L2 영어 학습자들의 연어 사용 능숙도와 텍스트 질 사이의 수치화

권준혁0, 김재준, 김유래, 박명관, 송상헌

동국대학교, 인천대학교

Kyunjh1272@gmail.com

Quantifying L2ers' phraseological competence and text quality in L2 English writing

Junhyeok Kwon0, Jaejun Kim, Yoolae Kim, Myung-Kwan Park (Dongguk University)

Sanghoun Song (Incheon National University)

Abstract

On the basis of studies that show multi-word combinations, that is the field of phraseology, this study aims to examine relationship between the quality of text and phraseological competence in L2 English writing, following Yves Bestegen et al. (2014). Using two different association scores, t-score and Mutual Information(MI), which are opposite ways of measuring phraseological competence, in terms of scoring frequency and infrequency, bigrams from L2 writers' text scored based on a reference corpus, GloWbE (Corpus of Global Web based English). On a cross-sectional approach, we propose that the quality of the essays and the mean MI score of the bigram extracted from YELC, Yonsei English Learner Corpus, correlated to each other. The negative scores of bigrams are also correlated with the quality of the essays in the way that these bigrams are absent from the reference corpus, that is mostly ungrammatical. It indicates that increase in the proportion of the negative scored bigrams debases the quality of essays. The conclusion shows the quality of the essays scored by MI and t-score on cross-sectional approach, and application to teaching method and assessment for second language writing proficiency.

Key words: phraseological competence, corpus, n-grams, Collgrams, quality of essay

1. L2ers' academic writing based on phraseology

Traditionally, the focus in second language acquisition has been on how L2 learners acquire morphological and grammatical knowledge rather than other levels of language. Although the focus on grammatical knowledge for l2 leaners is pervasive, recently corpus linguistic research has taken lexis to a central role in second language acquisition. Since corpus linguistics is mainly lexical, which is based on corpora from real world text, it is easy to deal with lexical items and a sequence of lexico-grammatical patterns.

Lewis' (1993) idea that "language consists of grammaticalized lexis, not lexicalized grammar" has shed light on lexicalization in language teaching. The notion of lexis is basically phraseological, which is on the premise that lexis is not just the study of single word, but phraseology is "the study of the structure, meaning and use of word combinations" (Cowie, 1994). Language production, including writing and speaking, is largely relies on pre-patterned phrases which are prefabricated in L1; therefore, the role of phraseology is quite essential for L2 learners.

L2 writers generally tend to use only limited repertoire of collocations that they mastered, but very little amount of native-like lexical bundles. "Failure to use native-like formulaic sequences is on factor in making their writing feel nonnative" (Li and Schmitt, 2009). As pointed out above, recently the focus on formulaic sequences in SLA has been issues. Formulaic sequences are contiguous word sequences which are stored and retrieved as one unit from memory, and they are rather than being generated by single words and their structures, considered as whole sequence at the time of use. Coxhead and Byrd (2007) enumerate three reasons for a focus on formulaic sequences in L2 academic writing: (1) for students, using ready-made formulaic sequences is easy; (2) formulaic sequences define standards of fluent academic writing; (3) formulaic sequences are easier to detect on the basis of corpus data. Besides these three reasons, using corpus data makes it possible to quantify its data and help search significant formulaic sequences for academic writing, and also those significant formulaic sequences are reliable in terms of 'usage based learning'.

2. N-gram method and problems with using frequent n-grams

N-grams are continuous word sequences of n items from a given text. A size of n-grams increases by one, and each of n-grams are referred to uni-grams, bi-grams, tri-grams, four-grams, and so on. Extracting n-grams from corpus data is one of the effective ways to detect formulaic sequences. From an extracted data, counting the same word combinations of n-grams shows how some kinds of n-grams are used in a text. However, focuses on the frequency of those n-grams only pay attention to absolute frequency, not the degree of relation within the given n-grams; for instance, bigrams 'of the' or 'it is' could be one of the most frequent bigrams from any texts in the real world. Both 'of' and 'the' or 'it' and 'is' are very frequent words themselves, but this high frequency does not present evidence for relevant phraseological status.

3. Mutual Information, t-score, and CollGrams

Since high frequency of n-grams only shows the absolute frequency of n-grams, frequency does not detect significant n-grams. To figure out the association within each n-grams, all of the n-grams extracted from a given text are assigned each MI (Mutual Information) and t-score, and each of the n-grams formulate CollGrams, which form the basis of evidence for phraseological association within the n-grams.

MI is a measurement for strength of association, which originated from information theory. The higher MI score each n-gram gets, the more infrequent words sequences appear within the n-grams. MI accords with the log transformed ratio between observed frequency of n-grams, including not only single words and word sequences, and its expected frequency. Following is the formula of MI from Church & Hanks, 1990:

$$Mutual\ Information = log2\frac{Freq\ xy}{(Freq\ x\ \times Freq\ y)/N}$$

(Church & Hanks, 1990)

On the other hand, t-score represents certainty of n-grams. The higher t-score each n-grams gets, the more frequent words sequences appear within the n-grams. t-score is the expected frequency of the square root of the observed frequency, including frequency of single words and word sequences. Following formula is from Church et al., 1991:

$$t \approx \frac{Freq(xy) - \frac{1}{N}Freq(x)Freq(y)}{\sqrt{Freq(xy)}}$$

(Church et al., 1991)

Each of the n-grams assigned MI and t-score forms CollGrams, which show their collocational status. This CollGram technique solve the problem of using frequent n-grams.

4. Previous study: Yves Bestegen and Sylviane Granger (2014)

In their study, they aim to assess L2 text quality in terms of phraseological competence. Using Collgrams, they calculated each bigram from the Michigan State University corpus of English based on the reference corpus, COCA (Corpus of Contemporary American English). The students who participated for making MSU corpus of English wrote essays three times from the beginning of a semester to the end of the semester, which made this study possible to search on both longitudinal and cross-sectional approach.

The results of the longitudinal study showed a decrease in mean t-score, which explains that L2 learners acquire more complex collocations and idioms instead of low-level binary chunks.

On the other hand, the results of the cross-sectional study presented that only MI and absent category from the reference corpus are statistically significant in terms of correlation between the quality of the texts and the mean scores. They assumed that bigrams consist of low frequency words might be more noticeable to raters who judged the texts, and positively influence their judgement.

5. Methodological approach and data

The technique, CollGram, aims to assign to each bigram association scores computed on the basis of a reference corpus. If a bigram extracted from a learner corpus does not exist in a reference corpus, it is classified in an absent category. CollGrams consist of three measures: the mean MI score, the mean t-score, and the proportion of absent bigrams from the reference corpus. Steps for profiling CollGrams are followed: (1) both the learner and the reference corpus are tokenized except for punctuation marks; (2) all bigrams extracted from the learner corpus; (3) each bigram extracted from the learner corpus are looked up in the reference corpus; (4) the bigrams existed in the reference corpus are assigned their MI and t-score; (5) if the bigrams do not exist in the reference corpus, they are

classified in the absent category.

In this study, Yonsei English Learner Corpus (YELC) is used for the learner corpus. YLEC is collected from January 2011 to February 2011. YELC where 3,286 freshmen at Yonsei University participated contains 1,081,280 words.

Details of YELC learner corpus	Beginner	Intermediate	Advanced
Number of texts	910	2256	120
Total number of words	233661	798907	48712
Number of words per proficiency (mean)	265.77	354.12	405.93
Rating	A1/A1+/A2	B1/B1+/B2	B2+/C1/C2

(*Table 1. details of YELC*)

All the texts from YELC are rated from A1 to C2, and for the study we classified the nine different ranks into three proficiencies: beginner, intermediate, and advanced. Total number of words from each proficiency shows that intermediate-level contains the largest words because more than the half of the texts rated from B1 to B2, but average words per each proficiency present that texts rated advanced-level have the largest words; 405.93 words per a text.

As to the reference corpus, we opted for Global Web based English Corpus (GloWbE), which has more than 1.9 billion words. About 60% of the corpus are from blogs, which means it is very informal (one of the reasons we opted for the corpus as a reference corpus). From the varieties of English in the corpus, we chose U.S. English since most of the Korean students are exposed to American English in Korea.

6. Results: highest-scoring bigrams and the absent category

Table below lists the top 20 highest-scoring bigrams in YELC learner corpus, advanced-level, beginner-level, respectively. The left-hand side of the table are sorted in decreasing order of the MI score. Many of the bigrams with high MI scores consist of either infrequent verbs and definite article 'the', or infrequent verbs and infinitive to (or prepositions). On the other hand, the right-hand side of the table shows the bigrams with high t-score which are composed of frequent prepositions and definite article 'the', or pronouns and verbs.

Top-scoring bigrams	MI	Top-scoring bigrams	t	Top-scoring bigrams	MI	Top-scoring Bigrams	t
circulates_the	40.4014	of_the	128.5095	refer_to	39.3586	i_think	62.4293
outweigh_the	40.0795	on_the	121.813	ought_to	39.3285	i_am	62.2972
throughout_the	39.7033	to_be	92.3940	disposing_of	39.2940	this_is	61.7925
violates_the	39.6424	on_the	81.7483	consequences_of	39.2940	at_the	61.2471
outweighs_the	39.5941	it_is	73.8763	opposed_to	39.2940	will_be	60.9732
according_to	39.5681	if_you	73.8233	tries_to	39.2688	i_have	60.6789
tends_to	39.5496	is_a	64.9545	refuse_to	39.2576	want_to	59.2859
able_to	39.5468	it_was	64.5843	due_to	39.2533	the_same	59.2609
subjeted_to	39.5289	for_the	63.5135	intend_to	39.2523	as_a	58.9083
tend_to	39.5020	i_do	62.75	distorting_the	39.2467	going_to	58.5038

(*Table2. top 20 highest-scoring bigrams in advanced-level*)

Top-scoring bigrams	MI	Top-scoring bigrams	t	Top-scoring bigrams	MI	Top-scoring bigrams	t
distorts_the	40.4014	of_the	128.5095	coffees_and	39.4517	i_think	62.4293
embodies_the	39.9420	in_the	121.813	embarrased_and	39.4517	i_am	62.2972
displace_the	39.8164	to_be	92.3940	drugging_and	39.4517	this_is	61.7925
pollute_the	39.8164	on_the	81.7483	willing_to	39.4437	at_the	61.2471
according_to	39.5681	it_is	73.8763	accustomed_to	39.4357	will_be	60.9732
able_to	39.5468	if_you	73.8233	supposed_to	39.4219	i_have	60.6789
unable_to	39.5200	is_a	64.9545	contaminate_the	39.4014	want_to	59.2859
tend_to	39.5020	it_was	64.5843	diminishes_the	39.4014	the_same	59.2609
lessen_the	39.4628	for_the	63.5135	unify_the	39.4014	as_a	58.9083
relating_to	39.4565	i_do	62.75	trying_to	39.3967	going_to	58.5038

(*Table3. top 20 highest-scoring bigrams in advanced-level*)

The other category that needs to be analyzed closely in terms of pedagogy is the absent category in the reference corpus. Since the bigrams in the learner corpus are absent in the reference corpus, this shows what types of erroneous bigrams are generated from L2 writers. The proportion of the absent category also represents quality of the given texts in accordance with CollGrams. In many cases, L2 writers made mistakes using two determiners at one time, and also, they made many erroneous bigrams with inappropriate prepositions. Following table shows randomly selected erroneous bigrams which are absent in the reference corpus:

W1W2	sentences	Error
a this	No I disagree a this opinion because that is a important thing	article
a students	Physical punishment can be a good medicine for a students who don't have a strong will	article
went Japan	I went Japan with my 4 best friends .	preposition
is need	I think physical punishment is need for them	past participle
animals	I think animals be used in medical experiments is need.	be verb
not use	so I want drivers not use their phone while driving.	to infinitive
doesn't free	Yet my country Korea doesn't free from fright that includes North Korea's threat.	negation

(*Table4. randomly selected erroneous bigrams in the absent category*)

7. discussion

Mean score	Beginner	Advanced
MI(Mean)	25.8309	26.4923
MI(SD)	6.5826	6.7566
t(Mean)	0.4526	2.6722
t(SD)	14.3550	12.4856
Number of absent category	42459	8657

(*Table5. Difference between Beginner and Advanced learners*)

The increase in the mean MI and the mean t-score indicates that the higher score L2ers get, the higher quality of the texts L2ers write. The decrease in the proportion of the absent category also presents that advanced-level L2ers made less erroneous formulaic sequences. These three indices constitute CollGrams, and it shows that both of infrequent and frequent formulaic sequences get higher score on each measurement.

The way to calculate CollGrams of the learner corpus based on the large reference corpus is a text-external measure. The text-external measure leads to operationalize formulaicity of L2 writing, which is helpful to make L2 writers more native-like and also for assessment.

For developing this study, we need to adopt statistical inferences to analyze the given results of CollGrams, and also the 'analysis of variance' (ANOVA) to measure the correlation between given results and rated text quality.

Reference

1. Bestgen, Yves, and Sylviane Granger. "Quantifying the development of phraseological competence in L2 English writing: An automated approach." *Journal of Second Language Writing* 26 (2014): 28-41.

2. Church, Kenneth Ward, and Patrick Hanks. "Word association norms, mutual information, and lexicography." *Computational linguistics* 16.1 (1990): 22-29.

3. Church, Kenneth, et al. "Using statistics in lexical analysis." *Lexical acquisition: exploiting on-line resources to build a lexicon* 115 (1991): 164.

4. Cowie, A. P. "Applied linguistics: lexicology." *Encyclopedia of Language and Linguistics. Pergamon, Oxford* (1994): 177-180.

5. Coxhead, Averil, and Pat Byrd. "Preparing writing teachers to teach the vocabulary and grammar of academic prose." *Journal of second language writing* 16.3 (2007): 129-147.

6. Lewis, Michael, and Cherry Gough. *Implementing the lexical approach: Putting theory into practice*. Vol. 3. No. 1. Hove: Language Teaching Publications, 1997.

7. Li, Jie, and Norbert Schmitt. "The acquisition of lexical phrases in academic writing: A longitudinal case study." *Journal of Second Language Writing* 18.2 (2009): 85-102.

8. Rhee, S., and C. Jung. "Yonsei English learner corpus (YELC)." *Proceedings of the First Yonsei English Corpus Symposium*. 2012.