Comparing Vocabulary Profiles of L2 Asian Written English in the ICNALE Corpus

Dax Thomas, Meiji Gakuin University, Japan

The IAFOR International Conference on Education – Hawaii 2019
Official Conference Proceedings

Abstract
This brief study examines the vocabulary profiles of Asian EFL students' written English in the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2018). The ICNALE corpus is a collection of written and spoken texts from 2600 learners of English across 10 different Asian countries. The texts included in this corpus were composed under controlled conditions (content and length) and are grouped by CEFR level (A2, B1-1, B1-2, B2+). The corpus also includes samples of native-speaker English. In this study, vocabulary profiles were constructed for written essays from the corpus along three bands - GSL1, GSL2, and AWL - using AntWordProfiler (Anthony, 2013). It was found that: 1) while there appeared to be a slight difference in AWL type percentages between low and high CEFR levels, these percentages varied much more greatly by country; 2) there was no statistical difference between the AWL type percentages of native speakers and those of CEFR levels B1-2, and B2+; and 3) when compared with the essays written by native speakers, those written by learners from Singapore and Hong Kong had higher percentages of AWL types, while those written by learners from the Philippines, Pakistan, and China showed no significant difference at all.

Keywords: AWL, corpus linguistics, ICNALE, GSL, vocabulary profiles

Introduction

Over the years, a large number of learner corpora have been constructed based on the written and spoken texts of non-native learners of English (Centre for English Corpus Linguistics, 2019). Generally, the purpose of these corpora is to give researchers access to larger amounts of authentic learner-produced English that can be analysed in place of, or compared with, their own students' English text collections. These corpora are useful in the study of error production as well as in grammar and structure analysis. They are also useful in the area of vocabulary research. Vocabulary profiles, a convenient way for instructors to evaluate the lexical development of their students, can be readily constructed from these corpora.

This study examines the vocabulary profiles - primarily Academic Word List (AWL; Coxhead, 2000) usage - of Asian EFL students' written English in the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2018). It shows that: 1) while differences in AWL type usage may seem to vary somewhat by CEFR level, they vary so much more dramatically according to country; 2) the AWL type usage of learners with CEFR levels B1-2, and B2+ was not significantly different from that of the native speakers; and 3) there was no significant difference between the AWL type usage in essays written by native speakers and learners in the Philippines, Pakistan and China, while learner essays from Singapore and Hong Kong had much higher AWL type usage scores.

The ICNALE Corpus

The ICNALE corpus is a collection of written and spoken texts from 2600 learners of English across 10 different Asian countries. It was developed by Dr. Shin Ishikawa of Kobe University, Japan, and is particularly valuable because of the great attention paid to both topic- and length- control during the production of the student texts it includes. Table 1 below shows the number of students per country in the corpus.

Country	*Texts*	*Country*	*Texts*
China	400	Pakistan	200
Hong Kong	100	The Philippines	200
Indonesia	200	Singapore	200
Japan	400	Thailand	400
Korea	300	Taiwan	200

Table 1: ICNALE composition

The corpus also includes 200 samples of native-speaker English texts (primarily from USA, UK, Australia, and Canada) for comparative purposes. Texts included in this corpus were composed under controlled conditions (content and length) and are grouped by CEFR level (A2, B1-1, B1-2, B2+; Figure 1).

This short study focused on the written English corpus and on just one of the two topics provided, namely the short essay relating to whether students should be permitted to have part-time jobs or not.

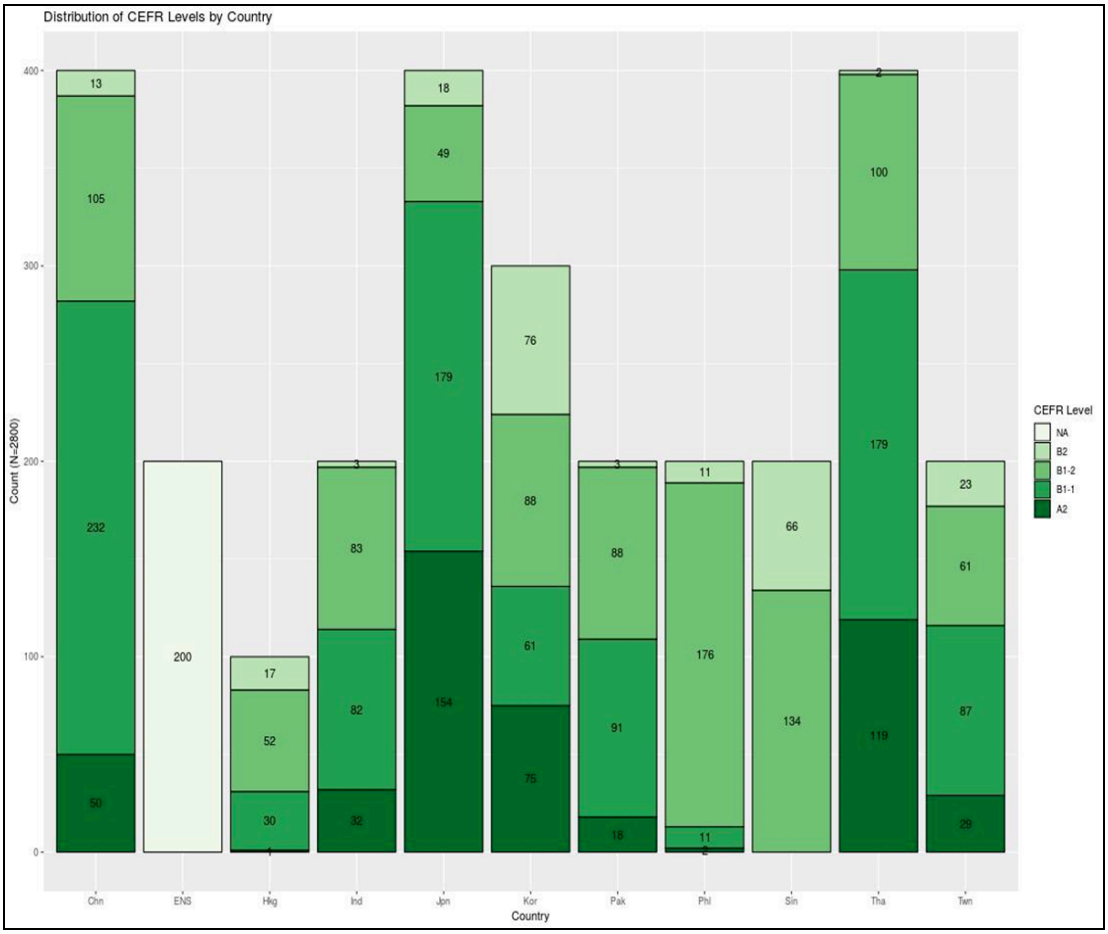

Figure 1: Distribution of CEFR levels across countries

Procedure

In this study, vocabulary profiles were constructed for each of the 2600 learner written essays, as well as the 200 native speaker written essays, using AntWordProfiler (Anthony, 2013). Three bands – GSL1, GSL2, and AWL – were used in the profile construction. The essays were then compared by CEFR level and by country.

The R scripting language was used to conduct a pairwise ANOVA test (with post-hoc Games-Howell) on the data and to build a boxplot for analysis. The data for the AWL was the main focus of this preliminary study.

Results and Discussion

AWL usage by CEFR level

The profiles were first compared by CEFR level. Tables 2 and 3 show the AWL type percentage means by CEFR level and pairwise ANOVA post-hoc results respectively. Figure 2 shows the same data graphically.

CEFR	*n*	*Means*	*Variances*
ENS	200	7.02	9.06
A2	480	4.52	6.62
B1_1	952	5.17	6.65
B1_2	936	6.75	10.62
B2+	232	7.82	14.35

Table 2: AWL type percentage means by CEFR

CEFR	*Diff*	*t*	*df*	*p-value*
B1_2-ENS	-0.266	1.12	307	1
B2-ENS	0.798	2.44	427	1
B2-B1_2	1.064	3.93	321	0.01
B1_1-A2	0.652	4.52	963	0.001
A2-ENS	-2.498	10.27	326	<.001
B1_1-ENS	-1.846	8.07	264	<.001
B1_2-A2	2.232	14.08	1181	<.001
B2-A2	3.296	11.98	337	<.001
B1_2-B1_1	1.58	11.67	1778	<.001
B2-B1_1	2.644	10.08	285	<.001

Table 3: ANOVA post-hoc

As can be seen, there was a statistically significant difference between the means of all levels with the exception of the highest two B levels and the native speaker group. At first glance, this seems to indicate that CEFR level does have some bearing on degree of AWL usage. However, according to the ICNALE supporting material:

The ICNALE team has required all the learners to take a standard L2 vocabulary size test (VST) covering the top 5K word levels (Nation & Beglar, 2007), and also to present their scores in the high-stake English proficiency tests such as TOEFL and TOEIC. Then, all the learners have been classified into four kinds of CEFR-linked proficiency bands: A2, B1_1 (B1 low), B1_2 (B1 high), and B2+, based on their scores in the proficiency tests ***or in the VST***. (Ishikawa, 2018, para 2.4; emphasis added)

It is then difficult to claim that CEFR level influences AWL type with any certainty when it may be that vocabulary level was what influenced the corpus designers' CEFR level assignment to individual students in the first place. Further investigation into the degree to which the VST influenced CEFR assignment, as well as the relationship of the VST to the AWL, is warranted.

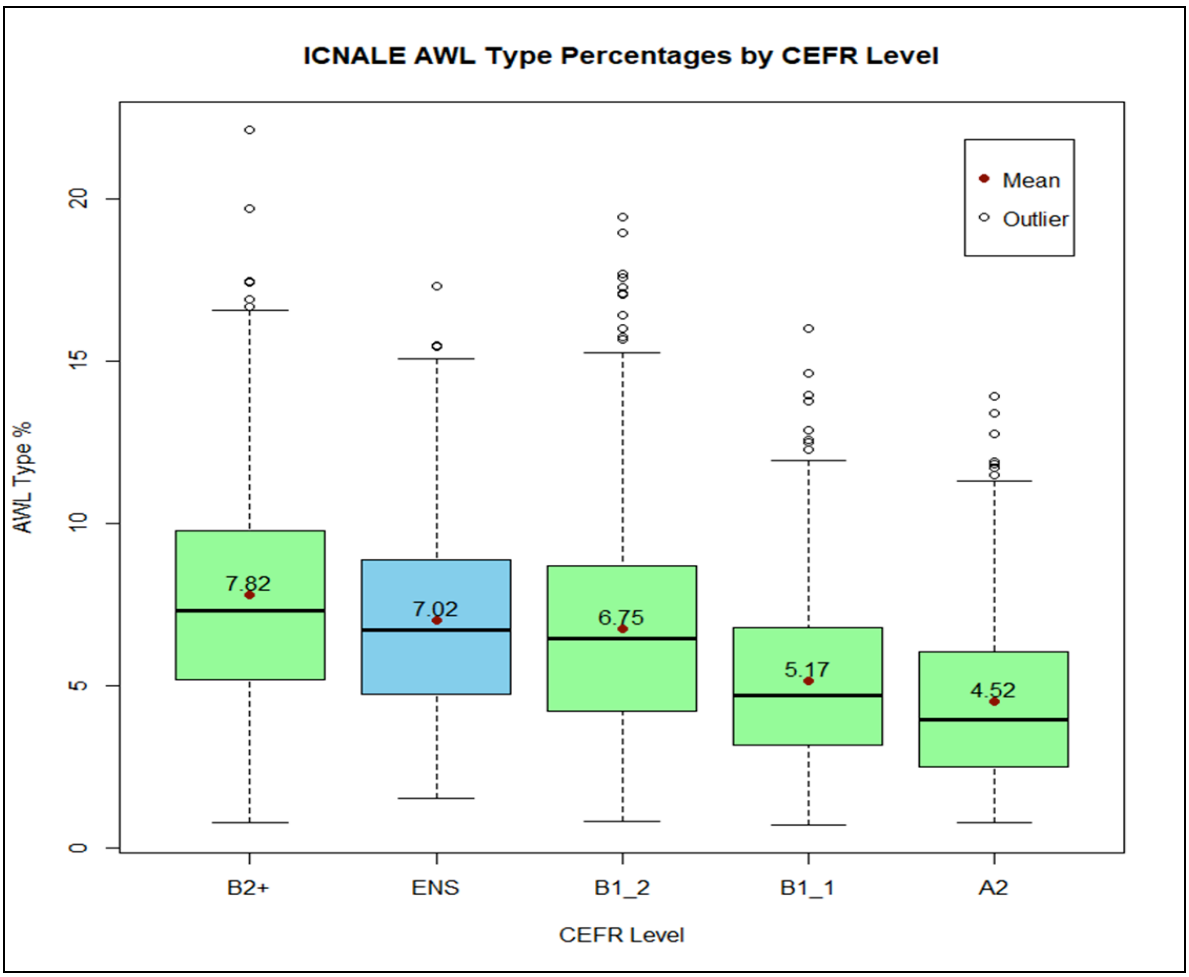

Figure 2: AWL usage by CEFR level

AWL usage by country

Next, AWL type usage was compared by country. Tables 4 and 5 show the AWL type percentage means by country and pairwise ANOVA post-hoc partial results (non-statistically-significant only) respectively. Figure 3 shows the same data graphically.

Country	*n*	*Means*	*Variances*
Chn	400	6.12	6.09
ENS	200	7.02	9.06
Hkg	100	8.73	9.25
Ind	200	5.57	7.38
Jpn	400	3.43	2.94
Kor	300	5.45	7.06
Pak	200	6.59	6.58
Phl	200	7.17	6.79
Sin	200	10.62	10.91
Tha	400	4.55	6.03
Twn	200	5.43	6.92

Table 4: AWL type percentage means by country

Countries	*Diff*	*t*	*df*	*p-value*
Ind-Chn	-0.5493	2.4055	366	1
Kor-Chn	-0.666	3.3834	617	1
Pak-Chn	0.4768	2.1728	385	1
Twn-Chn	-0.6883	3.0827	376	1
Pak-ENS	-0.4226	1.5107	388	1
Phl-ENS	0.1506	0.5352	390	1
Kor-Ind	-0.1167	0.4747	420	1
Twn-Ind	-0.139	0.5195	398	1
Twn-Kor	-0.0222	0.0923	429	1
Phl-Pak	0.5732	2.2169	398	1
ENS-Chn	0.8993	3.6552	336	0.723
Pak-Ind	1.0261	3.8826	397	0.311
Twn-Tha	0.8757	3.9282	375	0.264
Twn-Tha	0.8757	3.9282	375	0.264
Phl-Hkg	-1.5641	4.3994	173	0.052

Table 5: ANOVA post-hoc, non-statistically-significant differences only

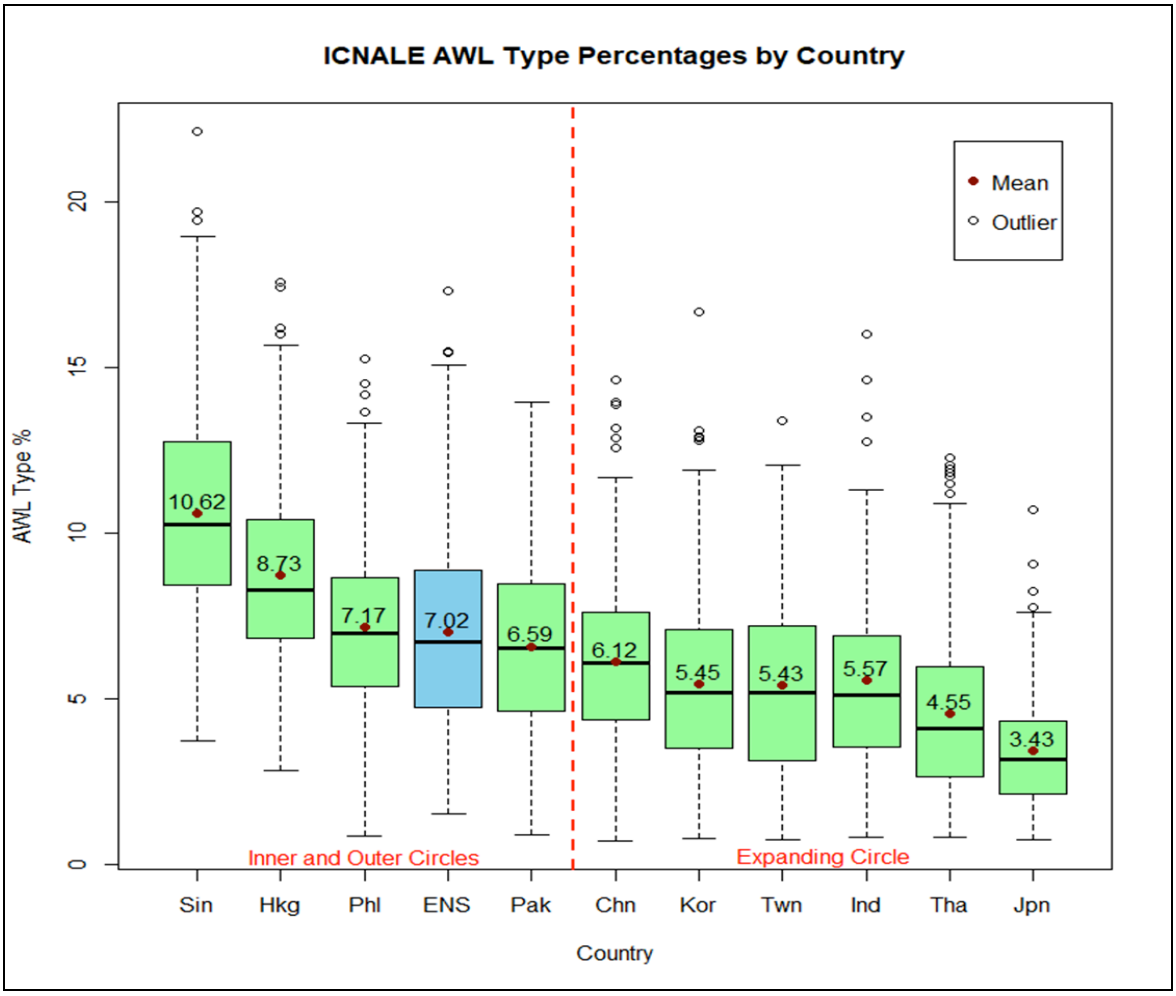

Figure 3: AWL type percentages by country

As can be seen, the boxes for each country arrange themselves neatly, with those from Inner and Outer Circle countries falling on the left side of the plot (higher AWL type usage) and those from the Expanding Circle countries falling on the right side (lower AWL type usage). Also of note, is that the essays written by Japanese learners had the lowest percentage of AWL types in the entire corpus while those written by learners in Singapore had the highest.

It is interesting to discover that the students from Singapore and Hong Kong had higher AWL type usage percentages than the native speakers. At this stage, it is unclear why this is the case, but one might speculate that it may relate to the essay writers' understanding of the essay register. It could be that the writers in Singapore and Hong Kong took a much more formal approach to the essay (thus including more AWL vocabulary) while the native speakers took a somewhat more casual approach. Degree of formality is something that should be marked for future study.

Finally, there was no significant difference between native speaker data and those of Pakistan, Philippines and China. This is an unexpected result but, as above, may be caused by the writers' understanding of what was expected in terms of the essay's register.

Conclusion

This preliminary study looked at the vocabulary profiles – primarily the use of AWL vocabulary – of Asian EFL/ESL learners' written essays in the ICNALE corpus. AWL type percentage means were compared by CEFR level and by country. It was found that, while there was some possible difference in AWL usage by CEFR level (complicated by the method of assigning CEFR scores to learners), there seemed to be a much clearer difference in usage when compared by country. The AWL usage percentages ranged from the highest in Singapore and Hong Kong, to the lowest in Thailand and Japan, and scores from some countries were higher or equal to those of native speakers.

Future study should look more closely at the relationship between the CEFR leveling criteria and the AWL, as well as at the specific register of the individual essays, as both of these may have an effect on AWL type usage.

References

Anthony, L. (2013). AntWordProfiler (Version1.4.0w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Centre for English Corpus Linguistics (2019). Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. Available from https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, Vol. 34, No.2 (Summer, 2000), pp. 213-238.

Ishikawa, S. (2018). ICNALE: The International Corpus Network of Asian Learners of English. Available from http://language.sakura.ne.jp/icnale/

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.

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