English Language Teaching; Vol. 9, No. 7; 2016
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

The Role of Native and Learner Corpora in Vocabulary Test Design

Eman Saleh Akeel[1]

[1] English Language Institute (ELI) – King Abdulaziz University (KAU), Jeddah, Saudi Arabia

Correspondence: Eman Akeel, English Language Institute (ELI) – King Abdulaziz University (KAU), Jeddah, Saudi Arabia. E-mail: eakeel@kau.edu.sa

Received: January 10, 2016 Accepted: April 29, 2016 Online Published: May 10, 2016

doi: 10.5539/elt.v9n7p10 URL: http://dx.doi.org/10.5539/elt.v9n7p10

Abstract

The growing field of corpus linguistics has been engaged heavily in language pedagogy during the last two decades. This has encouraged researchers to look for more applications that corpora have on language teaching and learning and led to the emersion of using corpora in language testing. The aim of this article is to provide an overview of using corpus data for the purpose of vocabulary test designing. It presents some native and learner corpora which are available for item writers to use. It covers the benefits and limitations of using corpora in language testing and argues for the importance and usefulness of using native as well as learner corpora as tools for designing a vocabulary test. The article aims to illustrate how both native and learner corpora can be used in language testing in general and in the development of vocabulary tests in particular.

Keywords: corpora, native corpora, learner corpora, vocabulary test, item writer

1. Introduction

The growing field of corpus linguistics has been engaged heavily in language pedagogy during the last two decades (e.g. Aijmer, 2002, Gavioli, 2006; Kettemann & Marko, 2006). Many studies have shown the impact of corpora on English language teaching and learning and have developed corpus-based teaching materials (e.g. Biber, Conrad, & Leech, 2002; Tilbury, Clementson, Hendrs, & Rea, 2015). This has encouraged researchers to look for other applications that corpora have on language teaching and learning and led to the emersion of using corpora in language testing (Alderson, 1996; Barker, 2006; Barker, 2010).

Sinclair defines a corpus as "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language, typically contains many millions of words" (Sinclair, 1991, p. 171). There are many different types of corpora (e.g. general language corpora, specialised corpora, historical corpora, etc.). There are also corpora in so many different languages. In this article, I will focus on two main types: native and learner corpora. The difference between those two types will be explained in section 4. At this point, it is useful to note that the main difference between them is the source of their data. In the first type, the data is collected from native speakers of English, whereas in the latter it is collected from learners of English.

This paper aims at discussing some possible ways of using corpus data for vocabulary test designing. It will also discuss the benefits and limitations of using corpora in language testing. Finally, it argues for the importance and usefulness of using native and learner corpora as tools for designing a vocabulary test. The purpose of the paper is not to develop a corpus-based vocabulary test, but to emphasize the role of corpora in the development of vocabulary tests.

2. Testing the Vocabulary of a Second Language

Vocabulary knowledge has been defined in many ways by different scholars. It has been referred to as the knowledge of word meaning, the collocational knowledge, and the knowledge of constraints in the use of a word (Nation, 2001; Richards, 1976; Elyas & Alfaki, 2014). From the early developments of assessing language proficiency, vocabulary has been one of the most crucial items in language testing. In fact, many learners used to believe that the more words they know the more linguistically competent they are.

In the literature of vocabulary testing, two main approaches are identified: one is concerned with vocabulary size and the other measures the quality of lexical knowledge. The focus in the first approach is the breadth of vocabulary whereas in the second one it is the depth of vocabulary knowledge.

There is a number of vocabulary test approaches proposed by different scholars, some of which view vocabulary as the knowledge of discrete words that are independent of context (Nation, 1990). However, Bachman (1990) proposes that language proficiency is a set of communicative skills, and therefore, measuring vocabulary knowledge should not only test the knowledge of separated items, but to include communicative competence as well.

There are several formats in which a vocabulary test can be constructed such as matching the words with their meaning, gap fill tasks (or cloze) where candidates are required either to write the correct word (open cloze) or choose it (multiple choice cloze) and much more. Henning (1991) suggested that a multiple choice cloze test is the best way to assess vocabulary knowledge in the Test of English as a Foreign Language (TOEFL) vocabulary items.

One of the earliest lexical measures is the English as a Second Language (ESL) Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey, 1981), where the vocabulary is part of other analytic scales that are used to evaluate the candidate's composition. Measuring vocabulary here is embedded within a larger construct. A contrasting way of measuring lexical knowledge is the Nation Vocabulary Levels Test (Nation, 1990) which provides a frequency profile for the candidates' vocabulary. Vocabulary here is assessed as a discrete and an independent construct (Chapelle, 2001). In such a test, candidates simply have to match words with their synonyms or definitions. The test takers are asked to choose the correct meaning that goes with the given words. Different versions of the Vocabulary Levels Test were developed to provide evidence of the validity of this kind of tests (Schmitt, Schmitt, & Clapham, 2001).

There is also The Lexical Frequency Profile (Laufer and Nation, 1995) where test takers compose a written text based on a prompt. The scoring in this type of tests is based on the correct use of word forms.

Another type of vocabulary test is the multiple-choice cloze test (Hale, Gordon, Stansfield, Charles, Rock, Donald et al., 1989 as cited in Read & Chapelle, 2001). This is a selective test where the measure of vocabulary knowledge is focused on specific vocabulary which forms the basis of a multiple choice item. In some other vocabulary tests, half of the words are deleted and the candidates should complete the missing half. Such type of tests is called the C-test (Singleton & Little, 1991 as cited in Read & Chapelle, 2001).

Nevertheless, no one can ignore the importance of context in vocabulary knowledge, and the role of context has become a major aspect in vocabulary testing. Therefore, it became highly significant to test candidates' ability to elicit lexical items from a broad and rich context rather than a narrow semantic level (Read, 1997). So, instead of having a sentence in which the target word takes place, a test should require the candidate to work out the item in a rich discourse context.

3. The Contribution of Corpus Linguistics to Language Testing

Using corpora to inform language testing has begun when examination boards have started compiling electronic collections of candidates' data. In the early 1990s, the English as a Foreign Language (EFL) Division of the University of Cambridge Local Examinations Syndicate and Cambridge University Press started to collect English exam scripts while keeping record of candidates' information such as first language, score, gender and so on (Barker, 2010).

One of the earliest investigations of the use of corpora in the area of language testing is Alderson's study (1996). He made attempts to find out some possible ways in which corpora provide potential contribution to language testing. His suggestions include: (i) test construction, compilation and selection, where he proposes to use the frequency function as a tool for selecting lexical items from the corpus and include them in exams, (ii) test presentation, to present concordance lines from a corpus to learners and ask them to make judgments about the language use such as identifying the genre of a given text, (iii) test scoring, by using authentic language in corpora as a model to compare candidates' responses in a given test, (iv) delivery of results, to use corpus data as a reference when norming the test.

Corpus-based studies have contributed to language testing in several ways. They have been used to generate tests or to reformulate them or even to establish descriptors of a test. Coniam's study (1997) is one of the earliest applications of corpora in test design. In his investigation, he used high frequent words from corpora to automatically produce multiple choice vocabulary cloze tests. In 2007, Kennedy and Thorp studied the common features and errors that are produced by IELTS writing test takers. Using IELTS exam response scripts, they have identified key linguistic features of L2 writing performance. Based on the results of their research, the band descriptors of IELTS writing have been modified (Barker, 2010). Hasselgren (2002) provided evidence of linguistic and mechanical markers of fluency based on a corpus analysis of learner and native spoken language. Her study could influence establishing fluency descriptors in the assessment of speaking ability.

4. What are Native and Learner Corpora?

4.1 Native Corpora

A native corpus is a collection of texts, whether written or spoken, that are produced by English native speakers in natural settings.

There are plenty of native corpora available for use and search. The following examples include texts that are written by English native speakers: the Lancaster Oslo-Bergen corpus of British English, The Brown corpus of written American English, the Australian corpus of English, and the Wellington corpus of written New Zealand English. The size of all these corpora is approximately one million words. They all involve 500 written texts of different genres; each text is about 2,000 words (Lee, 2010). They all can be used when searching for samples of English natives' written language.

However, David Lee (2010) stated that the most widely used and researched native corpus is the British National Corpus (BNC). It contains 100 million words; ninety million words are the written component and ten million words is transcribed speech. The written part of the BNC was mainly drawn from published sources, but it also includes a small number of unpublished texts. It involves a huge variety of text types which makes it a valuable tool for research (Lee, 2010).

4.2 Learner Corpora

A learner corpus consists of spoken or written texts that are produced by English language learners. Learner corpora are usually designed for a specific purpose (e.g. to archive test takers responses), but they can be used for many reasons. In the case of this article, it will provide some possible ways of using learner corpora in developing language testing materials and vocabulary test items in particular.

A major learner corpus is the International Corpus of Learner English (ICLE) which was developed at the Centre for English Corpus Linguistics. It contains 3.7 million words of English learners writing, consisting of argumentative essays written by graduate and undergraduate learners of English from 16 different language backgrounds. This corpus involves types of writing that can represent the general language ability of university level English students. Another learner corpus which is specialised in young learners of English is the International Corpus of Cross linguistic Interlanguage (ICCI). It contains samples of primary school up to pre-university learners' writings from different language backgrounds. Another learner corpus is the Varieties of English for Specific Purposes Database. It consists of texts that are written for ESP by various L1 backgrounds (Lee, 2010).

However, the world's largest and most heavily used learner corpus is the Cambridge Learner Corpus (CLC). This corpus was mainly developed for the purpose of informing test development for Cambridge ESOL and Cambridge University Press. It includes 40 million words, consisting of thousands of exam scripts written by students who take Cambridge English exams. It covers a wide range of proficiency levels and includes information about the candidate's first language, age, gender and date of exam. All this makes the CLC an invaluable source for the development of language testing.

5. Implications

After highlighting the differences between native and learner corpora, this section will focus on how to utilize the data in such corpora in the development of vocabulary test items.

5.1 The Role of Native Corpora in Vocabulary Test Design

Native corpora can be used for the development of vocabulary tests in various ways. One of the possible ways is using the frequency tool. Barker (2006) points out that "50 citations for a selected word supply a sufficient number of items in context which can usually confirm whether the item will provide a fair target for testing purposes" (p.3)". This tool could be used for multiple-choice cloze tests. An item writer can check the number of occurrences of a certain word that he/she wants to include as a test item. Based on the frequency of that word in the corpus, a decision can be made whether or not to include it in the test. For instance, in the BNC, the frequency of the word *hurt* is 4263, whereas as *afflict* is only 54. Both words have the same meaning, but why would a test writer include a low frequency word that is rarely used by native speakers where it could be replaced by another one which is commonly used by them? Certainly, word frequency is not the only factor to be taken into consideration when designing a vocabulary test, but it does play a significant role especially with low levels where low frequency words are usually not appropriate for the learners' level or are not suitable for their learning needs.

Another way that native corpora can be used for designing vocabulary tests is to use the concordance function.

Concordances are useful in many ways. For example, they can provide information about the collocational patterns of a target word. Recurrent collocates provide evidence of fixed expressions (e.g. of course, out of order, etc.). Crawford and Csomay (2016) investigated the collocations of two synonymous words: *equal* and *identical*, and found out that *equal* is more likely followed by abstract words such as *opportunities*, *rights* and *access*, whereas *identical* is followed by concrete nouns like *twins*, *copies*, and *items*. Such corpus-based findings are very useful information for item writers as they can look for common collocations of the word he/she wants to include in a test and use them as items in multiple-choice questions, for instance. Concordance lines can also provide authentic contexts for a given item. By reading samples of native speakers' use of a certain lexical item, the test writer can develop a real life context for the word instead of inventing examples which may sound artificial.

Moreover, native corpora can be used to identify pattern differences between nearly synonymous words. Taking the words *lethal* and *deadly* as an example, they are both defined by Oxford Dictionary as 'able to cause death'. Whereas *lethal* is used to convey the same meaning by the dictionary, the most frequent meaning of *deadly*, according to the BNC, is *very* (e.g. *deadly serious* means *very serious*). By studying the variation between nearly synonyms, item writers can use this information to create distractors for multiple choice questions in a vocabulary test.

5.2 The Role of Learner Corpora in Vocabulary Test Design

Learner corpora are also relevant and useful for vocabulary test design because they provide insights on "the needs of specific learner populations" (Meunier, 2002, p. 125). In addition, learner corpora help teachers and test item writers decide whether a particular grammatical structure or a collocation is difficult or not for language learners (Granger, 2002, p. 22).

Unlike native corpora which are mainly used in the test designing process itself, learner corpora are used in other stages of a test. As Barker (2010) suggests, learner corpora play a major role in defining user needs, in designing tests, and in rating tasks.

In terms of identifying users' needs, a learner corpus provides information about what learners can do at a particular level of proficiency. This is very important because it helps making decisions concerning test specifications and about including items in a test. Before deciding on a vocabulary item in a test, it is useful to check whether or not this particular item is difficult for language learners, who are going to take the test, in the specified level.

In addition, learner corpora provide information about the collocational patterns of learner's written language which can help item writers look for suitable distractors for multiple-choice items.

Moreover, learner corpora can be used in rating or scoring tests. A collection of test takers' responses is helpful for this purpose. A norm referenced test for instance, requires a comparison of candidates' responses to other test takers. Thus, based on the corpus data, a decision can be made regarding the ratings of candidates.

6. Advantages and Limitations of Using Corpora in Language Testing

6.1 Advantages

1). Authentic items as opposed to invented ones: native corpora play an important role in vocabulary testing. With huge collections of native language samples that show how real people use language, item writers will be able to develop genuine items instead of inventing examples which may sound unreal. A corpus such as BNC involve many different registers (e.g. news, academic, fiction, spoken, etc) which makes it suitable for providing real-life language and context for many linguistic domains.

2). The ability to check the frequency of a lexical item: In addition, any corpus has a frequency tool, a function that shows the number of occurrences of any given word. This function is clearly significant for the development of tests especially vocabulary tests. It provides accurate number of words or forms of a word.

3). Register variation: corpora show the number of occurrence of words in different registers (e.g. newspaper, fiction, academic, spoken, written, etc.).

6.2 Limitations

1). It requires some technical skills in using corpora: Generally, using corpora does not require advanced computer skills, but it needs some basic skills in using corpus tools and software. An item writer should be aware of features and basic functions of corpora in order to get the best out of it and to avoid having incorrect or inaccurate results. This could be solved by providing training on how to use corpus linguistic tools for those who will be using them in developing vocabulary tests.

2). Not all corpora are accessible: some corpora are closed where a person has to register in order to gain access. Others are paid corpora, where an amount of money must be paid before providing access.

3). Not all concordance-derived data are grammatically correct: a corpus will never tell us if a sentence is grammatically or syntactically correct because the data in corpora are all naturally occurring, so there might be some errors in the texts.

4). Large amount of data can be challenging and time-consuming: "a corpus is a large collection of texts, and the minimum number of words in a representative corpus is 20,000" (Oostdijk, 1991, p. 50). Therefore, searching a corpus for grammatical patterns, test items or contexts is very much time consuming. After getting concordance lines, one has to investigate and analyse each and every line in order to get the correct interpretation of lexis.

7. Future Applications of Corpora in Language Testing

The use of corpora in language testing has a promising future, and many areas in this field require further investigations. Taylor and Barker (2008) (as cited in Barker, 2010) discussed a range of future applications of corpora on language testing. One of them is the use of corpus tools in the evaluation of essays and candidates' responses using an automated system. With the rapid changes and developments in computer technologies, even spoken data can now be collected much easier than before.

Using corpora is an effective way of researching language. Despite the limitations it may have, using corpus data to inform language testing in general and vocabulary test design in specific is a useful way to develop authentic language and context for test takers.

Acknowledgments

I would like to express my gratitude to the Ministry of Higher Education and to King Abdulaziz University for granting me a scholarship to pursue my postgraduate studies in the United Kingkdom. Special thanks to Dr. Parvaneh Tavakoli for her valuable comments and constructive feedback on the earlier outline of this paper. My sincere appreciation goes to my dear family and friends who supported my throughout my study.

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/lllt.6.07aij

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas, & M. Short (Eds.) *Usingcorpora for language research,* 248-259.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Barker, F. (2006). Corpora and language assessment: trends and prospects. *Research notes 26*, 2-4. Cambridge: UCLES.

Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe, & M. McCarthy (Eds.). *The Routledge handbook of corpus linguistics* (pp. 633-645). London: Routledge. http://dx.doi.org/10.4324/9780203856949.ch45

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English.* Harlow: Pearson Education.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal, 14*/2-4, 15-33.

Elyas, T., & Alfaki, I. (2014). Teaching Vocabulary: The Relationship between Techniques of Teaching and Strategies of Learning New Vocabulary Items. *English Language Teaching, 7*(10), 40-56. http://dx.doi.org/10.5539/elt.v7n10p40

Crawford, W., & Csomay, E. (2016). *Doing Corpus Linguistics*. New York: Routledge.

Gavioli, L. (2006). *Exploring Corpora for ESP learning.* Amsterdam, The Netherlands: John Binjamins.

Granger, S. (2002). A bird's-eye view of learner corpus research. InS. Granger, J. Hung, & Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam, John Benjamins. http://dx.doi.org/10.1075/lllt.6.04gra

Hasselgren, A. (2002). Learner corpora and language testing: smallwords as markers of learner fluency. In S. Granger, J. Hung, & Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and* *foreign language teaching* (pp. 143-174). Amsterdam, John Benjamins. http://dx.doi.org/10.1075/lllt.6.11has

Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to TOEFL vocabulary test items, *TOEFL research reports, 35,* educational testing service, Princeton, NJ.

Kennedy, C., & Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS academic writing task. In L. Taylor, & P. Falvey (Eds.), *IELTS collected papers: research in speaking and writing assessment (studies in language testing* (vol. 19, pp. 316-377). Cambridge: UCLES and Cambridge University Press.

Kettemann, B., & Marko, G. (Eds.). (2006). *Planning, gluing, and painting corpora. Inside the applied linguist's workshop.* Frankfort. Germany: Peter Lang.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied linguistics, 16,* 307-322. http://dx.doi.org/10.1093/applin/16.3.307

Lee, D. (2010). What corpora are available? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 107-121). London: Routledge. http://dx.doi.org/10.4324/9780203856949.ch9

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, & Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreignlanguage teaching,* (pp. 119-141). Amsterdam, John Benjamins. http://dx.doi.org/10.1075/lllt.6.10meu

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Heinle and Heinle: New York.

Oostdijk, N. (1991). *Corpus linguistics and the automatic analysis of English.* Amsterdam and Atlanta, GA: Rodopi.

Read, J. (1997). Assessing vocabulary in a second language. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education* (vol. 7, pp. 99-107). *language testing and assessment.* Dordrecht: London. http://dx.doi.org/10.1007/978-1-4020-4489-2_10

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment, *Language testing, 18/1,* 1-32. http://dx.doi.org/10.1191/026553201666879851

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10,* 77-89. http://dx.doi.org/10.2307/3585941

Römer, U. (2011). Corpus research applications in second language teaching. *Annual review of applied linguistics, 31,* 205-225. http://dx.doi.org/10.1017/S0267190511000055

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language testing, 18,* 55-88. http://dx.doi.org/10.1191/026553201668475857

Sinclair, J. (1991). *Corpus concordance collocation.* Oxford, UK: Oxford University Press.

Tilbury, A., Clementson, T., Hendra, L. A., & Rea, D. (2015). *English Unlimited.* Cambridge University Press.

Copyrights