they must first compile a corpus to create their word list. A corpus (plural: corpora) is "a collection of texts that is designed to be representative of some aspect of language" (Webb & Nation, 2017). For example, a learner corpus is a corpus that represents the type of language used by learners of a second or foreign language (Webb & Nation, 2017). Moving forward, whenever there is mention of "high frequency words, or top 1,000 words", it is implied that it is word families and not just running words/lemmas (unless otherwise stated). The debate to be had is how many words should be considered high frequency. Information regarding the frequency of the word can "help course designers decide what words to include in a language course" (Nation, 2016). This in turn will guide the instructors on what to teach the students. Without guidance, teachers will not have the knowledge required to help students be able to communicate and/or comprehend effectively (Webb & Nation, 2017).

While some say that it should be the top 1,000 words (Dang & Webb, 2016), others want to include words up to the 3,000 level (Engels, 1968; Schmitt & Schmitt, 2014). The argument between the number of high frequency words is based on the coverage level by those word lists. While knowing the top 1,000 words covers 85% of words encountered in TV programmes, knowing the top 3,000 words covers 95% of spoken vocabulary (Webb & Nation, 2017). Since studies have shown that for learners to understand speech, they need to know 95% of the words used (Van Zeeland & Schmitt, 2013), Schmitt & Schmitt (2014) feel that 3,000 words should be the classification for high frequency words. On the other hand, Webb & Nation (2017) as well as Dang & Webb (2016) feel that high frequency words should only be classified up to 1,000 words because "beyond the 1,000 word families, the relative value of knowing each additional set of 1,000 word families drops substantially" (Webb & Nation, 2017).

While there is an argument to be had about whether high frequency words should include the top 1,000 words or 3,000 words, the beneficiaries are those trying to learn English. There are other types of vocabulary, such as low frequency vocabulary, technical vocabulary, and academic vocabulary. According to Schmitt & Schmitt (2014), low frequency vocabulary is identified as any word that is below the 9,000 most frequent word family. Although low frequency words do not occur as much in the English language generally, they can occur more frequently within a special topic or area of study; that is considered to be technical vocabulary (Webb & Nation, 2017). For example, 'puck', 'rink', and 'arena' are much higher frequency when discussing hockey than discussing other sports or other topics (Webb & Nation, 2017). Academic vocabulary is frequent across different academic contexts, yet not frequent outside of academic text (Webb & Nation, 2017). To the course designer or the student learning, the significance of the words is dependent on the subject being learned. While high frequency words might be helpful to a student learning

English for everyday purposes, academic vocabulary would be more useful to a student in university, and technical vocabulary might be what is needed for a student learning a new sport. It is true that knowing 'flower' is useful for a student starting out how to learn English (Webb & Nation, 2017), but knowing 'participate' is more useful for a nursing student (Yang, 2015). Depending on the individual's needs and knowledge, different word lists are required to help the student achieve their learning goal.

2.2 Corpus of Contemporary American English (COCA)

In 1953, Michael West published the General Service List, which helps identify the most common words in the English Language. This list was created with the intention of helping English language learners and ESL teachers (West, 1953). While this list served its function, newer lists have come out since then. While it is not a list, it is possible to derive the most frequent words from the Corpus of Contemporary American English (COCA). With more than 130,000 people using it per month in over 140 countries, COCA is the most used corpus in the world (Davies, 2020). The COCA is a billion word corpus that is divided between spoken, fiction, popular magazines, newspapers, and academic journals (Davies, 2010, 2020). Every year, over 25 million words are added to the corpus, yet the balance between each genre remains approximately equal at 20%. Not only does the genre remain balanced, but so does the sub genre (ex, newspaper – sports or Academic – Medicine) (Davies, 2010).

The benefit of the COCA is that the user has the ability to see how the word is being used. For example, by looking at the word seldom, the COCA would identify it a as a formal word (used academically or in books) and not common in spoken discourse. It would also statistically demonstrate that the use of the word seldom is declining over time (Davies, 2020). As Davies (2020) mentions, if the corpus only told the user that seldom is used 87,000 times in a 17 billion

word corpus, “students would never know that if they used this word, they will sound like a 70-80 year old person and/or someone in a formal setting”. A similar example would be the word “Morph”. When looking at the trend, it can be seen that while in 1990, the word morph was never used, in 2008, the word morph was at an all time high in terms of its usage. Starting the year 2009, the word morph has been used less often and it is continuously declining. It can also be seen that in spoken discourse, the word morph is used 0.9 times per million words. This shows that the word morph is not often used in spoken language and is mostly seen in popular magazines (Davies, 2010). This kind of data can help an English language learner identify if the word in front of them is valuable to them in their context. The COCA is compiled with so much detail, that it is able to compare the most commonly used verbs between 2005 – 2009 and 1990 – 1994. From 1990 to 1994, the word multitask does not show up in the corpus; yet it shows up 39 times between 2005 – 2009. The word Moralize shows up 17 times between 1990 – 94, while it only shows up twice between 2005 – 09 (Davies, 2010). Understanding how often a verb is used in a specific time period is not the only thing that the COCA can do. It is also able to look at recent changes between morphology, syntax, semantics, and many more (Davies, 2010). All these features make the COCA a detailed and meticulously constructed corpus, which can serve the purpose of identifying the most used words in the English language generally, and in specific registers/genres.

2.3 Academic Word List

Awareness for academic language proficiency has been growing for students’ success in schools (Nagy & Townsend, 2012). Academic vocabulary plays an integral role in school success not only for non-native speakers of English, but also for students who speak English as their first language. The importance of academic vocabulary is seen at all grade levels, including primary, middle school, secondary, and higher education (Biemiller, 2010; Schmitt et al., 2011; Townsend

& Collins, 2009). It is for that reason general academic word lists were developed in the hopes of helping the learners. In the past, there were multiple attempts at creating word lists, but due to limitations with technology, the word lists were compiled by hand (Campion & Elley, 1971; Ghadessy, 1979; Lynn, 1973; Praninskas, 1972). As technology started improving, Xue and Nation (1984) created a University Word List (UWL) by editing and combining the four word lists mentioned above. Coxhead (2000) and other researchers found the issue with the UWL is that it lacked consistent selection principles and had many weaknesses. As well as that, the corpora used to create the UWL were small and "did not contain a wide and balanced range of topics" (Coxhead, 2000).

With the academic word lists that were available either outdated or not well organized, Coxhead (2000) felt that there is a need for a new academic word list "based on data gathered from a large well-designed corpus of academic English". From Coxhead's (2016) perspective, "The ultimate aim was to create a tool to guide decisions around learning, teaching, and curriculum and materials design". The Academic Word List (AWL) was developed from a corpus of 3.5 million running words (Coxhead, 2000). Since the Academic Word List is meant to be a representation of all the academic texts, all 3.5 million running words in the corpus were divided into four sub corpora (equally divided containing approximately 875,000 running words) of arts, commerce, law, and science (Coxhead, 2000).

For a corpus to be informative, not only must the material be representative of the language, but the organization, size, and word selection also needs to be considered. While compiling the AWL, Coxhead (2000) followed four main criteria to ensure an effective collection of text. First, in terms of representativeness of the language, both long and short texts needed to be included (Coxhead, 2000). Coxhead tried to maintain a balance between short, medium, and long texts in

all four sections. Second, the corpus was organized and divided into the four disciplines of art, commerce, law, and science. Those four disciplines were divided into 28 subject areas (Coxhead, 2000). Third, with regards to size, "a corpus should include millions of running words (tokens) to ensure that a very large sample of language is available" (Sinclair, 1991). Based on an arbitrary measure, Coxhead (2000) attempted to include four million words into the corpus. Unfortunately, due to time constraints, only 3.5 million running words were used in the corpus. The idea behind including a large sample of writing is to have texts written by multiple authors. By having multiple authors, idiosyncrasies in the corpus can be reduced (Pryzant et al., 2020). Finally, word selection was an issue due to the different definitions of a word (Coxhead, 2000). Coxhead decided to adopt the concept of word family as the unit for count, that is, the word's stem plus its affixed forms. The latter followed the level 6 definition of an affix by Bauer and Nation (1993). An affix includes its inflections as well as its most frequent, productive, and regular prefixes and suffixes (Bauer & Nation, 1993).

With such a large corpus, rather than compiling the word list by hand similar to some earlier studies (Campion & Elley, 1971; Ghadessy, 1979; Praninskas, 1972), Coxhead opted to use the corpus analysis programme RANGE (Heatley & Nation, 1996). This program "was used to count and sort the words in the academic corpus" (Coxhead, 2000). The words for the AWL were based on the three criteria of specialized occurrence, range, and frequency. For the first category of specialized occurrence, Coxhead used the General Service List (GSL) (West, 1953) to ensure that the word families included in the AWL were not part of the 2,000 most frequently occurring words in English (high frequency vocabulary) (Coxhead, 2000). The GSL was created with the intention of being a "core vocabulary for second language learners" (Browne, 2014). In terms of the second category of Range, the word families had to occur in every one of the four main sections of the

corpus at least 10 times and it also had to occur in a minimum of 15 of the 28 subject areas. The last category is the frequency, which demanded that the words must occur a minimum of 100 times in the Academic corpus (Coxhead, 2000).

The final product of the AWL consisted of 570-word families. While in academic text, the AWL covered 10% of running words, it only accounted for 1.4% of running words in a fiction collection of the same size (Coxhead, 2000). The difference in coverage is statistically significant and provides evidence that the AWL is a list containing academic words (Coxhead, 2000). "By highlighting the words that university students meet in a wide range of academic texts, the AWL shows learners with academic goals which words are most worth studying" (Coxhead, 2000). Years later, Coxhead (2016) believes that the AWL has proven to be an effective tool used by teachers, researchers, learners, textbook publishers and many more. The benefits of the AWL are not just limited to the actual word list developed. Also, the steps taken to develop the word list served as a useful example for others who wish to develop a specialized word list. Many of the steps used by Coxhead in the development of the AWL can be seen in Nation and Webb's (2011) "list of steps involved in making a word list" (Coxhead, 2016).

2.4 Academic Vocabulary List

When Coxhead's AWL was published in 2000, no one could have predicted that it would be the gold standard by which other word lists would try to measure up to; word lists such as the Medical Academic Word List (MAWL), Nursing Academic Word List (NAWL), Science Specific Technical Vocabulary in Fiction Fantasy Texts, and many more (Rolls & Rodgers, 2017; Wang et al., 2008; Yang, 2015). All these word lists have followed similar procedures to those of the AWL. Yet with so much praise to the AWL, Gardner and Davies (2014) seem to have an issue with its construction. As a result, they have created their own Academic Vocabulary List (AVL), where

they feel they have identified the correct academic words and using the correct methods. They also talk about the shortcomings of the AWL and why it is not doing what it set out to do.

According to Gardner and Davies (2014), the two main issues with the AWL are their use of word families when determining frequencies, and the relationship between the AWL and the GSL. When creating the AWL, Coxhead (2000) mentions that "For the creation of the AWL, a word family was defined as a stem plus all closely related affixed forms, as defined by Level 6 of Bauer and Nation's (1993) scale". This means that a word family often has members belonging to different parts of speech (e.g., nouns, verbs, adjectives and adverbs). For example, the verb proceeds (continues) and the noun proceeds (profits) are counted as the same unit. However, the issue with this is that these members of a word family do not always share the same core meaning (Nagy & Townsend, 2012). For example, there is a difference in meaning between react (respond), reactivation (to make something happen again), and reactor (a device or apparatus). The differences in their meanings are emphasised even more as the members of the word families cross over to other academic disciplines (Hyland & Tse, 2007). Since understanding one member of a word family does not guarantee comprehension of the other members, Gardner and Davies (2014) feel that rather than using word families, lemmas should be used. Lemmas are "words with a common stem, related by inflection only, and coming from the same part of speech" (Gardner & Davies, 2014). Using lemmas, the verb proceeds, and the noun proceeds would be considered as two distinct units.

Gardner and Davies are not the only researchers to disagree with the methodology of the AWL. In the creation of the AWL, the GSL was used to identify the high frequency words in the corpus. Unfortunately, due to the GSL being derived from an outdated corpus (from the 1990s), the GSL is no longer considered an accurate reflection of high-frequency English (Cobb, 2010;

Gardner & Davies, 2014; Neufeld et al., 2011; Schmitt & Schmitt, 2014). Learning from the criticism received by Coxhead (2000) regarding the use of an outdated corpus, Gardner and Davies (2014) used the COCA. One advantage that the AVL has over the AWL is that the collection of texts was derived from the COCA, which is material published in the USA and representative of a broad range of written academic materials, whereas the AWL was based mostly on material published in New Zealand. At the time of development of the AVL, the COCA consisted of over 425 million words. Out of 425 million words to choose from, the AVL Corpus contained more than 120 million words from nine academic disciplines, including education, humanities, history, social science, philosophy/religion and psychology, law and political science, science and technology, medicine and health, and business and finance (Gardner & Davies, 2014). The AVL is broadly divided between academic journals and academically oriented magazines. The academic journal corpus consisted of 85 million running words, while the academically oriented magazines contained approximately 31.5 million running words (Gardner & Davies, 2014).

With a focus on lemmas and not word families, Gardner and Davies (2014) eliminated general high frequency words (lemmas) from their list by specifying that the word (lemma) frequency must be at least 50% higher in the academic corpus compared to the non-academic portion of COCA. As well as that, the lemma must have a minimum of 20% of the expected frequency in at least seven of the nine academic disciplines. Similar to Coxhead's (2000) criteria, where the words needed to appear a minimum of 100 times, the 50% ratio and 20% range specified are nothing more than arbitrary numbers. Gardner and Davies (2014) observed what words were included/excluded in the list depending on the different percentages, and the above values were deemed to be sufficient. Any lemma entered into the AVL needed to have a minimum dispersion rate of 0.80. The dispersion value ranges from 0.01 (the word only occurs in extremely small part

of corpus) to 1.00 (there is a perfectly even dispersion across all parts of the corpus) (Gardner & Davies, 2014). Finally, Gardner and Davies (2014) stated that the word cannot occur more than three times the expected frequency (per million words) in any of the disciplines. For example, the word student occurs in the education discipline approximately 6.8 times the expected frequency. Although this is a frequent academic word, it is too specific to be included in the "core" list. Essentially, while having a ratio of greater than 50% is meant to exclude general high frequency words; having a minimum range of 20%, a dispersion value of greater than 0.80, and excluding any words with greater than three times the expected frequency will help exclude discipline specific and technical words (Gardner & Davies, 2014).

Using the above stated steps, the AVL was created consisting of 3000 lemmas. The coverage of the AVL on the academic material, newspaper, and fiction, respectively are 13.8%, 8%, and 3.4%. Clearly, the AVL coverage of academic material is more significant. Compared to the AWL, the AVL has a > 6% greater coverage of academic material (13.8% for AVL and 7.2% for AWL) (Gardner & Davies, 2014). Overall, Gardner and Davies (2014) set out to create an updated academic word list and based on their results, it seems that they have succeeded.

2.5 Field Specific Word Lists

The AWL and AVL are not the only word lists that have been created. There exists a Medical Academic Word List (MAWL), a Nursing Academic Word List (NAWL), a Science Fiction Fantasy Word List, and many more (Rolls & Rodgers, 2017; Wang et al., 2008; Yang, 2015). In the MAWL, approximately 1.1 million running words were used in the corpus from a collection of 288 texts that were downloaded from the database ScienceDirect Online. The final word list contained 623-word families and accounted for 12.24% of words in the medical research articles (Wang et al., 2008). For the NAWL, a collection of 1,006,934 running words were

compiled from 252 nursing research articles. The final word list consisted of 676-word families with 13.64% coverage in the nursing research articles (Yang, 2015). While the MAWL and NAWL had greater than 10% coverage in their respective fields, that is not always the case. “Coverage of the word list in the science fiction fantasy corpus was found to be 0.50%, which was 46% higher than coverage of the same list in a corpus of fiction texts (0.27%), and 70% lower than coverage of the same list in a corpus of academic science journals (1.68%)” (Rolls & Rodgers, 2017). While the percentage may seem low, by reading 500,000 words, 21% of the science words will be met 10 times and 83% of the science words will be met at least once (Rolls & Rodgers, 2017). All three of the word lists above followed Coxhead (2000) and defined their words as word families and not lemmas. It seems more and more researchers are focusing on specific word lists instead of general ones because there is an argument stating, “the best way to prepare students for their academic studies is by exposing them to their own discourse” (Hyland & Tse, 2007). For example, by introducing prospective medical students to the target vocabulary in the MAWL, when finally accepted into medical school, the field specific jargon is not as novel and, theoretically, the students are more prepared.

2.6 Chemistry Academic Word List

As this study is focused on science vocabulary in the high school setting, it is necessary to understand what other science specific word lists have been created. In 2013, Valipouri and Nassaji created a chemistry specific word list with two main objectives. They wanted to create a list of the most frequently used academic words in chemistry research articles for EFL chemistry students (Valipouri & Nassaji, 2013). They also wanted to compare the coverage of high frequency words in the AWL (Coxhead, 2000) and the GSL (West, 1953) in their corpus (Valipouri & Nassaji, 2013).

Valipouri and Nassaji (2013) posed three research questions: "1. What are the most frequently used academic words in a large corpus of chemistry research articles? 2. Are the words that occur with high frequency in the corpus of chemistry articles also identified as high frequency words in AWL and GSL word lists? 3. Are there any words that are not identified as high frequency in AWL and GSL, but occur with high frequency in the corpus of chemistry research articles?" (Valipouri & Nassaji, 2013). Using research articles from the database ScienceDirect Online, they built a corpus consisting of 4 million words. This corpus was composed of 1185 written texts in the discipline of chemistry subdivided into 4 main subject areas: analytical chemistry, inorganic chemistry, organic chemistry, and physical/theoretical chemistry (Valipouri & Nassaji, 2013). Using a collection of 320 published volumes in chemistry, 38 volumes were selected. Of those 38 volumes, 8 were randomly selected from each of inorganic chemistry, organic chemistry, and physical/theoretical chemistry. Analytical chemistry was the exception in which 10 volumes were selected. Extra volumes were required for the analytical chemistry section to ensure an equal number of words for each sub-discipline; about 1 million words (Valipouri & Nassaji, 2013).

Similar to Coxhead's (2000) AWL, Valipouri and Nassaji (2013) used the RANGE program to develop their word list. They used the program to "identify the frequency and range of each word in the whole corpus and in each subject area and also those that were in the corpus but not in the AWL and GSL word lists" (Valipouri & Nassaji, 2013). The methodology followed by Valipouri and Nassaji (2013) is almost identical to that of Coxhead's (2000) AWL. Valipouri and Nassaji (2013) started off by separating their text into short, medium and long. They used the three criteria followed by Coxhead (2000) of frequency, range, and specialized occurrence. They classified the word families based on level 6 of Bauer and Nation's (1993) scale. Since Coxhead's (2000) corpus consisted of 3.5 million words and the minimum requirement for a word to occur

needed to be at least 100 times, Valipouri and Nassaji (2013) used the same proportion to deduce that with a corpus of 4 million words, the minimum frequency of the word family needed to be 114 times. With some uncertainty on what is considered technical and academic words, Valipouri and Nassaji (2013) used Chung and Nation's (2003) rating scale as a guide, with the help of chemistry professors in the discipline.

Overall, Valipouri and Nassaji (2013) developed their Chemistry Academic Word List (CAWL) by identifying 1400 academic word families that were used with high frequency in the corpus. Out of 1400 words in the CAWL, more than 600 were part of the 1st and 2nd thousand words families in the GSL. In addition, more than 300 words were part of the AWL. Therefore, the number of field specific terms was just under 400 (Valipouri & Nassaji, 2013).

2.7 Pilot Science Specific Word List

While combining the AWL and GSL to determine their percent coverage of texts from various academic disciplines, Coxhead and Hirsh (2007) found this coverage was comparatively small for science texts. They found that while the combination of AWL and GSL has 86.7%, 88.8%, and 88.5% coverage over the corpora of arts, commerce, and law respectively, the two word lists only had 80% coverage of the science corpus. While there might be multiple reasons why the sciences do not have the same coverage, one thing to note is that all 4 corpora are the same size. After noticing this gap, Coxhead and Hirsh decided to create the pilot science-specific word list (Coxhead & Hirsch, 2007).

This pilot science-specific word list was created by building onto the AWL science sub-corpus. The AWL science sub-corpus was initially made up of over 875,000 running words and consisted of biology, chemistry, computer science, geography, geology, math, and physics

(Coxhead & Hirsch, 2007). In order to improve onto the science sub-corpus, Coxhead and Hirsh (2007) added seven more subject areas: agricultural science, ecology, horticultural science, engineering, and technology, nursing and midwifery, sport and health sciences, and veterinary and animal sciences. All texts were collected electronically and verified by professors and staff at multiple universities in New Zealand (Coxhead & Hirsch, 2007).

Since this study is taking resources from the AWL, Coxhead and Hirsh (2007) followed a similar methodology. They used the criteria of range, frequency, and dispersion when developing their word list. Using the range program, they identified the most frequently occurring words outside of the GSL and AWL. They also made sure that the words occurred in at least seven subject areas. In terms of frequency, the words had to occur at least 50 times in the scientific corpus. Finally, the words needed to have a minimum dispersion factor of 35 (Coxhead & Hirsch, 2007). By expanding on the science corpus designed for the AWL, Coxhead and Hirsh created a corpus containing more than 1.7 million running words. They were able to then create a word list consisting of 318 words and covering approximately 4% of the science specific corpus (Coxhead & Hirsch, 2007). Using the AWL, GSL, and Pilot science-specific word list, the coverage of science specific vocabulary goes up from 80% to 84%. While it is still slightly lower than coverage over the arts, commerce, and law, it is still an improvement.

Chapter 3 – Methodology

This is a corpus-based study. “Corpus-based methodologies have been informed by genre principles of text analysis” (Flowerdew, 2005). Corpus-based studies “create lists, concordances, or data concerning the clustering of linguistic items in coherent, purposeful texts” (Coxhead, 2000). The exact step-by-step process in the methodology is an accumulation of resources. The

steps followed are not informed by a specific article but instead, by a combination of multiple publications explaining the development of word lists (i.e., Chung & Nation, 2004; Coxhead, 2000, 2016; Coxhead & Hirsch, 2007; Gardner & Davies, 2014; Rolls & Rodgers, 2017; Valipouri & Nassaji, 2013; Wang et al., 2008; Yang, 2015).

3.1 Developing the Corpus

Choice of appropriate material when developing the corpus was essential. When creating the corpus, this study took into consideration the representation of text, size, organization, and word selection. This study followed the steps taken by Coxhead (2000) as well as Gardner and Davies (2014) when creating their corpus. In terms of representation, the collection of textbooks used in the corpus are listed in Appendix B. The textbooks selected are those approved by the Ontario government as representative of the Ontario curriculum. The complete list of every book approved by the Ontario government can be found on www.trilliumlist.ca. As well as textbooks, some written public lecture material and assignments were used in the corpus. Many high school teachers create websites and upload their material, so that the students can gain access to them. By simply writing the course code of a class in Google, open access resources, such as written lecture material, assignments, labs, and tests are easily found. The inclusion of public material as part of the corpus is to diversify the language used in the corpus, as well as to help increase the size of the corpus. Not only does the inclusion of public material help diversify the language and increase the number of running words, but most importantly, it is more representative of the language students are exposed to in the classroom setting on a day-to-day basis. While different teachers may use different textbooks approved by the Ontario Government, some teachers do not even use any textbooks in their classrooms and therefore, the textbook is secondary to the classroom lecture material. All public lecture material included was written and not orally presented. In terms of

organization, the material collected was separated into biology, chemistry, physics, and general science. Unlike the AWL, where the word family was used to create the word list (Coxhead, 2000), this study used lemmas instead. By using lemmas, it is possible to identify grammatical parts of speech; for example, being able to differentiate between the verb "used" in "he used a rake" and the adjective "used" in "he bought a used car" (Gardner & Davies, 2014). All text was collected electronically, and the bibliography was removed. Any citations or other words that do not pertain to the sciences (e.g., numbers, author's acknowledgments, etc.) were also not included. The final corpus was composed of 3,235,149 running words. Each field of study (biology, chemistry, physics, and general science) had approximately 800,000 running words. To be exact, the biology sub-corpus contained 832,051 running words, chemistry contained 842,953 running words, physics contained 767,742 running words, and general science contained 792,403 running words. Once the collection of the material for the corpus was done, the next step was to develop the word list.

3.2 Developing the Word List

Rather than analysing the corpus by hand, the program WMatrix was used in the study. The WMatrix is a corpus analysis and comparison software. Unlike the RANGE program where it tags the word family, the current WMatrix tag set has over 130 different tags depending on the word and its location in the sentence. Using the WMatrix software, tagging the words by lemmas instead of word family is made easier. The WMatrix program is the same software used by Gardner and Davies (2014) when creating the AVL. Following the three criteria given by Coxhead (2000) and Gardner and Davies (2014), the words (lemmas) selected were chosen based on specialized occurrence, range, and frequency. For specialized occurrence, this study followed Schmitt and Schmitt's (2014) definition of high frequency words, and therefore, the lemmas included in the

word list were outside of the top 3,000 most frequent English word families of the COCA. By following Schmitt and Schmitt's (2014) definition, it is expected that the learners using the word list understand 95% of the words used in speech (Van Zeeland & Schmitt, 2013). While this study removed the high frequency words in the COCA, it did not exclude words either from the AWL or AVL. Using the ratio by Coxhead (2000) for the range, the word (lemma) had to be present in three of the four subject areas. When creating the word list, a flaw was found that some words would appear more than 90 times in one field, and it would appear a handful of times in other fields. To avoid such a skewed ratio, any lemma included in the word list must occur a minimum of 10 times in three of the four fields. In terms of frequency, there must be a minimum requirement of occurrences for a word before it is considered for inclusion in the OHSWL. The size of the corpus determined the minimum requirement of the frequency of words. As an example, Coxhead (2000) had a corpus of 3,513,330 running words and the requirement was that the words had to occur a minimum of 100 times. When creating their pilot science specific word list, Coxhead and Hirsh (2007) had a corpus of 1,761,380 running words and as a result, made the minimum frequency of the words had to be 50 times. For the OHSWL, the corpus size was 3,235,149 running words. Therefore, based on the ratio explained above, the words needed to appear a minimum of 92 times across the entire corpus.

Chapter 4 – Results and Discussion

4.1 Results

The OHS corpus consisted of 3, 235,149 running words. After running the corpus through the WMatrix, the program produced 55,912 lemmas. It must be noted that 55,912 lemmas are as defined by the WMatrix program and not an accurate definition of a lemma (see below). Only

3,433 of those lemmas occurred 92 times or more. After removing the high frequency words and following the range criteria that the lemmas must occur 10 times or more in a minimum of 3 of the 4 subject areas, approximately 1,283 lemmas were left. Manual analysis of the word list was done to make sure there were no irrelevant, non-science related words present. Elements from the periodic table such as hydrogen, carbon, and oxygen were removed. While some may believe that their inclusion would be valuable since they are commonly used in the science classroom, it is neither a curricular expectation nor an expectation from the teachers to memorize the elements. Students are always given a periodic table during a quiz, test, or assignment, so that they may refer to it. Additionally, all locations (e.g., Canada, Ontario, moon), names (e.g., Newton), numbers in written form (e.g., one, two, thousand), unit of measurement (e.g., Kg, ^{o}C), acronyms (e.g., aq), and any non-sense (e.g., Fi, NH, P., Oi) were also removed from the list. At this point, the OHSWL consists of 977 words. As mentioned earlier regarding the WMatrix program, it contains over 130 classifications of a word. The WMatrix program differentiates between singular nouns and plural nouns. It also differentiates between a verb, verb ending in -ing, and verb ending in -s. Since the OHSWL is intended to identify the different lemmas, a second round of manual analysis was conducted to group inflected word forms under a single lemma. The final OHSWL (Appendix A) contains 803 lemmas.

4.1.1 Occurrence of Science Words

The first research question asked which lemmas beyond the top 3,000 words in the COCA occur most frequently across a range of Ontario high school science material. In the OHS corpus, 803 lemmas met the criteria to be included in the OHSWL. Some of the most frequent lemmas in the OHSWL are atom, acid, and molecule. They appear in the OHSWL at a frequency of 5673, 5214, and 5188 respectively. The words that appear the least in the OHSWL are diverse, tap, and

improved. They all appear at a frequency of 92 times. It is worth noting that there are a total of 10 lemmas that appear with a frequency of 92 times. The three lemmas mentioned were just examples and should by no means be considered less significant than other lemmas with the same frequency.

4.1.2 Percent coverage across the science corpus

Originally when creating the corpus, one of the goals was to ensure that all four topics had a similar number of running words. When collecting textbooks and lecture material, biology seemed to have a much larger number of running words compared to other topics. Topics like physics had approximately 865,000 running words while biology had more than 1.5 million running words. Coxhead (2000) mentioned that any text that met their criteria, but was not included in the corpus, was "kept aside for use in a second corpus to test the AWL's coverage at a later stage". Following their methods, some material (evenly distributed between textbooks and lecture material) was kept aside to be used to answer the second research question. The second research question asks what the percent coverage of the OHSWL in the science corpus is. Initially, the Compleat Lexical Tutor website was used where there exists a program which provides lexical comparison of text. Unfortunately, there happens to be an unspecified word limit for the program. When uploading a file to the program, some files were analyzed and deemed as below the word limit, and others simply produced a blank screen. Rather than uploading one large file containing the science corpus, the file was divided into multiple files, which where then uploaded and compared to the OHSWL. While this method of lexical comparison is satisfactory, there were still some doubts to the accuracy of the final percentage provided by the program. As a result, another method employing excel was used. With the list of words already separated by the WMatrix file and the number of occurrences listed, an excel formula was used to highlight any words that appeared both in the OHSWL and the science corpus. Through the filtering of only the highlighted

words, the total frequency of the words present in the OHSWL can be calculated. By comparing it to the total frequency of all the words in the science corpus, an accurate percent coverage of the OHSWL in the science corpus was produced. This coverage of the OHSWL on the science corpus is 7.79%.

4.1.3 Percent Coverage across non-science Corpus

The third research question asks what the coverage of the OHSWL is in non-science material. Since the OHSWL is specifically aimed at high school students in Ontario, a decision was made that the non-science material must also be intended for high school students in Ontario. The non-science corpus consisted of History and Geography textbooks as well as their written lecture material. The Ontario curriculum approved textbooks were found using the website www.trilliumlist.ca. The percent coverage calculation for the non-science material followed the same calculations used to find the percent coverage of the science corpus. The OHSWL had a coverage of 1.52% across the non-science material. With the OHSWL having more than 6% coverage in the science corpus compared to the non-science corpus, it seems that the word list is fulfilling its purpose. The OHSWL is covering the Ontario High School Science Curriculum as opposed to any random text.

4.1.4 Value of top 3,000 words in science corpus.

The fourth research question asks for the coverage of the top 3,000 words of the COCA in the science corpus. As stated above, the OHSWL coverage of the science corpus is 7.79%. Using the same technique to figure out the OHSWL coverage, the coverage of each 1,000 word can be determined. When analyzing the coverage of the first 1,000 words of the COCA (words 1 to 1,000), it can be seen that it has a 60.17% coverage of the science corpus. When analyzing the second

1,000 words of the COCA (words 1,001 to 2,000), it has a coverage of 9.41% of the science corpus. Based on the massive decline in coverage percentage from the first 1,000 words to the second 1,000 words, it was hypothesised that the coverage of the third 1,000 words would also have a massive decline and be close to 3%. After analyzing the third 1,000 words of the COCA (words 2,001 to 3,000), it was found to have 8.07% coverage of the science corpus. By adding all three numbers together, we can conclude that the percent coverage of the top 3,000 words of the COCA on science material is 77.65%.

4.1.5 Coverage of top 3,000 words in non-science corpus.

The final question of this research tries to determine the percent coverage of the top 3,000 words of the COCA on the non-science corpus. Unlike the analysis of the top 3,000 words in the science material, where each 1,000-word level was analyzed, this question cares to answer the coverage of all 3,000 words. When examined, it was determined that the percent coverage of the top 3,000 words of the COCA for non-science material is 74.15%. This value is very similar to the coverage of the 3,000 words on the science corpus.

4.2 Discussion

4.2.1 OHSWL

The creation of the OHSWL was an interesting process. Beginning the research process, the intention was to have the OHSWL made up of word families similar to Coxhead's (2000) research. However, Gardner and Davies's (2014) justification using lemmas instead of word families was appealing. Furthermore, the idea of following in Gardner and Davies's (2014) footsteps by creating a list with the word families and a list with the lemmas was very enticing. Through personal experience both as an ESL student and teacher, the decision to create the word

list consisting of lemmas made the most sense. But the issue was whether the OHSWL contained words with different lemmas, because learning the headword of a word family does not guarantee comprehension of all the members of the family. While the OHSWL identified some lemmas under a single word family that had very similar meanings, such as the noun and verb form of "breathing", most of the lemmas identified were significantly different. For example, the adjective "charged" in "the negatively charged particle rod attracted the positively charged balloon", is referring to a net amount of positive or negative charge, whereas the verb "charged" in "A star is an electrically charged gas, that shines because nuclear fusion is taking place at its core", is defined as to cause to be agitated, excited, or aroused. There are many other lemmas in the OHSWL of the same word family with different meanings. Some examples are the noun and verb form of heating, the adjective and verb form of labelled, the adjective and verb from of measured, and many more. Due to the majority of the different lemmas under the same word family having different meanings, there was a justification to keep the words in the OHSWL as lemmas and not as word families.

As mentioned above, a lemma consists of the headword and its inflected forms (Webb & Nation, 2017). Theoretically, by knowing the headword, the learner is able to understand its inflected forms. For example, by knowing the headword walk, the learner understands walks and walked. In Coxhead's (2000) AWL, since it consists of word families, regardless of what inflection occurred the most, the headword was the one listed in the AWL. By contrast, in Gardner and Davies's (2014) AVL, which consisted of lemmas, the most frequent form of the word was included in the list regardless of whether it was a headword or an inflected form. Accordingly, a decision needed to be made on how the OHSWL will present its list. A goal of the OHSWL is to be representative of the science corpus. This means that the words listed in the OHSWL should be the most common form found in the science corpus. As a result, the OHSWL includes the most

common form of the lemmas in the science corpus regardless of whether it is a headword or one of the inflected forms. Furthermore, it can be deduced that in the same away that knowing the headword allows the learner to understand the inflections, the opposite can be true where knowing the inflection allows the learner to understand the headword.

Just as there was a debate when creating the OHSWL on whether it should consist of lemmas, word families, or both, a similar debate occurred on what to exclude and what not to exclude from the list. Many of the authors of publications explaining the process of how to create a word list chose to exclude high frequency words. After understanding that fact, the challenge at hand was how to define high frequency words. As mentioned above, there are debates on whether they should be the top 1,000 words or top 3,000 words. The task became even more complicated when reading Coxhead's (2000) article where the top 2,000 words of the GSL were excluded. Initially, a decision was made to follow Dang and Webb's (2016) definition of high frequency and exclude the top 1,000 words from the COCA. This decision was made in order to ensure that regardless of the proficiency level of the learner, they may be able to use the OHSWL. Unfortunately, after creating the word list, the final OHSWL included more than 1,600 words. With such a large number, it was clear that the definition of high frequency needed to be changed. Webb & Nation (2017) explains that "when we read or listen, our focus is on understanding the message, but we might gradually learn words that are encountered in the message by seeing and hearing them again and again in context. Thus, vocabulary learning is seen as being incidental rather than intentional". Since an Ontario high school learner is naturally immersed in English (at least in school), they will encounter the top 3,000 words repeatedly both in the science curriculum and outside of the science curriculum. Therefore, less attention is required by the learner to

understand the high frequency vocabulary. As a result, it was decided that the OHSWL will not include words from the top 3,000 word families of the COCA.

The next step was to determine whether words in the AWL and/or AVL should be included or excluded from the final OHSWL. This was less of a debate and the straightforward. From the beginning, there was no intention to remove any words in the AWL and/or AVL from the final OHSWL. The reasoning behind this choice was because the corpus created by Coxhead (2000) for the AWL and Gardner and Davies (2014) for the AVL consisted of material at the post secondary level. It included papers and research articles taught aimed at different learners than those who will be using the OHSWL. Since no one has attempted to compare the percent coverage of the AWL and/or AVL in the post secondary level vs secondary level, it is not guaranteed that these lists are just as valuable in the secondary level as they would be in the post secondary level.

4.2.2 OHSWL and AVL

After the Chemistry Academic Word List (Valipouri & Nassaji, 2013) and AVL (Gardner & Davies, 2014) were created, the authors compared their lists to the AWL. While the OHSWL was compared to another list, it was not compared to the AWL. Since the OHSWL consists of lemmas, the AVL is the only other list that it can be compared to, as it also consists of lemmas. As mentioned above, the OHSWL consists of 803 lemmas and the AVL consists of 3,000 lemmas. There was an overlap of only 200 lemmas between the two lists. These are lemmas such as axis, beneficial, buffer, catalyst, circuit, and many more. Therefore, 603 lemmas of the OHSWL did not exist in the top 3,000 words of the COCA or the AVL. The benefit of the OHSWL is that rather than the learner needing to know the 3,000 lemmas in the AVL, then learning the OHSWL, they simply need to know the OHSWL, which already includes the most valuable lemmas from the AVL that are specific to the Ontario high school science curriculum.

4.2.3 Value of OHSWL and each of the top 3,000 words of the COCA

When creating a word list, understanding the value it provides the learner is important. This value is determined by its percent coverage compared to other word lists in its specific field or other word lists that may overlap with it in vocabulary. For example, when looking at vocabulary used in television programs, by knowing the first 1,000 words of the British National Corpus (BNC)/COCA, the learner is able to understand 85.35% of words used. The percent coverage of the second 1,000 words (1,001 to 2,000) in television programmes is 4.12%, while knowing the fifth 1,000 words (4,001 to 5,000) covers only 0.59% of vocabulary used (Webb & Nation, 2017). Based on the percent coverage, it can be concluded that the value of learning the first 1,000 words of the BNC/COCA is much greater than learning the second 1,000 words and hence, should be prioritized in learning over the second. Similarly, the value of knowing the second 1,000 words of the BNC/COCA is much greater than knowing the fifth 1,000 most words. While it may seem insignificant to learn the second 1,000 words simply because it covers only 4% of vocabulary, after having learned the first 1,000 words which cover 85.35%, "knowing 4% more of the vocabulary that is encountered will have a positive impact on comprehension" (Webb & Nation, 2017). It must be noted that these values are only true for television programmes. While the idea remains true that knowing the first 1,000 words are more valuable than knowing the second 1,000 words, the percent coverage is different.

Similar to how the relative value of the top 5,000 words of the BNC/COCA were observed, the same can be done for the top 3,000 words of the COCA and the OHSWL. When looking at the COCA, the top 1,000 words have a coverage of 60.17% in the science corpus. With such a high coverage, it is clearly evident that any Ontario high school science learner needs to understand these words first. The second and third 1,000 words of the COCA had a coverage of 9.41% and

8.07% respectively. In comparison to the coverage of the OHSWL in the science corpus which was 7.79%, the second 1,000 words of the COCA are more valuable to the learner. In regard to the third 1,000 words of the COCA, although it may seem similar in percent coverage with only 0.28% greater coverage than the OHSWL, it is still considered more valuable to the learner. Therefore, it can be concluded that the greatest value for a learner in the Ontario high school science curriculum is by learning the first 1,000 words of the COCA, then the second and third 1,000 words, and finally, the OHSWL.

4.2.4 OHSWL and High Frequency Word Coverage in Science vs Non-Science Corpus

As mentioned earlier, this study follows Schmitt and Schmitt's (2014) definition of high frequency words, which are the top 3,000 words. By excluding the top 3,000 words of the COCA from the OHSWL, it is expected that the learner must know these before using the OHSWL. This expectation is made stronger by the fact that each 1,000 words of the COCA in the top 3,000 words provide greater value to the learner than the OHSWL. However, the question at hand is whether the value of the top 3,000 words of the COCA and OHSWL is greater in the science corpus or non-science corpus. By adding the percent coverage of each 1,000 word in the top 3,000 words of the COCA, it can be seen that knowing the high frequency words alone covers 77.65% of the science corpus. In addition to the percent coverage of the OHSWL, the total coverage of the top 3,000 words of the COCA and OHSWL in the science corpus is 85.44%. When adding the percent coverage of the top 3,000 words in the non-science corpus, 74.15%, to the percent coverage of the OHSWL in the non-science corpus, 1.52%, the total coverage equals to 75.67%. With an almost 10% greater coverage in the science corpus compared to the non-science corpus, it is undeniable that the high frequency words of the COCA and the OHSWL would provide a much greater value

to those learning the Ontario high school science curriculum versus the non-science related curriculum.

Chapter 5 – Conclusion

5.1 Implications

Having finally created the OHSWL, the next logical question to ask is: how should it be used? In Canada, there has been a steady increase in the immigrant population and more specifically, in Ontario (Statistics Canada, 2017). The top four places of birth for immigrants coming into Ontario are India, China, United Kingdom, and Philippines (Statistics Canada, 2017); with the exception of UK, the native language of the other countries is not English. Therefore, odds are that the students will need to enter the ESL stream. With immigrant population in Ontario constantly rising, creating field specific word lists to help both current and incoming native and non-native English speakers will be very beneficial. Ultimately, the goal of the OHSWL is to help reduce language as a barrier in the science curriculum by improving comprehension of text, and providing the learner with the opportunity to understand and enjoy the content being taught, rather than it being blindly memorized and turned into a call-and-response game. For that goal to be achieved, there are three levels for which the OHSWL can be utilized. The three levels are at the course designing, teaching, and learning stages.

5.1.1 OHSWL and Course Design

"If a course or text contains too many words that are unknown to the learners, then learners will struggle to focus on the meaning of the text because of the need to deal with the unknown words" (Nation, 2016). It is for that reason; the first level is to introduce the course designers to

the OHSWL. By having the OHSWL, when considering the vocabulary component of the course, they are able to standardize the language used and ensure that the textbooks utilize the vocabulary at an even higher frequency. This means that regardless of whether the student is using a grade 12 physics book or a grade 10 science textbook, the base vocabulary is relatively the same. By doing so, it helps the students learn the target vocabulary with greater ease.

5.1.2 OHSWL and Teaching

The next stage where the OHSWL can be used is at the teaching level. "A well-balanced course has four equal strands of meaning-focused input, meaning focused output, language-focused learning, and fluency development. The language-focused learning strand includes the deliberate teaching and learning of vocabulary, and frequency-based word lists can act as a useful checklists or source lists for such learning" (Nation, 2016). As mentioned earlier, while teachers may feel that they know which words are important and which ones are not, it is merely intuition and not facts (Alderson, 2007). By using the OHSWL at the pre-high school level (grade 7 and 8), the teachers can begin introducing the target vocabulary to the students at an earlier stage to prepare them for the high school science curriculum. Although grade 7 and 8 curriculum may not have these words as the target vocabulary, it is mandatory for every student in Ontario to take grade 9 and 10 science. Since one of the functions of the grade 7 and 8 teachers is to prepare the students for high school, the use of the OHSWL at this level would be beneficial.

This is not to say that the burden of the responsibility of teaching the OHSWL to students falls solely on the grade 7 and 8 teachers. The OHSWL contains field specific vocabulary used in the grade 9 and 10 science courses, as well as grade 11 and 12 physics, biology, and chemistry courses. Therefore, grade 9 and 10 teachers can also use the OHSWL not only to ensure that they are using the same target vocabulary taught by the grade 7 and 8 teachers, but to also maximize

student comprehension of their respective science courses and any future science courses taken by the student.

With teachers from grade 7 up to grade 10 sciences using the OHSWL, the students should have extensive knowledge of most of the target vocabulary in sciences. As long as grade 11 and 12 teachers use the OHSWL, the expectation is that language as a barrier is reduced and the students' focus is no longer on comprehension of text, but rather the understanding of content and applying their knowledge inside and outside of the classroom.

Classification and Linked Skills are two methods explained by Webb and Nation (2017) on how to teach vocabulary. Through the classification method, the students would organise a group of 30 to 40 words under a pre-determined way such as headings, categories, etc. "It is said it is one of the most efficient ways of getting learners to focus on thematically related vocabulary" (Webb & Nation, 2017, p. 81). In terms of the linked skills method, student would need to be exposed to the target vocabulary and use them through different language skills. For example, if the students learned about the Amazon Jungle, they would talk, read, listen to information, and write about it. Since there are many more techniques to teaching vocabulary, teachers may choose the method that most suits them and their students to help them understand the target vocabulary from the OHSWL.

5.1.3 OHSWL and Learning

The final stage where the OHSWL can be used is at the learning stage. As mentioned earlier, a well-balanced course has four equal strands. "During meaning-focused input, when learners meet unknown words in their listening and reading, the words can be checked against word lists to see if they are frequent enough to be worth learning" (Nation, 2016). Currently, with

no changes to any of the course textbooks or teaching methods by the educators, the OHSWL already proves to be a word list which provides the learner with field specific terminology that is of value. Therefore, learners are currently able to use the OHSWL to help identify target vocabulary in the Ontario science curriculum. However, should course designers use the OHSWL as a template when creating their textbooks and teachers emphasize the use of the OHSWL in the classroom, the value gained by using the OHSWL is greater.

5.2 Limitations

The methodology of the OHSWL was proposed only after researching many articles and books on the creation of word lists. However, regardless of how meticulous the research process may have been, when taking research and applying it to produce a product, there is always room for improvement. When looking at the OHSWL, there are two possible limitations. The first limitation would be in relation to the minimum frequency required per sub-corpus. For a word to be included in the OHSWL, it must meet the frequency requirement of occurring a minimum of 92 times in the entire corpus with a minimum of 10 times in three of four sub-corpora. While the number 92 was chosen based on a ratio proposed and accepted by others who have created a word list, the number 10 was chosen arbitrarily. The OHSWL was not created to identify the highly technical terminology relevant to one specific field, but rather, the most common words across the entire high school science curriculum. To ensure even distribution of the word, a minimum requirement per sub-corpus was required and it was estimated that a frequency of 10 is sufficient. Since the minimum frequency was randomly chosen, it is possible some extra science specific words were excluded from the OHSWL or vice versa, where low-frequency words entered the OHSWL. Choosing an arbitrary number is not out of the norm as was demonstrated in Coxhead's (2000) article when choosing the minimum frequency across the entire corpus to be 100. Gardner

and Davies (2014) also randomly chose a 50% ratio and 20% range (as mentioned above) in the development of their AVL.

The second limitation of the OHSWL is the WMatrix program, which was used to categorize the different lemmas. While to a certain degree, there is an element of trust in the program since it was the same one used by Gardner and Davies (2014) in their development of the AVL as well as many other corpus analyses, a mistake was observed in the development of the OHSWL. Accidently, the program separated the capitalized word "Ions" from the non-capitalized "ions" even though they were the same lemma. While ions is not a lemma that is included in the OHSWL, and the mistake by the WMatrix was only observed in this one instance, there is a level of uncertainty on whether the program identified every word correctly. Ultimately, this doubt will always be present for every word list that chooses to use a program to separate their words. It is still a limitation of the study and can only be resolved by manual analysis.

5.3 Future Research

Currently, there seems to be a lot of research on word lists and their coverage, yet there does not seem to be any research focused on effectiveness of teaching a word list or how educators have used their word lists in the classroom. So while we know that the AVL has a 13.8% coverage in academic text (Gardner & Davies, 2014) and the NAWL has a 13.64% coverage in the nursing corpus (Yang, 2015), no one truly knows how effective they are in the classroom. For the OHSWL to be advertised as a group of field specific terminology that help the learners in their science education, future research is required. Two types of research can be done, a short-term study and a long-term study.

5.3.1 Short Term Study

For the short-term study, two science classes that are learning the same subject are chosen. To limit the variables that may act on the study, the same educator should be teaching both classrooms. For Class A, the teacher follows the regular science curriculum, but any lecture material and assignments should incorporate the target vocabulary in the OHSWL more frequently. For Class B, the teacher follows the Ontario curriculum without prioritizing any words from the OHSWL. At the end of the semester, the students are given an exam which comprises of application-based questions. Typically, a student struggling with vocabulary will have difficulty articulating their thoughts using the scientific terminology. Based on the results of the test, a conclusion can be made on the use of the OHSWL in the classroom.

5.3.2 Long Term Study

In terms of the long-term study, it will be over the duration of six years. An OHSWL learning stream will be created for students from grade 7 up to grade 12. While the students will go about their education normally, during their science class, the teacher will include the use of the vocabulary from the OHSWL more frequently. By the time the students reach the grade 12 level, theoretically, they would have mastered the target vocabulary. This hypothesis will be tested by examining a student in grade 12 who entered the OHSWL learning stream compared to someone who did not. Similar to the short-term study, the exam will comprise of application-based questions. The results of the short- and long-term studies should be enough to prove whether the use of a word list in the classroom is beneficial or not.

5.4 Conclusion

The methods followed to create the corpus, as well as the word list, were similar to that of Coxhead's (2000) AWL and Gardner and Davies (2014) AVL. Exact steps could not be followed, since this study is different from those mentioned above due to the fact that there is one particular field of focus, that is, high school sciences, as opposed to a general word list for all academic purposes. The word list created is derived from a corpus that is well balanced and representative of the field. The corpus used was derived from Ontario approved science textbooks and public lecture material specific to the Ontario science curriculum. The main goal when designing this project is to help those learning the Ontario high school science curriculum. For the word list to actually help, it must truly be an Ontario High School Science Word List, as opposed to a general word list. When analysing the coverage of the OHSWL in science vs non-science corpus, the coverage was 7.79% and 1.52% respectively. Fortunately, the difference is a bit over 6%. Not only that, but the coverage of the top 3,000 words of the COCA and the OHSWL on the science corpus vs non-science corpus is 85.44% and 75.67% respectively. With such a large difference in coverage, it is undeniable that the OHSWL truly does what it was set out to do. Interestingly, while there is greater coverage by the first and second 1,000 words of the COCA in the science corpus, the coverage of the third 1,000 words is only marginally greater than that of the OHSWL.

References

Alderson, J. C. (2007). Judging the frequency of english words. *Applied Linguistics*, *28*(3), 383–409. https://doi.org/10.1093/applin/amm024

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap*. McGraw-Hill SRA Columbus, OH.

Browne, C. (2014). *A New General Service List: The BetterMousetrap We've Been Looking for?* *3*(2), 5–14.

Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. New Zealand Council for Educational Research.

Chung, T. M., & Nation, I. S. P. (2003). Technical Vocabulary in Specialised Texts. *Reading in a Foreign Language*, *15*(2), 103–116.

Chung, T. M., & Nation, P. (2004). *Identifying technical vocabulary*. *32*, 251–263. https://doi.org/10.1016/j.system.2003.11.008

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology*, *11*(3), 38–63.

Cobb, T. (2010). *Learning about language and learners from computer programs*.

Coxhead, A. (2000). A New Academic Word List Linked references are available on JSTOR for this article : *TESOL Quarterly*, *34*(2), 213–238.

Coxhead, A. (2016). *Academic Word List ”*. *50*(1), 181–185. https://doi.org/10.1002/tesq.287

Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, *12*(2), 65–78.

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL - International Journal of Applied Linguistics*, *167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Davies, M. (2020). *English-Corpora . org : a guided tour*. *November*, 1–27.

Engels, L. K. (1968). The fallacy of word-count. *IRAL - International Review of Applied Linguistics in Language Teaching*, *6*(1–4), 213–232. https://doi.org/10.1515/iral.1968.6.1-4.213

Flowerdew, L. (2005). *E NGLISH FOR An integration of corpus-based and genre-based approaches to text analysis in EAP / ESP : countering criticisms against corpus-based methodologies*. *24*, 321–332. https://doi.org/10.1016/j.esp.2004.09.002

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327.

Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum*, *17*(1), 24–27.

Heatley, A., & Nation, I. S. P. (1996). *Range (computer software). Wellington: Victoria University of Wellington*.

Hyland, K. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes*, *21*(4), 385–395.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253.

Lynn, R. W. (1973). Preparing word-lists: a suggested method. *RELC Journal*, *4*(1), 25–28.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108.

Nation, I.S.P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nation, Ian S P, & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning Boston, MA.

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System*, *39*(4), 533–538.

Praninskas, J. (1972). *American university word list*. Longman.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 480–489.

Rolls, H., & Rodgers, M. P. H. (2017). English for Speci fi c Purposes Science-speci fi c technical vocabulary in science fi ction-fantasy texts : A case for ' language through literature .' *English for Specific Purposes*, *48*, 44–56.

https://doi.org/10.1016/j.esp.2017.07.002

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. In *Language Teaching* (Vol. 47, Issue 4, pp. 484–503). https://doi.org/10.1017/S0261444812000018

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

Statistics Canada. (2017). *Focus on Geography Series, 2016 Census - Province of Ontario*. https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-PR-Eng.cfm?TOPIC=7&LANG=Eng&GK=PR&GC=35

Townsend, D., & Collins, P. (2009). Academic vocabulary and middle school English learners: An intervention study. *Reading and Writing*, *22*(9), 993–1019.

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. https://doi.org/10.1016/j.jeap.2013.07.001

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074

Wang, J., Liang, S., & Ge, G. (2008). *Establishment of a Medical Academic Word List q*. *27*, 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Webb, S., & Nation, I. S. P. (2017). How Vocabulary Is Learned. *Oxford University Press*. https://doi.org/10.1002/tesq.479

West, M. (1953). *A General Service List of English Words. London: Longman, Green and Co.”*.

Xue, G., & Nation, I. S. P. (1984). *1984: A university word list. Language Learning and Communication 3, 215-229*.

Yang, M. (2015). English for Speci fi c Purposes A nursing academic word list. *English for Specific Purposes*, *37*, 27–38. https://doi.org/10.1016/j.esp.2014.05.003

Appendix A

Word	Part of Speech	Frequency
atom	Noun	5673
acid	Noun	5214
molecule	Noun	5188
chemical	Adjective	4526
react	Noun	4470
equation	Noun	4047
organism	Noun	2791
diagram	Noun	2497
chemical	Noun	2270
bond	Noun	2216
particle	Noun	2211
elements	Noun	2199
electron	Noun	2163
contains	Verb	2026
circuit	Noun	1839
gas	Noun	1798
membrane	Noun	1784
nucleus	Noun	1723
velocity	Noun	1655
graph	Noun	1581
scientists	Noun	1571
equilibrium	Noun	1542
solutions	Noun	1426
concepts	Noun	1409
electrical	Adjective	1386
atomic	Adjective	1350
ionic	Adjective	1340
proteins	Noun	1312
magnetic	Adjective	1260
bacteria	Noun	1247
acceleration	Noun	1190
molecular	Adjective	1175
quantity	Noun	1168
enzyme	Noun	1166
waves	Noun	1158
observations	Noun	1141
characteristics	Noun	1105
metals	Noun	1078
friction	Noun	1054
format	Noun	1049
objects	Noun	1030
kinetic	Adjective	1029
proton	Noun	1022
organic	Adjective	1020
inquiry	Noun	1016
frequency	Noun	1010
displacement	Noun	1008
humans	Noun	1003
thermal	Adjective	1000
ray	Noun	981
periodic	Adjective	973
liquid	Noun	953
analyse	Verb	940
experimental	Adjective	936
vector	Noun	921
prediction	Noun	890
transport	Noun	882
combustion	Noun	881
measured	Verb	860
tube	Noun	852
strand	Noun	848
density	Noun	847
measurement	Noun	843
reproduction	Noun	843
contents	Noun	839
gravitational	Adjective	835
orbit	Noun	832
dissolve	Verb	826
aqueous	Adjective	822
summary	Noun	815
net	Adjective	810
radiation	Noun	810
convert	Verb	795
structures	Noun	787
lens	Noun	776
fossil	Noun	767
polar	Adjective	767
observed	Verb	746
solubility	Noun	737
allele	Noun	723
hypothesis	Noun	722
absorb	Verb	721
voltage	Noun	719
nutrient	Noun	708
variables	Noun	706
cellular	Adjective	702
covalent	Adjective	695
gravity	Noun	690
expectations	Noun	688
calculation	Noun	686
photosynthesis	Noun	682
functions	Noun	674

tissues	Noun	673	solvent	Noun	502	masses	Noun	425
container	Noun	671	associated	Verb	497	emission	Noun	420
respiration	Noun	664	mechanical	Adjective	497	discovered	Verb	419
components	Noun	658	neutral	Adjective	494	fuels	Noun	415
synthesis	Noun	657	connected	Verb	490	identical	Adjective	414
trait	Noun	653	shell	Noun	484	digit	Noun	411
stem	Noun	651	composition	Noun	479	features	Noun	409
oxidation	Noun	648	external	Adjective	479	reasoning	Noun	409
structural	Adjective	637	calculated	Verb	470	referred	Verb	409
solute	Noun	626	secondary	Adjective	470	empirical	Adjective	405
neutron	Noun	609	findings	Noun	464	located	Verb	401
applications	Noun	605	acidic	Adjective	462	viruses	Noun	401
concentrations	Noun	604	isotope	Noun	455	efficiency	Noun	400
fluid	Noun	585	patterns	Noun	454	undergo	Verb	400
reacts	Verb	585	consist	Verb	453	diffusion	Noun	398
magnitude	Noun	582	collision	Noun	450	technological	Adjective	398
liquid	Adjective	581	axis	Noun	449	temperatures	Noun	398
polymer	Noun	579	restriction	Noun	448	respiratory	Adjective	396
charged	Verb	577	symbols	Noun	448	toxic	Adjective	395
wavelength	Noun	574	experiments	Noun	446	techniques	Noun	394
fertilizer	Noun	570	formulas	Noun	443	electromagnetic	Adjective	393
spectrum	Noun	570	pollution	Noun	440	heated	Verb	393
conductor	Noun	566	quantitative	Adjective	440	vapour	Noun	393
horizontal	Adjective	565	connections	Noun	438	plasma	Noun	391
classify	Verb	562	coefficient	Noun	436	transferred	Verb	390
precipitate	Noun	550	crystal	Noun	435	depends	Verb	387
chemist	Noun	545	carbohydrate	Noun	433	radius	Noun	387
pathway	Noun	536	conservation	Noun	431	ionization	Noun	384
attached	Verb	533	insect	Noun	431	replication	Noun	383
maximum	Adjective	514	travels	Verb	430	procedures	Noun	380
beaker	Noun	508	magnet	Noun	428	outer	Adjective	379
decrease	Verb	507	radioactive	Adjective	426	samples	Noun	379

cart	Noun	378	soluble	Adjective	330	conclusions	Noun	289
theoretical	Adjective	378	evolutionary	Adjective	328	attraction	Noun	288
fungi	Noun	376	obtained	Verb	328	titration	Noun	286
relating	Verb	375	excess	Adjective	322	efficient	Adjective	285
vertical	Adjective	375	renewable	Adjective	321	reflection	Noun	285
evaluating	Verb	368	rod	Noun	321	grades	Noun	283
bulb	Noun	366	adaptation	Noun	318	predator	Noun	283
composed	Verb	364	fusion	Noun	317	cathode	Noun	282
decay	Noun	363	studying	Verb	316	interpreting	Verb	280
yield	Noun	363	cylinder	Noun	313	intestine	Noun	280
functional	Adjective	362	medium	Noun	313	solids	Noun	280
orbital	Adjective	362	biodiversity	Noun	311	surrounded	Verb	279
nerve	Noun	358	genome	Noun	311	exothermic	Adjective	278
intensity	Noun	356	qualitative	Adjective	308	stored	Verb	277
stable	Adjective	353	dependent	Adjective	305	ecological	Adjective	276
dominant	Adjective	351	redox	Noun	303	investigations	Noun	276
exploration	Noun	351	harmful	Adjective	302	hazard	Noun	275
flame	Noun	351	interactions	Noun	302	digestion	Noun	274
flask	Noun	349	balloon	Noun	300	chains	Noun	270
resulting	Adjective	345	curve	Noun	300	directions	Noun	270
performing	Verb	342	melting	Noun	300	mineral	Noun	268
commonly	Adverb	339	charged	Adjective	299	artificial	Adjective	267
decomposition	Noun	339	conduction	Noun	299	distances	Noun	267
muscles	Noun	336	consumption	Noun	297	specialized	Adjective	267
conducted	Verb	335	divided	Verb	296	exerts	Verb	265
predicting	Verb	335	heating	Noun	296	homeostasis	Noun	265
dynamics	Noun	334	greenhouse	Noun	295	researchers	Noun	265
advantages	Noun	333	transformation	Noun	294	inheritance	Noun	261
mechanisms	Noun	332	workplace	Noun	294	prey	Noun	261
prefix	Noun	332	researching	Verb	292	rapidly	Adverb	260
circulatory	Adjective	331	combined	Verb	291	feedback	Noun	258
nucleotide	Noun	330	fats	Noun	291	slope	Noun	257

uniform	Adjective	257	endothermic	Adjective	233	probability	Noun	219
ideal	Adjective	254	units	Noun	233	surroundings	Noun	219
pole	Noun	254	alcohols	Noun	232	alternatives	Noun	218
partial	Adjective	251	hazardous	Adjective	232	impulse	Noun	218
photon	Noun	251	omitted	Verb	232	surfaces	Noun	218
principles	Noun	251	sunlight	Noun	232	eukaryotic	Adjective	217
circular	Adjective	249	electronegativity	Noun	229	fermentation	Noun	217
differ	Verb	249	pigment	Noun	229	indicator	Noun	217
fission	Noun	249	graphic	Adjective	228	colourless	Adjective	216
momentum	Noun	249	limiting	Adjective	227	crops	Noun	216
loop	Noun	246	links	Noun	227	bacterial	Adjective	215
accurately	Adverb	245	sustainability	Noun	227	continuous	Adjective	215
listed	Verb	245	notebook	Noun	226	linear	Adjective	215
flowers	Noun	244	protists	Noun	226	tutorial	Noun	215
lung	Noun	244	classification	Noun	225	disorders	Noun	214
proportion	Noun	244	illustrated	Verb	225	sum	Noun	214
introduction	Noun	242	liver	Noun	225	boiling	Adjective	213
homologous	Adjective	241	safely	Adverb	225	paragraph	Noun	213
interference	Noun	241	selective	Adjective	225	succession	Noun	212
apparatus	Noun	240	transmission	Noun	224	batteries	Noun	211
disadvantages	Noun	240	vibration	Noun	224	effectiveness	Noun	211
tire	Verb	240	wires	Noun	224	gaseous	Adjective	210
saturated	Adjective	239	cation	Noun	223	negatively	Adverb	210
expressed	Verb	238	cord	Noun	223	reactor	Noun	210
layers	Noun	238	physicist	Noun	222	separated	Verb	210
metabolic	Adjective	238	terminology	Noun	222	females	Noun	209
genetics	Noun	237	tools	Noun	222	synthetic	Adjective	209
spontaneous	Adjective	235	instruments	Noun	221	biologists	Noun	203
ozone	Noun	234	symptoms	Noun	221	protective	Adjective	203
reactive	Adjective	234	volumes	Noun	221	precise	Adjective	202
diameter	Noun	233	aquatic	Adjective	220	vehicles	Noun	202
distinguish	Verb	233	mitochondria	Noun	219	consumers	Noun	201

generator	Noun	201	suitable	Adjective	188	fixed	Adjective	177
positively	Adverb	201	versus	Preposition	188	induction	Noun	176
rapid	Adjective	200	correctly	Adverb	187	lipids	Noun	175
conventional	Adjective	199	dipole	Noun	187	mixed	Verb	175
males	Noun	199	evolved	Verb	187	thermometer	Noun	175
burning	Adjective	198	lakes	Noun	187	combinations	Noun	174
gradient	Noun	197	refraction	Noun	187	communicating	Adjective	174
recording	Verb	197	veins	Noun	187	filter	Noun	174
summarize	Verb	197	minimum	Adjective	186	requirements	Noun	174
collected	Verb	196	cytoplasm	Noun	185	defined	Verb	173
signals	Noun	196	topics	Noun	185	recording	Noun	173
transcription	Noun	196	agents	Noun	184	catalyst	Noun	172
locations	Noun	195	atmospheric	Adjective	184	condensation	Noun	172
valve	Noun	195	cloning	Noun	184	generations	Noun	172
haploid	Adjective	194	dissolved	Adjective	184	implications	Noun	172
metallic	Adjective	194	exposed	Verb	184	osmosis	Noun	172
prokaryotes	Noun	194	tail	Noun	184	resonance	Noun	172
categories	Noun	192	abiotic	Adjective	183	activation	Noun	171
genetically	Adverb	192	communicating	Verb	183	homozygous	Adjective	171
notation	Noun	192	heterozygous	Adjective	183	similarities	Noun	171
trends	Noun	192	tested	Verb	183	wastes	Noun	171
beneficial	Adjective	191	interval	Noun	182	precautions	Noun	170
strip	Noun	191	shapes	Noun	182	antibiotics	Noun	168
absorption	Noun	190	silicon	Noun	181	incomplete	Adjective	167
burner	Noun	190	advances	Noun	180	codon	Noun	166
instructions	Noun	190	maintaining	Verb	180	decrease	Noun	166
neutralization	Noun	190	contributions	Noun	179	flows	Verb	166
resulting	Verb	190	digestive	Adjective	179	vocabulary	Noun	166
definitions	Noun	189	urine	Noun	179	sketch	Noun	165
salts	Noun	189	corrosion	Noun	178	nonpolar	Adjective	163
breeding	Noun	188	sensor	Noun	178	aspects	Noun	162
performing	Noun	188	briefly	Adverb	177	concentrated	Adjective	162

consequences	Noun	162	hypothesizing	Verb	153	generated	Verb	146
binding	Adjective	161	nucleic	Adjective	153	binary	Adjective	145
equivalent	Adjective	161	rewrite	Verb	153	corresponding	Adjective	145
forests	Noun	161	labelled	Verb	152	faster	Adverb	145
glycolysis	Noun	161	ultraviolet	Adjective	152	fur	Noun	145
rotation	Noun	161	imaging	Noun	151	lamp	Noun	145
chloroplasts	Noun	160	reptiles	Noun	151	agricultural	Adjective	144
eukaryotes	Noun	160	distilled	Adjective	150	consumed	Verb	144
elastic	Adjective	159	enables	Verb	150	disposal	Noun	144
geometric	Adjective	158	partially	Adverb	150	labelled	Adjective	144
plasmid	Noun	158	ratios	Noun	150	chlorophyll	Noun	143
proportional	Adjective	158	readily	Adverb	150	completed	Verb	143
starch	Noun	158	shaped	Adjective	150	dense	Adjective	143
unsaturated	Adjective	158	spinal	Adjective	150	sphere	Noun	143
engineers	Noun	157	voltmeter	Noun	150	translation	Noun	143
extinction	Noun	157	copies	Noun	149	dilute	Adjective	142
organizer	Noun	157	flammable	Adjective	149	interact	Verb	142
pencil	Noun	157	pesticides	Noun	149	resistant	Adjective	142
replaced	Verb	157	stationary	Adjective	149	sketch	Verb	142
aircraft	Noun	155	tertiary	Adjective	149	transformed	Verb	142
cycles	Noun	155	accuracy	Noun	148	gel	Noun	141
excess	Noun	155	diploid	Adjective	148	detailed	Adjective	140
terminal	Noun	155	labels	Noun	148	prevents	Verb	140
carrier	Noun	154	xylem	Noun	148	biochemical	Adjective	139
electroscope	Noun	154	apron	Noun	147	initiating	Verb	139
ethical	Adjective	154	elevator	Noun	147	breathing	Noun	138
genotype	Noun	154	pancreas	Noun	147	communicating	Noun	138
mixtures	Noun	154	skeleton	Noun	147	dimensions	Noun	138
vertebrates	Noun	154	arranged	Verb	146	reverse	Adjective	138
branches	Noun	153	bloodstream	Noun	146	ribosome	Adjective	138
diagnostic	Adjective	153	deposits	Noun	146	substitute	Verb	138
electrostatic	Adjective	153	fibres	Noun	146	principal	Adjective	137

distinct	Adjective	136	generating	Adjective	129	decimal	Adjective	122
geothermal	Adjective	136	illustration	Noun	129	introduced	Verb	122
hydrolysis	Noun	136	removal	Noun	129	poster	Noun	122
measured	Adjective	136	surrounding	Adjective	129	textbook	Noun	122
terrestrial	Adjective	136	freezing	Adjective	128	transgenic	Adjective	122
poisonous	Adjective	135	phloem	Noun	128	hybrid	Adjective	121
technician	Noun	135	steam	Noun	128	legs	Noun	121
varies	Verb	135	binds	Verb	127	metric	Adjective	121
agriculture	Noun	134	chemically	Adverb	127	pump	Noun	121
arrow	Noun	134	insoluble	Adjective	127	representation	Noun	121
assuming	Verb	134	lactose	Noun	127	vital	Adjective	121
perpendicular	Adjective	134	spontaneously	Adverb	127	beam	Noun	120
phases	Noun	134	subatomic	Adjective	127	exploring	Verb	120
fluorescent	Adjective	133	uncertainty	Noun	127	linked	Verb	120
grid	Noun	133	cholesterol	Noun	126	secretion	Noun	120
passive	Adjective	133	gently	Adverb	126	teeth	Noun	120
photosynthetic	Adjective	133	glycerol	Noun	126	biotic	Adjective	119
adapted	Verb	132	invasive	Adjective	126	candle	Noun	119
fraction	Noun	132	separation	Noun	126	dolphin	Noun	119
caution	Noun	131	sucrose	Noun	126	dynamic	Adjective	119
hemoglobin	Noun	131	oceans	Noun	125	handling	Verb	119
input	Noun	131	precision	Noun	125	wooden	Adjective	119
monomers	Noun	131	ruler	Noun	125	cardiac	Adjective	118
outline	Verb	131	sufficient	Adjective	125	significance	Noun	118
testable	Adjective	131	biotechnology	Noun	124	combined	Adjective	117
attractive	Adjective	130	configuration	Noun	124	continuously	Adverb	117
inorganic	Adjective	130	corrosive	Adjective	124	items	Noun	117
laser	Noun	130	random	Adjective	124	pathogens	Noun	117
phenotype	Noun	130	anatomy	Noun	123	analogy	Noun	116
stopper	Noun	130	endocrine	Adjective	123	farmers	Noun	116
clothing	Noun	129	faster	Adjective	123	columns	Noun	115
derived	Verb	129	polarity	Noun	123	cyclic	Adjective	115

detected	Verb	115	transported	Verb	110	gradually	Adverb	103
gathered	Verb	115	vitamins	Noun	110	helix	Noun	103
reactant	Adjective	115	improvements	Noun	109	cellulose	Noun	102
soap	Noun	115	probe	Noun	109	definite	Adjective	102
warmer	Adjective	115	wildlife	Noun	109	ears	Noun	102
acceptable	Adjective	114	embryonic	Adjective	108	foil	Noun	102
diffraction	Noun	114	reliable	Adjective	108	isolated	Adjective	102
homework	Noun	114	remote	Adjective	108	turbine	Noun	102
mainly	Adverb	114	rubber	Adjective	108	investigating	Verb	101
puck	Noun	114	deer	Noun	107	longitudinal	Adjective	101
references	Noun	114	footprint	Noun	107	supplied	Verb	101
discoveries	Noun	113	lifestyle	Noun	107	ventricles	Noun	101
greatly	Adverb	113	microorganisms	Noun	107	contraction	Noun	100
thermodynamics	Noun	113	abundance	Noun	106	infrared	Adjective	100
amphibians	Noun	112	classmates	Noun	106	rings	Noun	100
capsule	Noun	112	dispersion	Noun	106	tasks	Noun	100
equals	Verb	112	fork	Noun	106	urea	Noun	100
gained	Verb	112	infectious	Adjective	106	valid	Adjective	100
pulse	Noun	112	ingredients	Noun	106	altitude	Noun	99
viral	Adjective	112	numerical	Adjective	106	buffer	Noun	99
bubbles	Noun	111	substitution	Noun	106	electrochemical	Adjective	99
extension	Noun	111	terminal	Adjective	106	reflecting	Verb	99
observational	Adjective	111	explosion	Noun	105	solving	Noun	99
rubber	Noun	111	freely	Adverb	105	solving	Verb	99
stability	Noun	111	incidence	Noun	105	angles	Noun	98
thoroughly	Adverb	111	synthesized	Verb	105	stirring	Adjective	98
desired	Adjective	110	triangle	Noun	105	wax	Noun	98
directed	Verb	110	automobile	Noun	104	approved	Verb	97
electrically	Adverb	110	coloured	Adjective	104	breathing	Verb	97
flat	Adjective	110	triple	Adjective	104	dissection	Noun	97
grains	Noun	110	variations	Noun	104	evaporation	Noun	97
independently	Adverb	110	dehydration	Noun	103	gloves	Noun	97

instantaneous	Adjective	97	divisions	Noun	92
loads	Noun	97	headings	Noun	92
permeable	Adjective	97	improved	Verb	92
repulsion	Noun	97	nail	Noun	92
convenient	Adjective	96	pollination	Noun	92
conversion	Noun	96	tap	Noun	92
upward	Adverb	96	touching	Verb	92
warming	Noun	96			
bladder	Noun	95			
coronary	Adjective	95			
diffuses	Verb	95			
emitted	Verb	95			
erosion	Noun	95			
flexible	Adjective	95			
organized	Verb	95			
questioning	Noun	95			
segment	Noun	95			
dissociation	Noun	94			
dropper	Noun	94			
muscular	Adjective	94			
damaged	Verb	93			
established	Verb	93			
niche	Noun	93			
nomenclature	Noun	93			
potentials	Noun	93			
pulmonary	Adjective	93			
transmitted	Verb	93			
unstable	Adjective	93			
widespread	Adjective	93			
anions	Noun	92			
distributed	Verb	92			
diverse	Adjective	92			

Appendix B

Number	Course	Book Name	Author
1.	SNC1D	Nelson Science Perspectives 9	M. DiGiuseppe, D. Fraser, D. Hayhoe
2.	SNC2D	Nelson Science Perspectives 10	M. DiGiuseppe, D. Fraser, D. Hayhoe
3.	SBI3U/C	Addison Wesley Biology 11	Ray Bowers ; Dean Eichorn ; Len Silverman ; Gail de Souza ; Rob Young ; Susan Green ; Cecilia Chan ; Eileen F. Pyne-Rudzik ; Louise MacKenzie
4.	SBI3U/C	McGraw-Hill Ryerson Biology 11	Don Galbraith ; Leesa Blake ; Jean Bullard ; Anita Chetty ; Eric Grace ; Adrienne Mason ; Donna Matovinovic ; Grace Price ; Catherine Little ; D'Arcy Little M.D. ; Keith Gibbons ; Chris Schramek
5.	SBI3U/C	Nelson Biology 11 University Preparation	M. DiGiuseppe; D. Fraser; J. Dulson; et. al
6.	SCH3U/C	Addison Wesley Chemistry 11	Geoff Rayner-Canham ; Sadru Damji ; Ute Goering Boone ; Susan Green ; Cecilia Chan ; Nancy Andraos ; Jackie Dulson
7.	SCH3U/C	McGraw-Hill Ryerson Chemistry 11	Frank Mustoe ; Michael P. Jansen ; Ted Doram ; John Ivanco ; Christina Clancy ; Anita Ghazariansteja
8.	SCH3U/C	Nelson Chemistry 11	Hans van Kessel ; Dr. Frank Jenkins ; Lucille Davies ; Patricia Thomas ; Dr. Dick Tompkins ; Oliver Lantz
9.	SPH3U/C	McGraw-Hill Ryerson Physics 11	Greg Dick ; Arthur N. Geddis ; Ed James ; Tom McCaul ; Barry McGuire ; Richard Poole ; Bob Holzer ; Rob Smythe
10.	SPH3U/C	Nelson Physics 11 University Preparation	M. DiGiuseppe; R. Vucic; C. Stewart; et. al
11.	SBI4U/C	McGraw-Hill Ryerson Biology 12	Trent Carter-Edwards; Susanne Gerards; Keith Gibbons; et al
12.	SBI4U/C	Nelson Biology 12 University Preparation	Douglas Fraser, Barry LeDrew, Angela Vavitsas
13.	SCH4U/C	McGraw-Hill Ryerson Chemistry 12	Barbara Nixon-Ewing; Mary Schroder; Katy Farrow; et al

14.	SCH4U/C	Nelson Chemistry 12 University Preparation	Maurice DiGiuseppe; Kristina Salciccioli; Milan Sanader
15.	SPH4U/C	McGraw-Hill Ryerson Physics 12	Greg Dick, Dr. Lois Edwards, David Gue, Eric Brown, Robert Callcott
16.	SPH4U/C	Nelson Physics 12	Al Hirsch, David Martindale, Charles Stewart, Maurice Barry

Curriculum Vitae

Name:

Mohamed Mahfouz

Education:

MASTER OF ARTS IN APPLIED LINGUISTICS | UNIVERSITY OF WESTERN ONTARIO

· Class of 2021

BACHELOR OF EDUCATION| UNIVERSITY OF WESTERN ONTARIO

· Class of 2019

HONORS LIFE SCIENCES | MCMASTER UNIVERSITY

· Class of 2017

Work Experience

LONDON INTERNATIONAL ACADAMY | JANUARY 2020 – PRESENT

· IB Biology, OSSD Biology and Science Teacher

QUEST LANGUAGE STUDIES| JULY– SEPTEMBER 2018, APRIL – AUGUST 2019

· English Teacher

SAUNDERS SECONDARY SCHOOL| FEBRUARY – MARCH 2019

· Biology and Science Teacher

CLARKE ROAD SECONDARY SCHOOL| MARCH – APRIL 2018, SEPTEMBER – OCTOBER 2018

· Biology and Science Teacher

MCMASTER UNIVERSITY | FEBRUARY – AUGUST 2016

· Science Instructor

COLUMBIA INTERNATIONAL COLLEGE | JUNE – AUGUST 2015

· Computer Instructor