An Objective Method of Identifying Teachworthy Multi-word Units for Second Language Learners

James Rogers

1 Meijo University, Nagoya Aichi 463-0012, Japan
jrogers@meijo-u.ac.jp

Abstract. This study aimed to confirm whether a frequency-based approach toward the identification of multi-word units (MWUs) most exemplary of lemma that frequently co-occur would be as reliable as one that relies upon native speaker intuition. This results showed that the frequency-based method produced very similar results to the method that relied upon native-speaker intuition, and thus could be considered more reliable in that potential subjectivity could be avoided.

Keywords: Multi-word unit, concgram, corpora

1 Introduction

Research has shown that MWU knowledge for second language learners improves language processing speed. However learners are often very weak in this aspect of fluency and learning materials do not focus on them. This stems from a lack of resources. In fact, high-frequency collocation and MWU identification is a very complex, time-consuming process in which much research is still needed. This current study will evaluate a method which aims to improve upon the objectivity of one of the steps taken in previous research to identify high-frequency MWUs.

2 Literature Review

Knowledge of MWUs has been found to aid in more efficient language processing (Conklin & Schmitt, 2008). While there is variation in simply defining MWUs, this current study will define them as the MWUs most exemplary of lemmatized concgrams, with a ‘concgram’ being “all the permutations of constituency and positional variation generated by the association of two or more words” (Cheng, Greaves, & Warren, 2006, p. 411).

Despite agreement on the value of MWU knowledge, many studies have shown that learners are not acquiring this fluency (DeCock et al., 1998; Nesselhauf, 2005). This lack of fluency is connected to the fact that learning materials simply do not focus on them (Gitsaki, 1996). This is due to a severe lack of large-scale studies which actually have identified MWUs. Identifying MWUs is a complex and time consuming process, and thus it is difficult to create large-scale comprehensive mater-
als (Rogers, 2017). Thus, much research is still needed simply in developing and evaluating efficient and reliable MWU identification methods.

In Rogers (2017), MWUs were identified by searching for exemplary chunks of high-frequency words and their collocates. Mini corpora were compiled for these, and were then analyzed to extract the most frequent MWU in which the high-frequency pivot word and collocate occurred. To identify this MWU, native speakers chose the most frequent chunk identified. However, if this core item occurred within any subsequent items and the native speaker judged it to be worthy of teaching, the native speaker could opt to extend the core and have that MWU be the exemplar for the lemmatized concgram instead. For instance, if *come to terms* was identified as the most frequent chunk for the lemma *come/terms*, and the next most common chunk was *come to terms with* and the native speaker felt that this extended version was worth teaching, then it was chosen to represent *come/term*. Extending core MWUs in this way was found to be an essential step in that 53 percent of the sampled items in Rogers’ (2017) study were extended upon. However, relying upon native speaker intuition can introduce subjectivity. Thus, a gap in the research exists as to whether or not an objective method can be relied upon for this step.

3 Research Question

Does an objective method of extending exemplary MWUs of lemmatized concgrams beyond their core produce results similar to those using native speaker intuition, and does it affect native speaker intuition-based judgments of teach-worthiness?

4 Procedure

This study evaluated an objective method to extend MWUs of lemmatized concgrams beyond their core in comparison to Rogers’ (2017) method which utilized native speaker intuition. It utilized corpus frequency data instead of native speaker intuition to decide whether or not MWUs should be extended or not. These results were compared to the results of Rogers (2017) to determine whether a similar percentage of items were extended or not. The results were examined to determine whether or not extending a MWU beyond its core affects native speaker intuition based judgments of teach-worthiness.

First, the most frequent 500 lemma in Gardner and Davies’ (2014) high-frequency academic vocabulary list were utilized as pivot words to search for lemmatized collocates in the academic section of the Corpus of Contemporary American English (COCA) (Davies, 2008). Mini corpora were then compiled for the lemmatized pivot words and collocates from the academic section of the COCA, and these corpora were analyzed to extract the most frequent MWU in which the pivot word and collocate occur.

Analyzing the data using conventional concordance software such as AncConc (Anthony, 2011) would result in a large amount of noise in each set that would have to be
removed manually. Rogers (2017) solved this issue by utilizing AntWordPairs (Antony, 2013), a custom piece of software designed specifically for his study. Thus, this same software was utilized in this current study.

This current research is part of a larger study which resulted in approximately 9,000 MWUs. Analyzing such a large amount of data is beyond the scope of what will be explored, and thus a random sample of that data was examined in this study. Data for 100 random pivot words of the initial 500 were utilized.

Collocation implies ‘frequent’ co-occurrence. This study thus followed Rogers’ (2017) parameter of approximately one occurrence per million tokens. However, since this current study is only working with the academic section of the COCA which is one-fifth of the entire corpus, this study’s frequency cut-off was set to 100 occurrences or more.

Other researchers have also used relative frequency of co-occurrence compared with the individual words’ total independent occurrences to help identify collocations. Church and Hanks (1990) referred to this measure as ‘mutual information’ (M.I.). Stubbs (1995) and Hunston (2002) both believe that an M.I. statistic of three or higher indicates that two words collocate. However, after analyzing initial data, this cutoff was deemed to be too exclusive, leaving out many clearly useful collocations. M.I. cutoffs of two and one were thus experimented with. An M.I. cutoff of one was deemed too inclusive, while a cutoff of two was found to be most balanced and thus it was used.

After processing with AntWordPairs, results were examined to determine if the core MWU should be extended or not. Any subsequent MWU that still contained the core while having half or more of the frequency of the core was considered the top MWU. Then, the data was further examined to determine if that new MWU identified could be extended any further. Its frequency was used to determine if any other subsequent MWUs contained it and also had half or more of its frequency. This process continued on until it was not possible to continue any longer. MWUs that were extended were then tallied.

Three native speakers then judged the MWUs in regards to their teach-worthiness for learners of academic English. They made a judgment since corpus data can contain noise and does not always produced what the user intends to find. The native speakers were instructed to mark items as not worthy of teaching if they are: odd formulations (noise in the data), proper nouns, too specific to a genre of academic research, have more of a tendency to occur in general English rather than academic English, grammatical formulations that are devoid of meaning, or items that are overly technical.

Then an interrater reliability (IRR) was conducted to assess the degree that coders consistently determined whether the lemmatized congrams were pedagogically valuable for learners of academic English. Cohen’s kappa was utilized to factor out agreement due to chance, and a mean of the kappas for each unique pair of coders was used. Landis and Koch’s (1977) scale of interpreting kappa was used to determine the level of agreement. Since ratings may not be evenly distributed across categories, the formula of IRR=2(PA)-1 (where PA is the probability of agreement due to chance) was used as recommended by Siegal and Castellan (1988). Byrt, Bishop and Carlin (1993) recommend reporting all three kappas when there may be doubt, and thus this was followed.
Finally, the judgments were tallied for items that were extended versus items that were not extended to determine whether or not extending MWUs beyond their core affects native speakers' judgments of teach-worthiness.

5 Results

The 100 pivot words analyzed in this study produced a total of 1,777 concgrams. When the MWUs of these were identified, native speakers deemed 79 percent to be worthy of teaching. 48 percent of these items examined were extended. 41 percent of items judged as being unworthy of teaching were extended, and out of these all cores of that extended item were also judged to be unworthy of teaching.

IRR analysis showed that marginal distributions indicated the presence of both prevalence and, to a lesser extent, bias in the data (Hallgren, 2012). Therefore, following Byrt, Bishop, and Carlin (1993), Cohen’s (1960) kappa was supplemented by two other measures of agreement: Siegel and Castellan’s (1988) kappa (which guards against bias) and Byrt et al.’s 2(PA) – 1 (which guards against prevalence). For each of these indices, IRR was computed separately for the three coder pairs then averaged to provide a single IRR index (Light, 1971). Based upon Landis and Koch's (1977) guidelines, IRR was moderate to substantial, $\kappa = .46$; Siegel and Castellan's $\kappa = .45$; 2(PA) - 1 = .62.

6 Discussion

This study showed potential for improved reliability of utilizing corpus data in comparison to native speaker intuition. In addition, it was revealed that extending an item did not make it more prone to be judged as unworthy of teaching. However, although more objective, the method still needed to be done manually and is time consuming. Future concordance software development certainly should take these issues into consideration.

IRR was shown to be reliable, and the judgements showed that the vast majority of items (78.5 percent) were deemed to be worthy of teaching. However, if only corpus data and the quantitative parameters used in this study were relied upon 21.5 percent of the results would be unworthy of instruction. This large percentage thus points to native speaker intuition still being essential in creating a good quality, ready to use resource.

7 Conclusion

This study found that an objective method utilizing corpus frequency data can be relied upon to produce results similar to what native speakers produce when they use their intuition. Approximately half of the items examined in this study were extended beyond the core top MWU identified using corpus frequency data, which is the nearly the same amount that were extended in a previous study which relied upon native speaker intuition. This study also found that native speaker intuition was shown to be an essential step needed to be taken to remove items deemed not worthy of instruction. While this
method is an improvement since it is more objective, the step still needs to be done manually and is very time-consuming. It is hoped that future concordance software development will take the results of this study into consideration to help improve upon the efficacy and quality of MWU identification.

References

19. Rogers, J. What are the collocational exemplars of high-frequency English vocabulary? On identifying multi-word units most representative of high-frequency lemmatized concgrams