Language Research Bulletin, 30, ICU, Tokyo

Lexical Features of the ELA Reader: What words are students exposed to?

Izumi Watanabe-Kim
English for Liberal Arts Program
International Christian University

This article reports on a study conducted to identify the lexical features of the ELA Reader, the common in-house textbook of the English for Liberal Arts Program at International Christian University. Adopting the corpus linguistics approach, the Reader was analyzed to identify the most frequently appearing words, keywords, lexical variation and proportion of academic vocabulary. Findings suggest that the current selection of articles is generally appropriate though considerable variation is present among individual texts. Several suggestions are made on the criteria for text selection and on helping students developing their repertoire of academic vocabulary.

Students in their first year of undergraduate study are commonly exposed to significant amount of English academic vocabulary in their courses. In fact, it is often reported that in their first year of undergraduate study, students find English academic vocabulary "a particularly challenging aspect of their learning" (Hyland & Tse, 2007, p.236) whether they are learning in their first or second language. Furthermore, it is rare that these words are explained, unlike the keywords directly related to the content (Flowerdew, 1993). Instead, students are expected to incidentally "pick- up" the meaning and usage of these academic words as they maneuver through difficult readings.

When students are assessed, they are yet again challenged by academic vocabulary as they are commonly asked to display their knowledge in writing. Incorporation of academic vocabulary is crucial in such forms of assessment as it can influence the evaluation. According to Hinkel (2003), limited lexical resources may "create an overall impression of textual simplicity in texts (…) that may therefore reflect negatively on the quality of L2 academic essay" (p.276). This can be backed by a quick survey of assessment criteria of the writing component of standardized tests, such as TOEFL and IELTS, which seem to put a great deal of importance on "lexical resource"

(British Council, 2015) and "appropriate word choice" (Educational Testing Service, 2015).

Though some question the validity of academic word lists, it is widely accepted that the knowledge of and ability to integrate academic vocabulary into writing are factors of academic success. Therefore, researchers and teachers of academic English have developed and employed academic vocabulary lists for explicit instruction to help students to achieve academic success in higher education. (Hyland & Tse, 2007).

A standard method of generating such a list is to rely on methodologies of corpus linguistics by identifying features of a particular language. Assuming that language input is largely influencing the vocabulary acquisition of L2 learners, it is worth examining the types of language our students are exposed to. Thus, the current study is an attempt to investigate the lexical features of the ELA Reader to identify the types of input our students receive by taking the corpus linguistics approach.

A brief history of corpus linguistics

Although the field of corpus linguistics is considered relatively young, it can be traced back to the early 1900s when word lists were developed to meet the language learning needs of US immigrants and the indigenous people of British colonies (Kennedy, 1998). In those days, with the absence of computer technologies, word lists were developed manually with the intuition of language teachers. (Ishikawa, 2008). Early lists of vocabulary include Harold Palmer's 3,000-word list published in 1931 and Michael West's 2,000-word General Service Lists (GSL) published in 1953 (Ishikawa, 2008). Despite being over half a century old, the GSL, which contains the most frequently used 2,000 English word families, continues to be the popular source of reference corpus (Coxhead, 2000).

It was not until the 1960s when electronic corpus was made available. The first of its kind was the Brown corpus of 1961 (Akano, et. al., 2014; Ishikawa, 2008), representing the "first generation" corpus. In the 1970s and 80s, with the advancement of computer technologies, the size, and variation of corpus grew rapidly. These massive databases, including the Lancaster/Oslo-Bergen Corpus (LOB) and Collins Birmingham University International Language Database (COBUILD), were known to be the "second generation" corpus (Anthony, 2015).

In the 1990s, in addition to language produced by native speakers of English, language produced by learners caught the attention of corpus linguists. Further advancement of computer tools allowed the corpus to grow larger in scale as seen in the 100-million-word British National Corpus (BNC) in 1995 (Anthony, 2015). In Japan, JACET 4,000 (1993) and JACET 8,000 (2003) were developed to meet the language learning needs of the Japanese audience (Ishikawa, 2008). Building upon the original LOB corpus, Freiburg-Brown (Frown) corpus and Freiburg LOB (FLOB) corpus were created (Akano, et. al., 2014).

More recently, the Academic Word List 570 was developed by Coxhead (2000) to meet the needs of English language learners and teachers of higher education. The list does not contain the most frequent 2,000 words listed in GSL; instead, it contains the most frequently encountered words in various academic texts (Watanabe, 2013). Also, following the sampling frame of LOB and the FLOB, AmE06 and BE06 were developed from the corpora of American and British English compiled between 2005 and 2007 (Lancaster, 2015).

Today, with the spread of the internet and development of easy-access, user-friendly corpus tools, even a novice can create a corpus in seconds by simply downloading texts from the internet (subsequently raising the issue of validity and the copyright). With the appearance of powerful search engines (e.g., Google) and abundant language data made available in cyberspace, it appears that the field has made a shift from the third generation to the fourth.

Corpus-based studies and tools for analysis

Corpus linguistics is defined as “the study of data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts” (McEnergy & Hardie, 2012). As such a definition shows, it is rather ambiguous to distinguish whether corpus linguistic is a "study" or "computer aided analysis." The study-or-method debate can be traced back to Chomsky's criticism of the approach of study (McEnergy & Hardie, 2012), but the discussion of the opposing views would be irrelevant here. Rather, it is important to point to the fact that scholars find the techniques and instruments of corpus linguistics useful, and they have been adopted and applied to analyze language in numerous contexts. The study (or the methodology) is similar to genre analysis in the sense that it identifies patterns and features of language use.

Naturally, more studies have been conducted on written texts than speech due to the ease of obtaining and handling data. Studies in the past can be categorized as 1) error analysis; 2) comparative analysis of spoken and written language, native and non-native language, British and American English and other variations of language use; 3) interlanguage attributes; and 4) pedagogical applications of corpus-based research (Tono, 2003). These studies more or less employ lexical profiling tools which explore frequencies of words, concordance tools that look into collocations and phrases in use, and analytical techniques of grammatically-tagged data to find patterns and syntactic features in discourse. For the current study, only the commonly employed lexical profiling tools are reviewed.

Basic lexical features of the text

Frequency count of words can "provide a general picture of a text" (Adolphs, 2006, p.40). It is concerned with the frequency of word appearance in a given text or speech. Recent electronic text analysis tools have made it possible for anyone to create a wordlist in seconds, and such a list can be utilized in a variety of ways in language teaching and learning.

Keywords are the "unusually frequently-appearing words" (Anthony, 2015) in a speech or text. The degree of 'unusualness' or 'keyness' is calculated by comparing the frequency of word appearance in the target corpus to the reference corpus (Adolphs, 2006). The reference corpus is considered to be the basis of calculating the "expected frequency" of a particular word appearance, and if the "observed frequency" deviates from the former, the program determines the statistical significance of the difference between the two (Adolphs, 2006). Experts advise that caution is required in choosing an appropriate reference list as it can have a significant impact on the results of the test (Adolphs, 2006).

Lexical variation or some referred to as "lexical density" is commonly measured by the type/token ratio. It is calculated by dividing the number of different types of words by the total number of words (or tokens) used in a text. For example, in the following sentence, "The ELA Reader is the common textbook for all students in the ELA program", there are 14 tokens, but 11 types of words because some words repeat. The type/token ratio (TTR) of this sentence would be 0.78. A higher ratio indicates that

there is a greater variety of words, implying that the repetition of words is low (Adolphs, 2006; Granger, 1998).

Proportion of academic words is a common indicator of the level or the readability of a text. It is determined by the proportion of high-frequency 'everyday' words and low-frequency 'academic' words composing the text. It is considered that the higher the rate of low-frequency words, the more challenging the text would be for a reader. The proportions of different levels of vocabulary can be calculated by referring to the General Service Lists and Academic Word List (Hinkel, 2003; Anthony, 2015).

The Study

The current study aims to identify the lexical features of the ELA Reader, the in-house textbook of the English for Liberal Arts (ELA) program at the International Christian University (ICU). Given that the Reader is a required reading for nearly 600 first-year students, attention must be paid to language in addition to the content to ascertain the types of language that students are exposed to. Adopting the corpus linguistics approach, an initial attempt was made to explore the lexical features of the ELA Reader.

Background

The English for Liberal Arts Program (ELA) is an intensive English for Academic Purposes (EAP) program, in which, all incoming students must enroll to acquire the necessary academic English language skills to be successful in courses offered in English. Incoming students are placed into one of four levels or streams, based on their English language proficiency. The first year experience at ICU is conceptualized as a flowing of stream beginning with a 'discovery' in the spring term, which focuses on the educational values of higher education. In the fall, the journey moves 'inward bound'. Students question and reflect on the process of perception and bias though the themes of culture and race. The stream ends with the 'outward bound' focus in winter when students are asked to think about their visions for the future through topics such as peace and bioethics.

Since the program is integrated, a common in-house textbook, the ELA Reader, is used by all full-time instructors in the core courses. The readings included in the book

are selected by instructors to meet the program objectives. Text selection is a labor-intensive, time-consuming process as it involves piloting and achieving consensus and approval of multiple teachers. Despite different backgrounds, teachers come to agree upon choosing texts that are authentic, thought provoking and academic, providing that they meet the umbrella themes of each term, and are level-appropriate and practical for teaching. The reader includes a collection of texts, usually selected from books and articles written by academics.

Data

The source of data was the ELA Reader 2015, comprised of 19 articles chosen by teachers to meet the program objectives. The data set contained a total of 107,332 words, which was converted to a single text file to create the "ELA corpus."

Data analysis

Lexical features of the ELA corpus were analyzed using the *AntConc*, an application tool developed by Laurence Anthony. The initial step was to generate a list of vocabulary in the order of frequency. Type/Token ratio for each of the 19 articles was calculated to find degrees of lexical variation. The second step was to identify keywords of the ELA corpus referring to the general corpus of American English. Since most texts in the ELA Reader were written by American authors, a large corpus of American English was considered appropriate for the comparison. For certain lexical items, recurrent word sequence was investigated. Further analysis was conducted to identify the ratio of academic vocabulary using the *AntWordProfiler,* also developed by Anthony. The following word lists were used for the analysis of this study:

General Service List (GSL) 1st 1000 level
General Service List (GSL) 2nd 1000 level
Academic Word List (AWL) 570
AmE06 Word List

Results and Discussion

The ELA Corpus

Frequency count of words generated a list of words in the ELA corpus in the order of frequency. The top 60 frequently appearing words are listed in Table 1. As expected, the top ten items on the list were function words: *the*, *be*, *of*, *and*, *to*, *a*, *in*, *that*, *or*, and *for*. The highest ranking pronoun was *we*, which appeared 583 times. Since most of the readings included in the reader are single-authored, the frequent use of *we* may be interpreted as an authors' conscious effort to engage the audience. The top three nouns were *human* (353 times), *people* (318 times) and *Japanese* (194 times). This appears to confirm that the reader is serving the interests of Japanese university students, and also, of humanity department.

Keywords of the ELA corpus were identified and listed in the order of keyness in Table 2. Similar to the result of frequency count, words with higher keyness value are closely related to the content of the text. Such words include *Japanese*, *human*, *belief*, *race*, *propaganda*, *perception*, *euthanasia*, *communication* and *genetic*.

Lexical variations were measured by type/token ratio (TTR). As shown in Table 3, the type/token ratio of the ELA corpus was 0.1. When the corpus was divided into spring, fall and winter readings, the ratio was 0.14 for the spring, 0.14 for the fall, and 0.16 for the winter as shown in Table 4. A slight increase in the lexical variation was seen in the winter term due to the inclusion of more lexically dense articles such as *Shannon & Kockler* (TTR=0.36), *Wager* (TTR=0.33), and *Ogata & Cels* (TTR=0.33) as listed in Table 3. One might wonder why the TTRs of individual articles do not average out to the TTR of the entire book/term. That is because TTR is sensitive to the text length. It does not take it into consideration that the longer the text (or larger the token size), the more likely that words are to be repeated (Jarvis, 2002). Assuming that students' reading proficiency and range of vocabulary increase as they proceed in the program, overall, the current line-up of articles appears to be pedagogically appropriate with a gradual sloping of lexical density. However, when examining individual articles, it is important to keep in mind that a higher TTR does not necessarily reflect a higher ratio of academic vocabulary. It may result from a higher proportion of content-specific, off-list words, such as *Blumenbach*.

Proportion of Academic vocabulary of the ELA corpus was calculated and compared with other studies of the academic corpus. The overall distribution of word tokens of the ELA corpus is shown in Table 4. In order to generate the proportion of academic vocabulary, words were divided into four groups: GSL first level, GSL second level, academic words and other "off-list" words. Of the 107,332 word tokens contained in the ELA corpus, 73.11% matched with the first GSL list, 4.66% matched with the second GSL, and 8.42% matched with the AWL. The remaining 13.81% were the off-list words, which were likely to be either proper nouns or content-specific vocabulary items.

In comparison to the corpus of other studies, the proportion of academic words in the ELA corpus was lower but comparable. For example, in studies of the multi-disciplinary corpus, academic words occupied 10% of the text in Coxhead's (2000) study, and 10.6% in Hyland and Tse's (2007) study. In discipline-specific studies, academic words accounted for 9.1% in agriculture research articles (Martinez, Beck & Panza, 2009), 11.7% in applied linguistic papers (Vongpumivitch, Huang & Chang, 2009), 11.6% in finance corpus (Neufeld, Hancioglu & Eldridge, 2011), and 10.07% in medical research articles (Chen and Ge, 2007). Considering that these corpora consisted of authentic academic/professional papers (for example, in the genre of applied linguistics, articles were sampled from journals such as *TESOL Quarterly*, *Applied Linguistics*, and *Modern Language Journal*), 10% (roughly taking an average of these studies) might be too challenging for the audience of the ELA Reader. Therefore, an 8.42% ratio in the ELA corpus seemed sensible and realistic. Of course, if teachers feel that the Reader should be more challenging for our students, it is possible to bring the ratio up to 9% by replacing a few articles with higher proportions of academic words.

However, when the ELA corpus was subdivided into individual texts for further analysis, it was found that the proportions of academic words varied considerably, ranging from 2.13% to 17.09% as shown in Table 4. The text with the lowest ratio of academic vocabulary was *Cross*, a sample political speech deliberately written to illustrate the tactics of propaganda. The one with the highest proportion of academic vocabulary was the *Ogata & Cels*, which was taken directly from an academic journal of global governance. Interestingly, when TTRs of the two texts were compared, Cross (TTR=0.40) had a higher value than the *Ogata & Cels* (TTR=0.33). This endorsed the point made earlier that TTR is sensitive to text length, and it only indicates a greater variety of words, not necessarily academic words. Being a model of political speech, *Cross* included a large number of off-lists words reflecting mid-American

cultural values such as *John Wayne*. But it was not the text with the highest ratio of off-list words.

The texts with the highest proportions of off-lists words were *McDaniel* (22.06%) and *Shipman* (20.26%). This might be explained by the fact that the former, *McDaniel,* was originally a conference paper including a large number of technical terminology, whereas the latter, *Shipman*, being a historical narrative, consists many proper nouns. Currently, *McDaniel* is under evaluation for a possible replacement. Though the proportion of off-list words may not be the most important factor in the decision-making process of text selection, it might be worth taking it into consideration, as it can increase the workload for both teachers and learners, with little return on investment. In another words, we may want to question, what is the relative value of teaching/learning the off-list words, such as *olfactics* in an EAP program?

Conclusion

This study focused on presenting descriptive data on the lexical aspects of the ELA Reader. A summary of the results of the study includes the following observations. In general, the selection of articles in the reader appear to be appropriate concerning lexical variation and the proportion of academic vocabulary, though there is considerable variation among individual articles. Also, the frequency count of words and keyword search confirmed that the content of the reader is inclined towards the humanities. Another finding is that many of the articles are composed of a high proportion of content-specific, off-list words, which could be a concern that the high proportion of content-specific words are absorbing students' attention, leaving little time and opportunity to learn more prevalent academic words presented in the readings. If so, it may be necessary to either 1) select articles with lower proportions of off-list words or 2) provide more explicit instruction to direct students' attention to use general academic vocabulary such as the ones listed in Table 5.

However, from another perspective, it is possible to make the supposition that students are building their academic vocabulary through the readings. Perhaps, they have not yet been able to shift those words from receptive to productive vocabulary for teachers to confirm their vocabulary development. If that is the case, more opportunities to integrate newly acquired academic vocabulary into student writings must be provided. At the very least, further study in both first-year courses as well as in the second-year research writing course is required to develop a more comprehensive picture of the

relationship between reading input and its effect on writing. It is hoped that this study has shed some light on the basic lexical features of the ELA Reader for the purpose of initiating a conversation about how we might best choose the most appropriate materials for our students.

References

Adolphs, S. (2006). *Introducing Electronic Text Analysis.* Oxon: Routledge.

Akano, I., Hori, M., & Tono, Y. (Eds). (2014). 英語教師のためのコーパス活用ガイド [*A practical guide to using corpora in English language teaching*] Tokyo:Taishukan-shoten.

Anthony, L. (2015, May). Applications of corpus linguistic in language materials design and in-class teaching and learning. Lecture presented at Temple University, Tokyo.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publishing Company.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5-35.

British Council. (2015). IELTS assessment criteria. Retrieved July 25, 2015, from http://takeielts.britishcouncil.org/find-out-about-results/ielts-assessment-criteria.

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

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30-49.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), 213-238.

Educational Testing Service. (2015). Writing scoring guide. Retrieved July 25, 2015, from https://www.ets.org/toefl/pbt/scores/writing_score_guide/.

English for Liberal Arts Program (2014). *ELA Staff Handbook 2014-2015*. Tokyo: International Christian University.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System*, 21, 231-244.

Granger, S. (Ed). (1998). *Learner English on computer*. New York: Addison Wesley Longman Limited.

Hinkel, E., (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, 37(2), 275-301.

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

Ishikawa, S. (2008). 英語コーパスと言語教育 [*English corpus and language teaching*] Tokyo: Taishukan-shoten.

Ishikawa, S. (2012). ベーシックコーパス言語学 [*Basics of corpus linguistic*] Tokyo:

Hitsuji-shobo.

Jarvis, Scott. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), 57-84.

Lancaster University. (2015). Compilation of the AmE06 reference corpus. Retrieved July 31, 2015, from http://www.research.lancs.ac.uk/portal/en/projects/compilation-of-the-ame06-reference-corpus%287d03286f-e363-427c-bc04-548b9834d8ba%29.html.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-322.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Martinez, I., Beck. & Panza, C. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, 28, 183-198.

McEnergy, T., & Hardie, A. (2012). *Corpus Linguistics*. New York: Cambridge University Press.

Masuko, M., Mizoguchi, S., Sano, M., Shiina, K., Thrasher, R., & Yoshioka, M. (1997). *English Vocabulary for Academic Purposes.*Tokyo: Liber Press.

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

Tono, Y. (2003). Learner corpora: design, development and applications. Proceedings of the 2003 Corpus Linguistics Conference, 800-809. Retrieved from http://www.tufs.ac.jp/ts/personal/corpuskun/pdf/2014/reading002.pdf

Tono, Y., Kaneko, T., Sugiura, M., & Izumi, E. (2013). 学習者コーパス活用ハンドブック [*Handbook of active use of learner corpus*] Tokyo: Taishukan-shoten.

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

Watanabe, Y. (2013). Profiling lexical features of teacher talk in CLIL courses: The case of higher education EAP program in Japan. *International CLIL Research Journal*, 2(1), 3-18.

Acknowledgements

I would like to thank Audrey Morrell and Jennie Roloff-Rothman for their valuable comments on an earlier version of this paper.

APPENDIX A

Table 1: Frequently appearing words

Rank	Frequency	Word
1	5958	the
2	3966	be
3	3951	of
4	3079	and
5	2792	to
6	2613	a
7	2194	in
8	1680	that
9	960	or
10	921	for
11	914	this
12	869	have
13	822	as
14	750	it
15	610	not
16	583	by
17	583	we
18	539	with
19	527	on
20	483	from
21	469	you
22	449	one
23	443	but
24	433	they
25	398	other
26	353	human
27	343	their
28	334	our
29	331	what
30	318	people
31	310	do
32	306	at
33	302	will
34	294	can
35	285	all
36	282	which
37	279	if
38	273	more
39	273	some
40	259	would
41	256	he
42	256	i
43	243	may
44	225	his
45	222	about
46	220	use
47	220	who
48	204	when
49	200	such
50	197	so
51	196	question
52	194	japanese
53	194	reason
54	192	make
55	187	no
56	186	conjurer
57	186	race
58	185	there
59	184	belief
60	180	than

Table 2: Keywords

Rank	Frequency	Keyness	Keyword
7	194	719.366	Japanese
8	353	658.389	human
9	184	631.959	belief
10	186	522.065	race
11	120	511.172	propaganda
13	139	485.568	perception
14	103	472.98	euthanaisa
15	141	471.182	communication
16	135	468.59	genetic
18	194	394.398	reason
20	196	344.633	question
22	77	301.804	perceive
23	156	298.952	college
24	172	298.34	person
26	66	292.528	bioethics
27	138	288.91	student
28	101	279.47	patient
29	57	267.877	slant
30	56	263.177	blumenback

Table 3: Type and token of texts in the ELA Reader

Author	Token	Type	TTR
Meiland	15667	2294	0.15
Morgan	2172	650	0.30
Larson	5443	1557	0.29
Cross	1214	491	0.40
Barna	6568	1874	0.29
Fisher	7278	1488	0.20
Birk & Birk	4161	1064	0.26
McDaniel	5409	1588	0.29
Gould	4427	1325	0.30
Shreeve	4045	1228	0.30
Diamond	4552	1170	0.26
Shipman	13118	3293	0.25
Shannon	1976	712	0.36
Shannon	5947	1342	0.23
Sandel	7474	1976	0.26
Morioka	4840	1219	0.25
Wennberg	5532	1485	0.27
Wagar	3910	1304	0.33
Ogata & Cels	3599	1174	0.33
Total	107332	10835	0.10

APPENDIX B

Table 4: Coverage of lexical items in the ELA Reader

Sprint Term	**Meiland**		**Morgan**		**Larson**		**Cross**	
TTR=0.14	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)
GSL 1st level 1000	83.67	50.96	84.79	67.38	72.5	49.23	78.91	64.42
GSL 2nd level 1000	4.58	9.98	5.13	10	5.15	10.09	3.57	7.16
AWL 570	7.02	18.35	3.79	8.31	6.67	12.47	2.13	4.5
Off-list words	5.86	20.71	6.29	14.31	15.69	28.21	15.39	23.93

Fall Term	**Barna**		**Fisher**		**Birk & Birk**		**McDaniel**		**Gcould**		**Shreeve**		**Diamond**		**Shipman**	
TTR=0.14	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)
GSL 1st level 1000	71.87	46.31	74.37	51.17	81.33	60.3	58.13	40	70.1	47.62	72.91	49.88	70.3	52.52	67.84	36.2
GSL 2nd level 1000	5.34	11.16	4.27	9.31	4.18	11.19	5.49	8.19	4.25	9.28	4.56	8.96	4.8	8.47	4.25	9.57
AWL 570	8.89	14.89	12.22	16.68	5.47	9.88	14.33	19.38	7.82	14.72	6.27	13.28	7.27	12.32	7.66	14.1
Off-list words	13.9	27.64	9.14	22.84	9.02	18.63	22.06	32.44	17.83	28.38	16.26	27.87	17.63	26.69	20.26	40.3

Winter Tearm	**Shannon & Kockler**		**Shannon**		**Sandel**		**Morioka**		**Wennberg**		**Wagar**		**Ogata & Cels**	
TTR=0.16	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)	Token (%)	Type (%)
GSL 1st level 1000	69.02	53.73	76.06	51.42	71.69	43.38	70.1	51.89	70.3	47.63	77.02	51.04	63.18	41.62
GSL 2nd level 1000	4.65	8.44	4.83	8.72	6.26	11.77	5.33	8.85	5.49	9.16	3.89	8.21	5.17	7.15
AWL 570	12.27	15.89	10.23	19.45	7.3	13.94	7.76	14.75	7.72	14.65	5.93	12.51	17.09	23.23
Off-list words	14.06	21.94	8.87	20.42	14.75	30.91	16.8	24.51	16.48	28.56	13.15	28.24	14.56	28

ELA Reader Total		
TTR=0.10	Token (%)	Type (%)
GSL 1st level 1000	73.11	24.39
GSL 2nd level 1000	4.66	10.99
AWL 570	8.42	13.7
Off-list words	13.81	50.92

APPENDIX C

Table 5: Frequently appearing academic words

ELA Reader		
Rank	Frequency	Words
1	216	perceive
2	213	culture
3	167	communicate
4	126	interpret
5	118	individual
5	118	secure
7	113	select
8	103	ethic
9	96	process
10	91	theory
11	79	identity
12	77	medical
13	76	vary
14	75	define
15	72	enhance
16	71	evident
17	68	principle
18	67	sex
18	67	similar
20	66	attitude
21	64	assume
22	61	text
23	60	create
23	60	institute
25	57	issue
25	57	require
27	56	psychology
28	55	theme
29	50	justify
30	49	method