Harnessing keyness: Corpus-based approach to ESP material development

JAIST JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY 1990

John Blake

Background

ESP course developers adopting a **corpus-based or corpus-driven approach**[1] can create a **focus corpus** of relevant texts and use **keyness**[2] to identify **lexical sets** to integrate into teaching materials.

Keyness

- Keyness is a measure of the frequency of disproportionate occurrence.
- Focus corpus is contrasted to a **reference corpus.**
- Operational definition of keyness[3] affects key words.
- Words may be key to whole corpus (**global**) or part (**localised or bursty**)[4].

This study investigates the effect of different **1. concordancers**, **2. reference corpora** and **3. statistical formulae** on the **keyword lists** for a focus corpus of research articles on international business.

Table 1: Focus corpus

Variable	Count
Tokens	2,516,051
Words	1,966,650
Sentences	77,547

1: Concordancer

Concordancers are designed for different purposes and budgets. Raw frequency counts also differ with concordancer[5]. **4th generation concordancers**[6] are web-based, fast and suitable for large corpora. This presentation compares the results using two popular concordancers: AntConc and Sketch Engine.

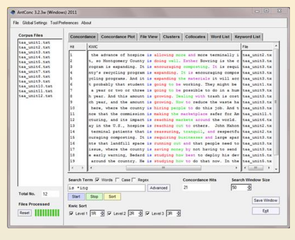

Fig 1: Sceenshot of AntConc

Table 2: Examples of 3G & 4G concordancers

3G	AntConc[7] UAM Corpus Tool[8] Wordsmith Tools[9]
4G	CQPweb[10] Sketch Engine[11] Wmatrix[12]

Raw frequency list
Sketch Engine

1	the	106,022
2	and	77,508
3	of	72,733
4	to	47,454
5	in	41,791
6	a	32,007
7	that	23,092
8	is	21,249
9	for	17,293
10	as	14,309

List 1

Raw frequency list
AntConc

1	the	106,064
2	and	77,542
3	of	72,990
4	to	47,834
5	in	42,056
6	a	32,336
7	that	23,092
8	is	21,245
9	for	17,303
10	as	14,329

List 2

2: Reference corpus

The **genre** & **diachrony** of a corpus significantly affect keyness[13, 14]. AntConc users need to provide their own reference corpus, while Sketch Engine provides access to 20 corpora, including **enTenTen12**, a 12 billion token corpus.

Table 3: Comparison of Brown and BAWE

Brown corpus	British Academic Written English Corpus (BAWE)
circa 1960s	circa 2000s
General corpus	Academic corpus
American English	British English
1,000,000 words	6,506,995 words

Keyword list
Sketch Engine
Brown
Midway (1000)

1	firms
2	firm
3	export
4	foreign
5	subsidiary
6	internationalization
7	FDI
8	subsidiaries
9	markets
10	MNEs

List 3

Keyword list
Sketch Engine
BAWE
Midway (1000)

1	firms
2	firm
3	export
4	table
5	variables
6	international
7	markets
8	knowledge
9	foreign
10	market

List 4

3: Statistical formula

Many linguists are ill at ease with statistics[15].

Main ideas

- Frequency bias of **log likelihood** and **chi squared**[3]
- Rarity bias of **mutual information**
- The **simple maths** version[16] in Sketch Engine names the formulae clearly and assumes language is not random[17].

Keyword list
AntConc
Brown
Log likelihood

1	the
2	firms
3	firm
4	al
5	et
6	in
7	knowledge
8	market
9	this
10	table

List 5

Keyword list
AntConc
Brown
Chi squared

1	the
2	firms
3	firm
4	et
5	al
6	in
7	knowledge
8	market
9	international
10	foreign

List 6

Keyword list
Sketch Engine
BAWE
Rare (0.01)

1	OFDI
2	offshoring
3	Vahlne
4	multinationality
5	full-size
6	MathML
7	Kogut
8	BOP
9	MathJAx
10	Ghoshal

List 7

Keyword list
Sketch Engine
BAWE
Midway (1000)

1	firms
2	firm
3	export
4	table
5	variables
6	international
7	markets
8	knowledge
9	foreign
10	market

List 8

Keyword list
Sketch Engine
BAWE
Common (1 million)

1	and
2	firms
3	firm
4	foreign
5	knowledge
6	international
7	market
8	country
9	table
10	performance

List 9

Summary

4G concordancers with the choice of reference corpora and formulae enable developers to tweak results to create the most useful list of keywords.

References

1. **Tognini-Bonelli, E. (2001).** *Corpus Linguistics at Work.* Amsterdam: John Benjamins.
2. **Scott, M. (1997).** PC analysis of key words – and key key words. *System, 25* (1), 1-13.
3. **Gabrielatos, C. and Marchi, A.** (2012) Keyness: Appropriate metrics and practical issues. Paper presented at Corpus-assisted Discourse Studies International Conference 2012. University of Bologna, Italy. 13-14 September, 2012.
4. **Katz, S. (1996).** Distribution of Common Words and Phrases in Text and Language Modelling, *Natural Language Engineering, 2* (1), 15-59.
5. **Anthony, L. (2012).** A critical look at software tools in corpus linguistics. *Linguistic Research, 30* (2), 141-161.
6. **McEnery, T, & Hardie, A. (2012).** *Corpus linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
7. **Anthony, L. (2012).** AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University.
8. **O'Donnell, M. (2013).** UAM Corpus Tool (Versions 2.8 & 3.1). Wagsoft Systems.
9. **Scott, M. (2012).** WordSmith Tools (Version 6). Liverpool: Lexical Analysis Software.
10. **Hardie, A. (2012).** CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics 17* (3), 380–409.
11. **Kilgarriff, A. *et al.* (2004).** The Sketch Engine. Lexical Computing.
12. **Rayson, P**. (2008). W-matrix corpus analysis and comparison tool. Lancaster University.
13. **Scott, M. (2009).** In search of a bad reference corpus. In D. Archer (ed.), *What's in Word-list? Investigating Word Frequency and Keyword Extraction (pp.79-92).* Oxford: Ashgate.
14. **Goh, G-Y. (2010).** Choosing a reference corpus for keyword extraction. *Linguistic Research, 28* (1), 239-256.
15. **Loewen, S. *et al.* (2014).** Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48* (2). 360-388.
16. **Kilgarriff, A. (2009).** Simple maths for keywords. In Mahlberg, M., González-Díaz, V. & Smith, C. (eds.), *Proceedings of the Corpus Linguistics Conference CL2009*. University of Liverpool, UK, 20-23 July 2009.
17. **Kilgarriff, A. (2005).** Language is never ever ever random. *Corpus Linguistics and Linguistic Theory* 1 (2): 263-276.