Corpus-informed test development: Making it about more than word frequency

Jonathan W. Trace[1] and Gerriet Janssen[1, 2]
jtrace@hawaii.edu
gjanssen@hawaii.edu

1. University of Hawai‘i at Mānoa
2. Universidad de los Andes–Colombia

Abstract

Given the rising popularity and usefulness of corpora in the field of applied linguistics, more and more there is a need to identify practical applications of the different tools available beyond just word frequency. One area where corpora seem ideal for this is in the realm of second language assessment. This study looks at the use of corpus-informed test items on an academic English vocabulary test (N = 203). Two different formats of the test (c-test and multiple-choice) are analyzed to explore possible relationships between item characteristics for difficulty and contextual information. First, Rasch measurement is used to determine the difficulty of a set of common items across both tests. These results are then compared with a series of mutual information scores based on collocations and multi-word constructions with the target items. The goal is to examine possible relationships between context and item difficulty, and more importantly provide teachers and test-designers with one way to utilize corpus linguistics to create more effective language assessment tools.

Keywords: corpus linguistics, language testing, formulaic language, vocabulary

Introduction

Corpus-based research is still a growing area in applied linguistics, though it boasts a long and productive almost 40 year history, with studies ranging from basic descriptions of language (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Sinclair, 1990), to more practical applications such as lexico-grammatical approaches to instruction (Liu & Jiang, 2009) and materials development (Chang & Kuo, 2011). One area that is slowly building momentum is the use of corpora in language assessment design and use (see Barker, 2005; Coniam, 1997; Sharpling, 2010). These two areas would seem to go hand in hand with one another given that both espouse such concepts as reliability and authenticity. Yet it still seems that most of the testing literature that incorporates corpus is still limited to conversations that seldom go beyond word frequency (Crossley, Salsbury, McNamara, & Jarvis, 2010).

Certainly there is more value to be had from these vast databases of authentic language than just how often a word is used? As teachers and test designers, we know there is more to knowing a language than just the individual parts on their own. Rather, it is language in use, with considerations of context or lexico-grammatical function that we should be interested in measuring and teaching. Even when our focus is on something narrow, such as the vocabulary test in this study, there are still several distinct constructs that stand out as important for successful mastery of a language that go beyond frequency, such as vocabulary depth (Nation & Snowling, 1997). Neither are we always interested in knowledge of words on their own, as language is not given to us in piecemeal but as part of a larger whole. This includes knowledge of formulaic language (e.g., *n*-grams or idiomatic expressions), which have been common topics of discussion in the field (Evert, 2009; O'Keeffe, McCarthy, & Carter, 2007) but remain relatively unexplored in the field of language assessment.

The goal of this study is to highlight one possible way of expanding our use of corpora in the field of language assessment from a very practical and authentic position. Using a vocabulary test, this study will explore one method of analyzing words in context, as well the outcomes of a corpus-informed test in the hopes of providing teachers and test designers tools to better measure language. To this end, the following research questions were asked:

1. What if any are the relationships between item difficulty and mutual information as identified by corpus-derived data?
2. To what degree are these relationships similar across different test formats?

Methods

Participants

Data were collected from responses on a vocabulary test from a total of 203 examinees at a South American University in Colombia. The test is part of a larger placement exam for incoming Ph.D. students in an English for Academic Purposes support program. Examinees were primarily L1 speakers of Spanish and ranged from low beginners to advanced users of English.

Instruments

The assessment tool discussed in this study was a test of academic English vocabulary. Along with writing and speaking subtests, the vocabulary test was part of a new reading pilot that included sections for grammar and reading comprehension. Data for this study were collected from three administrations of the vocabulary subtest.

As this test is still in a piloting stage, revisions to the test are ongoing and variations exist across each of the three administrations examined here. Most apparent among these is that the first two administrations (n = 124) used a c-test (CT) format, where examinees were given a passage with several missing words that they are required to supply. Unlike a traditional cloze procedure, in a CT the first letter of each missing word is provided both as a clue for test takers, as well as to limit the possible number of accepted responses. For the final administration (n = 79), the test was converted to a multiple-choice format (MCT) based on apparent difficulty problems with earlier versions of the test.

The original design of the test incorporated a 500-word passage with 26 missing words using a rational pattern of deletion. Items were selected based on corpus data from the *Corpus of Contemporary American English* (COCA; Davies, 2008) using statistics such as word frequency and mutual information scores. As the researchers wanted to choose a topic that was academic but also general, to avoid biasing any particular academic background or major, an article on the history of the wheel was selected from an online academic journal.

After the first administration of the test, item analysis was conducted using classical test theory. Six items were removed from the test as problematic and two new items were introduced (k = 22). Results of the revised test were also analyzed, and based on these findings the researchers decided to change to a MCT format. Examinees were presented with four possible choices in-text and were required to circle the correct answer rather than producing any language. The new test removed ten of the previous items and added eight new items (k = 19). In total, twelve items were shared across all three tests, and these were used as the basis for the final linguistic analysis detailed below.

Procedure

Because of our interest in the relationships between an item's target word and the immediate context surrounding that word, mutual information (MI) scores were used as reported by the COCA. MI is a statistical measure of association that indicates the degree to which a set of words or a phrase is likely to appear in the same pattern together in the language (Biber, 2009; Evert, 2009). It works by comparing the frequency of a multi-word pairing or phrase. A high MI value is found when there is a strong likelihood of words or a phrase to appear together within the corpus. According to Davies (2008), MI values of 3.00 or higher indicate a high chance of a set of words being bound together semantically in naturally occurring language. Scores are dependent upon the individual word frequencies, as those words with very high frequencies show up in many different contexts and their appearance with other words may be more due to random chance than any kind of formulaic-ness.

In order to measure collocations between the target word and nearby function words, MI scores were gathered for all function words within a fixed area around the target word. While MI is typically used to look at fixed semantic phrases (e.g., *strong coffee*), taking into account only immediate pairings of words, it seems logical that words that are still local (e.g., within the same *T*-unit) but not immediate might also have a triggering effect for a particular target word, and this has been explored in the psycholinguistics literature (e.g., Duffy, Henderson, & Morris, 1989). To reflect this, analysis of MI scores included all function words within four collocations to the left and right of the target word within the same *T*-unit. In order to make comparisons about the relationship of collocation on item difficulty, the maximum left and right MI scores were used in the analysis.

Biber (2009) differentiates between collocations of function words and multi-word formulaic sequences. Given our interest in how different kinds of context influences item difficulty, it is important to consider MI values for both of these kinds of constructions. As with collocations, multi-word formulaic sequences required a bit a preparation to measure in a systematic and authentic way. Formulaic sequences can be quite varied, both in relation to the content words within a fixed set of function words (Renouf & Sinclair, 1991), as well as in length (e.g., *fairly certain* vs. *fairly certain that*). As the target items were typically function words, and both the CT and MCT formats limited the number of possible answers, the main concern was determining what is or isn't part of a phrase. One possible way of accomplishing this is to look at multi-word constructions in three areas: (a) before the target; (b) after the target; and (c) including the target. By isolating these three patterns, we can check MI values for constructions with the target word as the base, and then expand outwards in the appropriate direction. Different number *n*-grams can be compared in terms of their MI scores, and the construction with the highest MI score and fewest number of words can be reasonably identified as a multi-word formulaic pattern in the data.

Analysis

Exact answer scoring was used in the analysis of the item data as a way of controlling for differences in responses by examinees. While more than one answer was possible for some of the items, exact answer scoring allowed us to focus only on the original constructions in their relationship to item difficulty and corpus linguistic features. Given the different formats and modes (e.g., receptive vs. productive) of the CT and MCT, analyses were conducted separately. For the CT analysis, only those items that were shared across both tests were included (k = 20). As there was only one version of the MCT, all 19 items were included in the analysis.

Rasch measurement was utilized to analyze item responses on both tests using *Winsteps* (Linacre, 2010). Unlike classical test theory, Rasch, which belongs to the item-response theory family of test analysis, can give a sample-free estimation of item difficulty as it relates to examinee ability levels along a true interval scale. The benefit of Rasch modeling has been discussed extensively in the literature (e.g., Henning, 1984; McNamara & Knoch, 2012), but suffice to say this method of analysis provides a more generalizable and readily comparable interpretation of item difficulty.

Corpus analysis of the 12 common items on the test was carried out according to the procedure outlined above, with five sets of MI scores for each item: (a) left MI; (b) right MI; (c) pre *n*-gram MI; (d) post *n*-

gram MI; and (e) mid *n*-gram MI. In addition, the frequency of the target word per one million words was recorded. While the COCA provides both spoken and written data, as these were items on a vocabulary test and measuring knowledge of the written language, only written corpus data was used for all analyses.

Results

Descriptive data for both test formats are displayed in Table 1, including means, standard deviations, minimum and maximum values for each the CT results and the MCT results. Notice that the mean score on the CT was markedly low (*M* = 3.90) with a relatively low degree of variation in scores (*SD* = 3.70), indicating that the test was quite difficult for the sample of examinees. By comparison, the MCT was more normally distributed (*M* = 12.90, *SD* = 3.86), with examinees appearing to perform much higher than on the CT. Cronbach's alpha reliability estimates are also included in the bottom row of Table 1, indicating the degree to which the scores on the test were internally consistent. Scores on the CT were more reliable, with an estimate of 88%, while the MCT was slightly less reliable at 73%. We should be careful in over-interpreting the reliability of the CT given the low degree of variance of scores and positively skewed distribution. These might be causing this value to be higher than it actually is, as reliability estimates work under the assumption of normally distributed data. A lack of variance might mean that the scores are consistent, but only consistently low, and have little to do with actual test function.

Table 1
Descriptive Statistics for the Vocabulary C-Test and Multiple-Choice Test

	CT	MCT
M	3.90	12.90
SD	3.70	3.86
Min	0.00	3.00
Max	14.00	19.00
N	124	79
k	20	19
α	.88	.73

Item analysis of the tests was performed using Rasch analysis, which displays item difficulty as logit measures. Measures are spread across an interval scale with a mean value of 0.00, ranging negative (less difficult) to positive (more difficult). A first step to determining item function in Rasch is to check the fit of the items, or the degree to which the item measures can be adequately predicted by the model. Misfitting items were determined by evaluating infit mean square values and identifying items more than two standard deviations from the mean (McNamara, 1996). A preliminary analysis found that one of the common items misfit the model for the CT, and so was removed from the final analysis for a new total of 11 items.

Table 2 displays results for the 11 remaining common items in the order they appear on the tests. The first column displays the target word, followed by item difficulty in logits. Corpus statistics are also included for each item, including the frequency of the target word per one million words in the COCA, the highest MI value for near context words to the left and right of the target word, and *n*-grams with the target for left, right, and surrounding context.

Looking first at the item analysis data for the 11 common items, we can see that there are clear differences between difficulty measures for both tests. Notice that in general measures for the CT were higher than those for the MCT, which again points to the CT being the more difficult test. While the CT had four items with logit measures above 2.00, the most difficult item on the MCT was for the target word *people*

(1.96). Most items on the MCT were either closer to the center of the scale or quite easy as reflected by high negative values.

Table 2
Item and Corpus Statistics for 11 Vocabulary Items

Item	CT Measure	MCT Measure	Word Freq*	Left MI	Right MI	Pre *n*-gram MI	Post *n*-gram MI	Mid *n*-gram MI
Exactly	-1.88	-2.28	74.94	5.97	4.19	7.29	8.06	10.12
Certain	2.38	-0.87	130.01	6.61	2.17	5.47	0.00	5.88
Beneath	3.37	0.22	47.55	3.27	2.11	2.77	0.00	5.18
Learned	-0.86	-0.46	93.45	4.42	3.65	1.09	6.56	8.88
People	-0.06	1.96	1006.95	1.47	0.00	0.00	0.00	0.00
Before	-3.55	-2.05	668.49	4.13	3.94	3.33	6.46	8.21
Indicate	0.31	0.67	36.23	4.60	5.39	0.00	0.00	10.27
Making	-0.20	-0.87	227.06	1.70	3.36	3.76	7.09	0.00
Although	2.38	-0.19	239.89	0.00	1.70	0.00	0.00	0.00
Keeping	2.78	1.24	55.08	2.38	3.30	3.17	2.96	7.69
Everyday	0.62	-0.10	18.84	3.87	8.33	7.86	7.81	10.22

Note. * Word frequency based on the occurrence of the target word per 1 million words in the COCA.

For the corpus data, we can see that the frequency of the target words was quite varied, ranging from about 18-1000 occurrences per million words, though most items had values below 250. MI scores for left and right collocations showed that most items had at least one semantically related word in the near context. Recall that MI scores of 3.00 or higher indicate a semantic relationship, and only *although* (1.70) and *people* (1.47) were below this threshold. As the former is a connecting device and not likely to be directly contextually linked, and the latter was the most frequent word, these results could be expected. We find similar patterns in the data for multi-word MI scores. Overall, the highest multi-word MI scores occurred when the target word was centered in the phrase.

Pearson product correlations were used to gauge the degree of relation between corpus data and item difficulty. Table 3 displays these results arranged by item difficulty, frequency, collocation, and multi-word formulaic sequences. Given the number of comparisons, an *a priori* alpha of $p < .002$ was used in determining statistical significance. As we might expect given the low number of items in our sample, none of the correlations were determined to be statistically significant, though there were some interesting trends worth exploring in the data.

Table 3
Correlation Matrix for Item Difficulty and Corpus Statistics for 11 Vocabulary Items

	CT Measure	MCT Measure	Word Freq*	Left MI	Right MI	Pre *n*-gram MI	Post *n*-gram MI	Mid *n*-gram MI
CT Measure	1.00	.56	-.41	-.27	-.25	-.13	-.66	-.28
MCT Measure		1.00	.19	-.50	-.28	-.54	-.61	-.30

Note. * Word frequency based on the occurrence of the target word per 1 million words in the COCA. A Bonferroni adjusted a priori alpha value of p < .002 was set to account for the number of comparisons.

Notice that items in the CT displayed the strongest relationship with multi-word sequences following the target word (r = -.66). A negative value indicates that as item difficulty on the CT increased, the likelihood that the target word was the beginning of a fixed phrase decreased. The same was true for items in the MCT (r = -.61), but expanded also to *n*-grams that preceded the target item (r = -.54). MCT difficulty also appeared related to collocations occurring before the target word (r = -.50). This seems to show a possible relationship between item difficulty and the presence of fixed word combinations or multi-word sequences, and that the pattern of this influence might change depending on the test format.

Discussion

Based on the proposed corpus-driven methodology described above, results seem to indicate tentatively that there is a relationship between item difficulty and the degree to which the item is connected to nearby context in the form of fixed expressions. Based on the correlation data, it appears that different contextual features are affecting difficulty as a whole across both tests. Difficulty seems to be influenced when the target is part of a multi-word phrase, either with a fixed sequence of words preceding or following the word. Alternatively, the degree to which single word collocations are related to item difficulty is not as clearly displayed, especially for the items in the CT. Word frequency was only weakly related to item difficulty, and only for the CT. This is likely due to the small sample of items in the analysis, but still worth nothing that collocations were more related than word frequency alone, as we might expect in a test of vocabulary in context (Crossley et al., 2010; Zareva, 2007).

Findings related to the CT and MCT also showed some apparent differences in formulaic language and item difficulty across test formats. As mentioned, the difficulty of the items on the CT seemed to be mostly unrelated to the presence or absence of collocations to the target. There was evidence of a possible relationship, however, between items on the MCT and the presence of collocations prior to the target. Items on both tests were sensitive to the presence of multi-word phrases, though again differences were found between test formats. While items on the CT seemed to be affected by sequences following the target, the MCT included sequences before and after the target. Neither was influenced when the target word was centered in a sequence, which was somewhat surprising as those values tended to be the highest and most common in written English (Biber, 2009).

It could be these differences were due in part to the presence of options in the MCT. Examinees might have been able to use the provided answer choices to help interpret the context, whereas in the CT examinees didn't have access to this added information and had to work from the context alone. We might think this would increase the effect of multi-word sequences on item difficulty on the CT, but the data doesn't seem to support this notion. This might be a result of the CT being too hard, or that examinees lacked knowledge about the context to make these kinds of judgments without more information. Unfortunately, without more information or better functioning items, it is impossible to be certain.

Conclusions

The goal of this study was to display one possible practical application of corpus-based test design through the use of different statistical procedures. While there remain a variety of questions about how corpus-informed tests function in different contexts, we hope that this can be a starting point for test designers to make more informed decisions when creating and selecting items.

This was a small-scale study with only a few items, and because of this we must be careful with the kinds of conclusions that can be drawn. That said, the results do seem to indicate possible benefits in using collocations to influence item difficulty, and point to the value in looking beyond frequency or individual words when testing vocabulary as authentic language use.

It is hoped that this information can lead to more in-depth studies of the use of corpora in test development. The next step in this research will be to look at a broader range of items that can more fully encompass different constructions of vocabulary in context, as well as incorporate eye-tracking methods to examine where test takers are looking when responding to items in a test, using online measures of processing to better understand the degree to which examinees use context in reading assessment.

References

Barker, F. (2005). *What insights can corpora bring to language testing? CRILE* (pp. 1–4). Lancaster University. Retrieved from http://www.ling.lancs.ac.uk/groups/crile/docs/crile%20lectures/barker0105.doc

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311. doi:10.1075/ijcl.14.3.08bib

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Chang, C.-F., & Kuo, C.-H. (2011). A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, *30*(3), 222–234.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Calico*, *14*(2), 15–34.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition 15*, 791–801.

Evert, S. (2009). Corpora and collocations. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 1–53). Berlin: Mouton de Gruyter.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing*, *1*(2), 123–133.

Linacre, J. M. (2010). *A user's guide to Winsteps*. Chicago, IL: Author.

Liu, D., & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *Modern Language Journal*, *93*(1), 61–78.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. doi:10.1177/0265532211430367

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*, 359–370.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching* (pp. 1–333). Cambridge: Cambridge University Press.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128–143). London: Longman.

Sharpling, G. P. (2010). When BAWE meets WLT: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage. *Journal of Writing Research*, *2*(2), 179–195.

Sinclair, J. (1990). *Collins COBUILD English grammar*. London: Harper.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers’ lexical organization? *Second Language Research*, *23*(2), 123–153.