Is native speaker intuition reliable for high-frequency context creation?

関西外国語大学 研究論集 第102号（2015年 9 月）
Journal of Inquiry and Research, No.102 (September 2015)

Is native speaker intuition reliable for high-frequency context creation?

James M. Rogers

Frank E. Daulton

Ian B. MacLean

Gordon A. Reid

Abstract

This study determined whether native speaker intuition could be relied upon to produce contextual content that mostly fell into what is considered high-frequency vocabulary. Native speakers wrote over 160,000 tokens worth of example sentences for high-frequency multi-word units derived from a corpus. The resulting database was examined to determine whether the content added by the native speakers mostly stayed within the high-frequency realm.

Results showed that not only did the vast majority of native speakers' tokens fall into the high-frequency realm, the percentage that fell into the high-frequency realm only dropped by 0.84 percent in comparison to the multi-word units alone despite the large amount of data being added. This study highlighted how the intuition of experienced ESL practitioners can be relied upon to produce high-frequency contextual content.

Keywords: vocabulary acquisition, formulaic language, corpora, native speaker intuition, high-frequency vocabulary

INTRODUCTION

Corpora can no doubt help improve upon our ability to select useful language to teach to second language learners. However, while corpora can certainly help inform singular vocabulary, collocation and formulaic language choices, the value of a native speaker's intuition should not be discounted. Corpora, by their very nature, are not perfect. For

some tasks, it may actually be preferable to rely on a native speaker. For instance, a native speaker's intuition may be more reliable when the task is to create example sentences to help teach keywords or formulaic phrases because the native speaker can take into account word frequencies in comparison with the target keywords/formulaic phrases. This is the key to helping students learn how a word or phrase is used in proper context while not increasing the learning burden of the item.

However no previous research has examined on a large scale the extent to which a native speaker's intuition can be relied upon to create example sentences whose contents mostly fall into the high-frequency realm. Thus this paper will examine the type of data native speakers create when writing example sentences for high-frequency formulaic sequences, using mostly high-frequency language while still producing natural, appropriate examples.

LITERATURE BACKGROUND

Many researchers agree that mastery of high-frequency vocabulary is key to second language fluency. This is because such vocabulary can cover 80 percent or more of the words in most texts (Nation, 2008). Because of the practical limitations of classroom time, Nation (2001) believes that only high-frequency vocabulary deserves direct teaching time and recommends a cut-off point at 3,000 word families.

But how should such vocabulary be taught? Learning collocations rather than isolated words has been found to actually be easier (Ellis, 2001; Lewis, 2000). For example, Bogaards (2001) found that multi-word expressions containing familiar words were retained 10 percent more than completely new single words immediately after a learning session and also 12.1 percent more in a delayed posttest three weeks later. Researchers also agree that formulaic language knowledge is key to native-like fluency in a second language (Cowie, 1998; Wray, 2002).

However, semantic knowledge of a word and knowledge of its common collocations and the formulaic sequences it commonly occurs in is still not enough for a learner to truly master a word. There is, in fact, a "wide range of lexical knowledge" (Schmitt, 2010, p. 152) that must be mastered to truly know a word. Nation (2001) refers to this as vocabulary depth knowledge. In addition to semantic and collocational knowledge, vocabulary depth knowledge also includes knowing constraints on word use. One of the key ways in which learners can master constraints on word use is by learning vocabulary, collocations and the

formulaic sequences they co-occur in within context (i.e., example sentences).

So if we are to provide learners with examples of contextualized high-frequency vocabulary/collocation, where should these example sentences be sourced from? Corpora can be tapped for this since it provides authentic examples of keywords in context. For instance, the Corpus of Contemporary American English's (COCA) (Davies, 2008) online interface allows for users to conduct searches of collocations to give users access to example sentences in which both collocates occur. However, this method is not ideal for second language learners because the material's creator has no control over the frequency of the contextualized sentence. For instance, if the keyword(s) being taught are in the 3,000 word frequency band, and words in the example sentence are very low frequency, a context may actually add to the learning burden of the item, the exact opposite of its purpose.

However, this issue can be easily circumvented by relying on native speaker intuition to create example sentences. Example sentence creation by native speakers with specific proficiency levels in mind may be preferable in comparison to sifting through thousands of examples of keywords in context from a corpus to find an appropriate contextualized example for students. But can native speaker intuition be as reliable as corpus data? For keyword vocabulary selection for direct instruction to learners, native speaker intuition has been proven to be reliable to a large extent (Rogers, 2010). It has even been shown to be essential in comparison with corpus data alone when judgments on whether a collocation should be considered as having balanced range, and also whether a collocation should be considered chronologically stable (Rogers et al., in press). However, researchers have yet to examine the reliability of native speaker intuition for context creation on a large scale that mostly falls into the high-frequency realm. Thus this study aims to make it salient whether or not a native speaker can be relied upon for such a task.

RESEARCH QUESTION

Can native speaker intuition be relied upon on a large-scale to create example sentences whose contents fall into the high-frequency realm?

MATERIALS

This study began utilizing Rogers et al.'s (in press) list of 12,604 high-frequency

lemmatized concgrams. A *lemma*, as defined by Nation and Meara (2002, p. 36), is a "set of related words consisting of the stem and inflected forms that are all the same part of speech". For instance, the verbs *run, runs, running* and *ran*, are all counted together as one lemma, while the noun *run* is counted separately. This is in comparison to counting words as *word families*, which is the larger grouping of all parts of speech of a word. Bauer and Nation (1993) provide a technical definition of word families as "a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (p. 1). A *concgram* "constitutes all the permutations of constituency and positional variation generated by the association of two or more words" (Cheng, Greaves, and Warren, 2006, p. 411). *Constituency variation* (AB, ACB) is when a pair of words not only co-occur adjacent to each other (*lose weight*) but also with a constituent (*lose some weight*). *Positional variation* (AB, BA) counts total occurrences of two or more particular lexical items, including occurrences on either side of each other. Thus, *provide you support* and *support you provide* would both be included in the total counts for a multi-word unit concordance search for the lemma *provide* and *support*.

This list was originally derived from Davies' (2010) *Word List Plus Collocates*, a list of collocations that occur with the most frequent 5,000 lemma of the COCA. To identify only items from this list that are useful for learners of general English, Rogers et al. (in press) delimited the list by frequency (approximately one occurrence per million tokens), and only included items with balanced range and chronological data.

Concordance data for each of the 12,604 concgrams was collected from the COCA. This study's unique approach necessitated the writing of custom concordance software to identify the most common multi-word units. Using normal concordance software, such as Anthony's (2011) *AntConc*, was not an option because the data this study aimed to identify were only multi-word units in which both lemma occurred in, a function not possible with *AntConc* or other concordance software. For example, examining 500 example sentences which all contain both the lemma *take* and *break* with *AntConc* would not reveal *take a break* as the most common multi-word unit, but rather various unrelated common multi-word units first, such as *of the*, etc. This is not ideal because removing such "noise" from the data would prove extremely time-consuming. Furthermore, the large amount of data (over 12,000 pairs) required a batch processing option, another feature not present on currently available concordance software. Thus, this study used the custom concordance software *AntWordPairs* (Anthony, 2013), a program written specifically for the purpose of this study.

This software utilizes Someya's (1998) *E-lemma list*. For coding purposes, Someya's lemma list could not contain duplicate entries, and thus was modified to remove any homonyms.

PROCEDURE

The first step of this study was to collect concordance data (example sentences) for each of the 12,604 lemma pairs. Lemmatized concordance searches were conducted using the COCA's online interface to collect data for instances when the collocate occurred either three words to the left or right of the node word. The rationale for this length (7 words) was influenced by findings on typical human memory limitations (Miller, 1956).

Next, 500 example sentences for each lemma pair were then collected from the COCA. Extracting 500 example sentences per lemma pair essentially created a mini corpus consisting of approximately 13,000 words per pair. Once this was complete, the data was then processed with *AntWordPairs* to identify the most common formulaic sequences each lemma pair occurs in. Because the amount of resulting data was unnecessarily copious, only formulaic sequences occurring in 5 percent or more of the corpora were collected.

After that, the data was examined to not only extract the most frequent formulaic sequence, but to also extend the sequence beyond the most frequent item to its left or right when the native speaker judged any additions to be part of the natural unit. For instance, the most frequent sequence for the lemma pair *come* and *term* was found to be *come to terms* at 243 occurrences (see Table 1 below). However, the next most common string in the data beyond *come to terms* was *come to terms with* (229 occurrences), then *to come to terms* (133 occurrences), and beyond that, *to come to terms with* (129 occurrences). *To come to terms with* was thus identified as the formulaic sequence most representative of the lemma pair *come* and *term* using native speaker intuition. In other words, the raw data indicates that *with* and then *to* are the next most common strings and a native speaker used his/her judgment to determine whether this corpus data matches his/her knowledge of what is or is not typical usage. Core multi-word units were identified in bold and any strings present in the data and also judged to be typically co-occurring with the multi-word unit were added in italics. To accomplish this, native speakers used their intuition to only add strings to the core formulaic sequence that truly represented common usage, but that also provided learners with useful information.

Table 1

Formulaic sequences identified from 500 example sentences in which the lemma pair come and term both occur in.

Multi-word unit	Occurrences in 500 sentences
come to terms	243
come to terms *with*	229
to **come to terms**	133
to **come to terms** *with*	129
coming to terms	96
coming to terms with the	86
to **come to terms** *with the*	44
come to terms *with* [pre-nominal possessive pronoun]	28
coming to terms with the	26

Subsequently, the 12,604 lemma pairs were distributed among the four native speakers—two Americans and two Canadians—who wrote an example sentence for each lemmatized concgram. These native speakers are experienced ESL practitioners, each with ten years or more experience teaching English as a second language. Each native speaker was instructed to choose high-frequency contextualized context when possible while still creating natural and appropriate sentences. Essentially, the goal of the native speaker was to create an example sentence that did not increase learning burden, but rather lowered the burden while also highlighting an item's typical usage in the language.

Then, the formulaic sequences alone were processed with Heatley, Nation and Coxhead's (2002) *RANGE* program to determine the extent to which the contents fell into the high-frequency realm. After that, the same analysis was repeated, but with the formulaic sequences within the example sentences created by the native speakers. The results were compared to each other. Finally, the formulaic sequences within the example sentences were processed with Cobb's (2013) *Vocabprofiler* to specifically determine which of the top 3,000 word families were not covered by the data.

RESULTS

Example sentences by all four native speakers were combined, which in total consisted of 160,932 tokens.

The formulaic sequences alone and the formulaic sequences with the example sentences were examined using *RANGE*, and Tables 2 and 3 below show their coverage of the top 34 groups of 1,000 word families of English.

Table 2

Word family frequency breakdown of formulaic phrases using RANGE

Word Family Frequency Level	Total Tokens / %	Total Types / %	Families
1	25369/78.01	1,940/43.98	923
2	4,497/13.83	1,204/27.30	721
3	2,078/ 6.39	808/18.32	588
4	277/ 0.85	216/ 4.90	203
5	99/ 0.30	85/ 1.93	85
6	39/ 0.12	34/ 0.77	32
7	10/ 0.03	10/ 0.23	10
8	11/ 0.03	11/ 0.25	10
9	4/ 0.01	4/ 0.09	4
10	0/ 0.00	0/ 0.00	0
11	5/ 0.02	4/ 0.09	4
12	3/ 0.01	2/ 0.05	2
13	1/ 0.00	1/ 0.02	1
14	1/ 0.00	1/ 0.02	1
15	0/ 0.00	0/ 0.00	0
16	0/ 0.00	0/ 0.00	0
17	0/ 0.00	0/ 0.00	0
18	0/ 0.00	0/ 0.00	0
19	0/ 0.00	0/ 0.00	0

20	0/ 0.00	0/ 0.00	0
21	0/ 0.00	0/ 0.00	0
22	0/ 0.00	0/ 0.00	0
23	0/ 0.00	0/ 0.00	0
24	0/ 0.00	0/ 0.00	0
25	0/ 0.00	0/ 0.00	0
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0
29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	6/ 0.02	3/ 0.07	3
32	3/ 0.01	3/ 0.07	3
33	64/ 0.20	40/ 0.91	38
34	0/ 0.00	0/ 0.00	0
not in the lists	55/ 0.17	45/ 1.02	
Total	32,522	4,411	2,628

Table 3

Word family frequency breakdown of formulaic phrases within example sentences created using native speaker intuition using RANGE

Word Family Frequency Level	Total Tokens / %	Total Types / %	Families
1	138,167/85.88	2,660/33.81	985
2	13,218/ 8.21	1,969/25.03	899
3	5,315/ 3.30	1,359/17.27	785
4	1,129/ 0.70	559/ 7.10	452
5	663/ 0.41	281/ 3.57	245
6	237/ 0.15	143/ 1.82	127
7	105/ 0.07	75/ 0.95	69

8	92/ 0.06	52/ 0.66	49
9	45/ 0.03	34/ 0.43	34
10	35/ 0.02	26/ 0.33	25
11	25/ 0.02	13/ 0.17	12
12	20/ 0.01	10/ 0.13	8
13	6/ 0.00	5/ 0.06	4
14	6/ 0.00	5/ 0.06	5
15	1/ 0.00	1/ 0.01	1
16	1/ 0.00	1/ 0.01	1
17	1/ 0.00	1/ 0.01	1
18	2/ 0.00	2/ 0.03	2
19	0/ 0.00	0/ 0.00	0
20	1/ 0.00	1/ 0.01	1
21	0/ 0.00	0/ 0.00	0
22	1/ 0.00	1/ 0.01	1
23	0/ 0.00	0/ 0.00	0
24	0/ 0.00	0/ 0.00	0
25	0/ 0.00	0/ 0.00	0
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0
29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	760/ 0.47	250/ 3.18	228
32	90/ 0.06	14/ 0.18	10
33	743/ 0.46	223/ 2.83	190
34	37/ 0.02	15/ 0.19	14
not in the lists	232/ 0.14	168/ 2.14	
Total	160,932	7,868	4,152

Tables 2 and 3 show that the phrases themselves consisted of 2,628 word families and after the example sentences were written, there were only 1,524 word families added by the

example sentences.

Table 4 below shows the percentage of items in the top 3,000 word families of English that were not covered by any of the words in the example sentences.

Table 4

Vocabprofiler breakdown of top 3,000 word family words not covered by example sentences created using native speaker intuition.

Word Family Frequency Level	Top 3,000 word family tokens not present in example sentences	Percentage of word amily not covered
K-1 families not in input:	13	1.3%
K-2 families not in input:	85	8.5%
K-3 families not in input:	203	20.3%
Totals	301	10%

DISCUSSION

The results of this study showed that native speaker intuition can be relied upon to create content using mostly high-frequency vocabulary since overwhelmingly the large amount of context created by native speakers fell into the high-frequency realm. In fact, in comparison to the percentage of items that fell into the high-frequency realm for the formulaic phrases alone, the addition of approximately 130,000 more tokens of example sentence context actually only reduced the percentage of tokens in the high-frequency realm by 0.84 percent (see token percentages for 1,000 word family groups 1-3 in Tables 2 and 3). This copious amount of high-frequency data creation revealed that native speaker intuition can be relied upon to supply contextual content when the goal is to create supporting context that does not add an addition learning burden in relation to the target formulaic sequence.

This study also confirms the value of a small but extremely frequent amount of word families. In total, the words used in the entire corpus of example sentences consisted of only

4,152 word families. This indicates that even when there is a great amount of data, certain high-frequency words are used repeatedly. Thus the value of high-frequency vocabulary and the collocations they occur with are confirmed. Furthermore, despite adding such a copious amount of context, only 1,524 word families were actually added since the phrases themselves consisted of 2,628 word families. Although a very large database, the vocabulary load (4,152 families) is feasible for learners.

One interesting aspect of this study was the style that the sentences were written in. All four native speakers wrote and used language in a subtly different style. For instance, one of the native speakers, an avid reader of fiction, more often included sentences which included quotes of what someone said in a way that is typical of fiction writing. Another more often wrote about economic issues in comparison to the other writers. Another writer, an American, created sentences involving gun violence more often that the others. It is certainly a possibility that this variety of native speakers writing sentences may have contributed to the high coverage of the top 3,000 word families of English.

Although the example sentences did cover a high percentage (90 percent, see Table 4) of the top 3,000 word families of English, why 10 percent was overlooked should be discussed as well. Ideally, writers would have included some of the words in this 10 percent in the sentences to expose learners to them. However corpora, by its nature, can never truly represent natural language perfectly. For instance, the ease with which corpora can be compiled with written texts already in digital form increases the potential for formal language to more often be included due to the nature of written texts. This is clear in how words such as *bacterium* exist within the top 3,000 words of English. Actually, the existence of the word *bacterium* in the top 3,000 word families of English is an issue, because such a word clearly has low value to learners of general English. Also, since *Vocabprofiler* utilizes word family lists partially derived from the British National Corpus, differences between British and North American English occasionally explained why these words were overlooked. A few examples found were *centimetre, flavour, duke, lord,* and *pub.* Furthermore, the vast majority of the words not found in the top-34 (1,000 headword) word family lists were items that the program has trouble counting, such as word with hyphens (*middle-aged, x-ray,* etc.). Such items highlight weaknesses in the corpus or the software rather than weakness in the example sentences.

CONCLUSION

This study aimed to determine whether native speaker intuition could be relied upon to create contextual content that mostly fell into what is considered high-frequency vocabulary. Native speakers wrote over 160,000 tokens worth of example sentences for high-frequency formulaic sequences derived from a corpus. The resulting database was compared to the formulaic sequences alone to determine whether the content added by the native speakers mostly stayed within the high-frequency realm.

The results showed that the tokens in the sentences not only covered the vast majority of the top 3,000 word families of English (90 percent of them), 97.39 percent of the words in the sentences also fell into these top 3,000 families. Therefore, this study affirmed that native speaker intuition can be relied upon for such a task, even on a large scale.

While this study highlighted how the intuition of experienced ESL practitioners can be relied upon to produce high-frequency contextual content, some unintended discoveries were also made. The content all four native speakers created had subtle differences in style and focus, and this variety of language may have contributed to the high coverage of high-frequency vocabulary. Therefore, future research should consider this and compare the type of language created by multiple native speakers versus only one to determine whether the subtle differences among writer styles are connected to high-frequency vocabulary coverage.

REFERENCES

Anthony, L. (2011). *AntConc* (*Version 3.2.2*) [Computer Software]. Tokyo, Japan: Waseda University, Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2013). *AntWordPairs* (*Version 1.0.2*) [Computer Software]. Tokyo, Japan: Waseda University, Available on request.

Bauer, L. & Nation, P. (1993). Word Families. *International journal of lexicography*, 6(4). Oxford: Oxford University Press.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23,* 321-343.

Cheng, W., Greaves, C. & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11,* 411-433.

Cobb, T. (2013). *Vocabprofiler*. Retrieved from http://www.lextutor.ca/vp/

Cowie, A. (Ed.) (1998). *Phraseology: theory, analysis, and applications.* Oxford: Oxford University Press.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 Million Words, 1990-Present.* Retrieved from http://corpus.byu.edu/coca/

Davies, M. (2010). *Word List Plus Collocates.* Retrieved from http://www.wordfrequency.info/purchase1.as p?i=c5a

Ellis, N. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Heatley, A., Nation, P. & Coxhead, A. (2002). RANGE program. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 10-27). Hove, England: Language Teaching Publications.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Nation, P. (2001). How many high frequency words are there in English? In M. Gill, A.W. Johnson, L.M. Koski, R.D. Sell, & B. Warvik (Eds.). *Language, Learning, Literature: Studies Presented to Hakan Ringbom. English Department Publications* (4) (pp. 167-181). Turku, Abo Akademi University.

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques.* Boston, Heinle.

Nation, P. & Meara, P. (2002).Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35-54). London: Edward Arnold.

Rogers, J. (2010). Is intuition enough when choosing vocabulary? 関西外国語大学研究論集, 第91号, 195-210.

Rogers, J., Brizzard, C., Daulton, F., MacLean, I., Florescu, C., Mimura, K., ... Shimada, Y. (in press). On using corpus frequency, dispersion, and chronological data to help identify useful collocations. *Research in Corpus Linguistics.*

Schmitt, N. (2010). *Researching Vocabulary.* New York: Palgrave MacMillan.

Someya, Y. (1998). *E-lemma list.* Retrieved from http://www.antlab.sci.waseda.ac.jp/software/resources/e_lemma.zip

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

（James M. Rogers　国際言語学部講師）
（Frank E. Daulton　龍谷大学教授）
（Ian B. MacLean　英語国際学部講師）
（Gordon A. Reid　国際言語学部准教授）