Poster Presentation

Harnessing Keyness: Corpus-based Approach to ESP Material Development

John Blake

Japan Advanced Institute of Science and Technology

Concordancers often provide an option to generate lists of keywords. Keywords are the words that occur disproportionately more frequently in a particular text type (e.g., business English) compared to another text type (e.g., general English). This is one way of distinguishing technical or domain-specific words from general words. Novice users of concordancers tend to expect that the keyword lists produced are identical, yet there are significant differences in the lists generated. This paper shows how keyword lists are affected by the choices of concordancer, reference corpus and statistical test. ESP materials developers can use this knowledge to make a more informed choice of the variables so that the most appropriate keyword list for the target audience can be created.

The identification of words that deserve inclusion in teaching materials is a difficulty that many materials developers face. There are many factors to consider in the selection of vocabulary, such as frequency, appropriacy, expediency, need and level. The most frequent words in a text are relatively easy to identify, but are not necessarily the most useful words to highlight in ESP materials. This is because grammatical words and high frequency general words are likely to occupy the top positions. Words that are key, however, are likely to merit inclusion. Specialized software programs (concordancers) can analyze collections of written texts (corpora) to identify the frequency and keyness of vocabulary.

Simply put, keyness is a measure of the frequency with which a word occurs in the corpus being analyzed (focus corpus) in comparison to another corpus (reference corpus). Words that occur more frequently show positive keyness and

are commonly called key words (Scott, 1997). For example, the words *export, firm* and *market* are more likely to occur in texts about business than in texts on general topics, so these words show positive keyness. Keyness is computed by using a concordancer to count words in the focus corpus, compare the counted frequency to counts in a reference corpus by using a statistical test. The choice of concordancer, reference corpus and statistical test affects the keyword lists generated.

Novice users may expect all concordancers to produce the same keyword list for a text. However, this is not the case. Different concordancers, reference corpora and statistical tests result in radically different keyword lists.

Concordancers can be classified into four generations (McEnery and Hardie, 2012) although the first two generations are now obsolete. Fourth-generation concordancers can deal with large corpora and are far more powerful than third-generation concordancers, such as AntConc (Anthony, 2012) and Wordsmith Tools (Scott, 2012) (Table 1). Some concordancers provide options to upload a reference corpus to which the focus corpus can be compared, while others provide a range of corpora from which the user can select. Concordancers may have a default statistical test (e.g., chi-squared in AntConc) or provide alternatives for the user to select. Keyword list generation is underpinned by comparing the ratios of words occurring in the focus and reference corpora using statistical tests. Kilgarriff (2012) highlights two statistical problems when comparing two corpora. The first is the resolution of dividing by zero when

Table 1
Current Generations of Concordancers

	3rd generation	4th generation
Location	Personal computers	Web servers
Size of corpora	Small corpora - low millions	Large corpora – 100 million+
Examples	AntConc (Anthony, 2012) UAM Corpus Tool (O'Donnell, 2013) Wordsmith Tools (Scott, 2012)	CQPweb (Hardie, 2012) Sketch Engine (Kilgariff et al., 2014) W-matrix (Rayson, 2008)

there are no occurrences of a word in the reference corpus. The second involves overcoming the domination of words which occur rarely in the reference corpus. Different tests use different methods to address these issues.

This paper explores how the choice of concordancer, reference corpus and statistical test generates different lists of keywords. Materials developers can use this knowledge to make more informed choices of which vocabulary to focus on in their tailor-made materials.

Method

A corpus of texts comprising all the research articles published in the journal *International Business Review* from February 2010 to October 2013 was manually collected and concatenated into a single text file. Table 2 shows the composition of this focus corpus.

The three variables (concordancers, reference corpora and statistical tests) were each tested in turn. A popular third-generation concordancer, AntConc 3.2.4w (Anthony, 2012), and a popular fourth-generation concordancer, Sketch Engine (Kilgarriff et al., 2014), were selected for comparison. The raw frequency word count for each concordancer was first calculated. Keyword lists were generated using the British Academic Written English (BAWE) corpus and the Brown corpus in Sketch Engine (Table 3). A keyword list was then generated using the Brown corpus in AntConc. This was undertaken using three simple maths statistical tests in Sketch Engine and two simple ratio tests in AntConc. The keyword lists were then evaluated from the perspective of an ESP materials developer.

Table 2
IBR Focus Corpus

	Count (made in AntConc 3.2.4w)
Tokens	2,516,051
Words	1,966,650
Sentences	77,547

Results

Concordancers

The raw count of frequency of words in both AntConc and Sketch Engine results in the same order for the top ten words, yet only the word count for *that* is identical (Table 4). This raw word count difference can be accounted for by differences in the operational definition of a word and the process of tokenization. For example, Anthony (2013) notes that Wordsmith Tools and AntConc count contractions differently, e.g., *we'll* is counted as one word in Wordsmith, but as two words in AntConc. Word count is just one variable in the calculation of keyness. Since results differ at the level of raw word count, this difference may be exacerbated by the choice of reference corpus and statistical test.

Each concordancer offers different functionality with regard to calculating keyness. For example, AntConc allows users to upload their own reference corpus and provides the standard choice of either chi-squared or log-likelihood for the statistical test, while Sketch Engine incorporates access to numerous reference corpora and a set of statistical tests based on simple maths (Kilgarriff, 2009). For most ESP material developers, the functionality of the concordancer is most likely of more importance than a thorough understanding of the definition of words and tokenization process used. Materials developers want to know which vocabulary deserves inclusion in materials and so selecting a concordancer that can produce key word lists that are pitched at the level and topic of the target audience is of primary importance.

Table 3
Outline of Reference Corpora Used

	BAWE corpus	Brown corpus
Date created	2000s	1960s
Type of corpus	Academic	General
Type of English	British	American
Words	6,506,995	1,000,000

Table 4
Raw Frequency Results

No		Sketch Engine	AntConc
1	the	106,022	106,064
2	and	77,508	77,542
3	of	72,733	72,990
4	to	47,454	47,834
5	in	41,791	42,056
6	a	32,007	32,336
7	that	23,092	23,092
8	is	21,249	21,245
9	for	17,293	17,303
10	as	14,309	14,329

Reference Corpora

Table 5 shows the keyword lists created in Sketch Engine using the Midway statistical test but with different reference corpora. Keyword lists created when using the BAWE corpus and Brown corpus shared five of the top ten results. The remaining five words in BAWE appeared more specialized than the Brown corpus. The BAWE keyword list, therefore, appears more appropriate for learners with a stronger vocabulary base.

Scott (2009) claims that there is no bad reference corpus. However, different reference corpora yield radically different keyword lists. Keyness is significantly affected by the genre and diachrony of a reference corpus (Goh, 2010). When the focus and reference corpus are more similar in terms of topic and time period, the keyword list is likely to contain words that are more obscure. As the BAWE corpus is more similar in terms of genre and diachrony than the Brown corpus to the focus corpus, the resultant key word list therefore contains words that are more obscure (Table 5). Given that different reference corpora impact the generated keyword lists, ESP materials developers would be well advised to compare the results using different reference corpora.

Statistical Tests

As shown in Table 6, selecting the log-likelihood and chi-squared tests in AntConc using the Brown corpus resulted in identical lists for the first eight keywords. Simple ratios, such as log-likelihood and chi-squared, produce keyword lists "dominated by rare words" (Kilgarriff, 2012, p.5). Both of these tests are based on the assumption of randomness; but language is not random (Kilgarriff, 2005), and so these tests are inappropriate (Gabrielatos **and** Marchi, 2012).

Table 7 shows the keyword lists generated in Sketch Engine using the BAWE Corpus, but selecting different statistical tests. The simple maths version (Kilgarriff, 2009) in Sketch Engine names the tests clearly (e.g., Common, Rare) and is not based on the assumption that language is random. Rare resulted in higher occurrence of rare words, while Common resulted in a skew to more common words. When selecting vocabulary for less proficient students, it may be prudent to use a keyword list generated using Common.

Table 5
Keyword Lists using BAWE and Brown in Sketch Engine with Midway Test

No	BAWE	Brown
1	firms	firms
2	firm	firm
3	export	export
4	foreign	Table
5	subsidiary	variables
6	internationalization	international
7	FDI	markets
8	subsidiaries	knowledge
9	markets	foreign
10	MNEs	market

Table 6
Keyword Lists using Log-likelihood and Chi-squared Tests in AntConc with Brown Corpus

No	Log-likelihood	Chi-squared
1	the	the
2	firms	firms
3	firm	firm
4	al	et
5	et	al
6	in	In
7	knowledge	knowledge
8	market	market
9	this	international
10	table	foreign

Table 7
Keyword Lists using Three Statistical Tests in Sketch Engine with BAWE Corpus

No	Rare	Midway	Common
1	OFDI	firms	and
2	offshoring	firm	firms
3	Vahlne	export	firm
4	multinationality	foreign	foreign
5	Full-size	subsidiary	knowledge
6	MathML	internationalization	international
7	Kogut	FDI	market
8	BOP	subsidiaries	country
9	MathJax	markets	Table
10	Ghoshal	MNEs	performance

Conclusion

The three variables of concordancer, reference corpora and statistical tests greatly affect the keyword lists generated. The functionality of fourth-generation concordancers far outweighs third-generation, and so if time is a priority, it is worth investing in a subscription. Sketch Engine provides an easy, quick, and affordable way to calculate a variety of keyword lists. The availability of 20 reference corpora and four appropriately-named statistical tests make it easy for novice users to tailor keyword lists to the intended learners. Selecting a general English reference corpus and the Common statistical test in Sketch Engine is likely to generate keyword lists that are more suitable for lower level students.

References

Anthony, L. (2012). AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research, 30*(2), 141-161.

Gabrielatos, C., & Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Paper presented at Corpus-assisted Discourse Studies International Conference 2012, University of Bologna, Italy. Retrieved from http://repository.edgehill.ac.uk/4196/

Goh, G-Y. (2010). Choosing a reference corpus for keyword extraction. *Linguistic Research, 28*(1), 239-256.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380-409.

Kilgarriff, A. (2005). Language is never ever ever random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263-276.

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference CL2009*. Retrieved from http://ucrel.lancs.ac.uk/publications/cl2009/

Kilgarriff, A. (2012, January). Getting to know your corpus. In *International Conference on Text, Speech and Dialogue* (pp. 3-15). Berlin: Springer Verlag.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7-36.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

O'Donnell, M. (2013). UAM Corpus Tool (Versions 2.8 & 3.1) [Computer Software]. Edinburgh, UK: Wagsoft Systems.

Rayson, P. (2008). W-matrix corpus analysis and comparison tool. Lancaster, UK: Lancaster University. Retrieved from http://ucrel.lancs.ac.uk/wmatrix/

Scott, M. (1997). PC analysis of key words – and key key words. *System, 25*(1), 1-13.

Scott, M. (2009). In search of a bad reference corpus. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp.79-92). Oxford: Ashgate Publishing.

Scott, M. (2012). WordSmith Tools (Version 6) [Computer Software]. Liverpool: Lexical Analysis Software.

Author bio

John Blake *is a research lecturer at the Japan Advanced Institute of Science and Technology. He has taught English at universities and schools for over 20 years in Japan, Thailand, Hong Kong and the UK. His current research interest is corpus analysis of scientific research articles. johnb@jaist.ac.jp*

Received: November 18, 2014
Accepted: September 29, 2015