Deriving Lists of Collocates from Web Corpora: Issues and Implications of Different Association Measures

Keith Barrs

(Received on October 31, 2019)

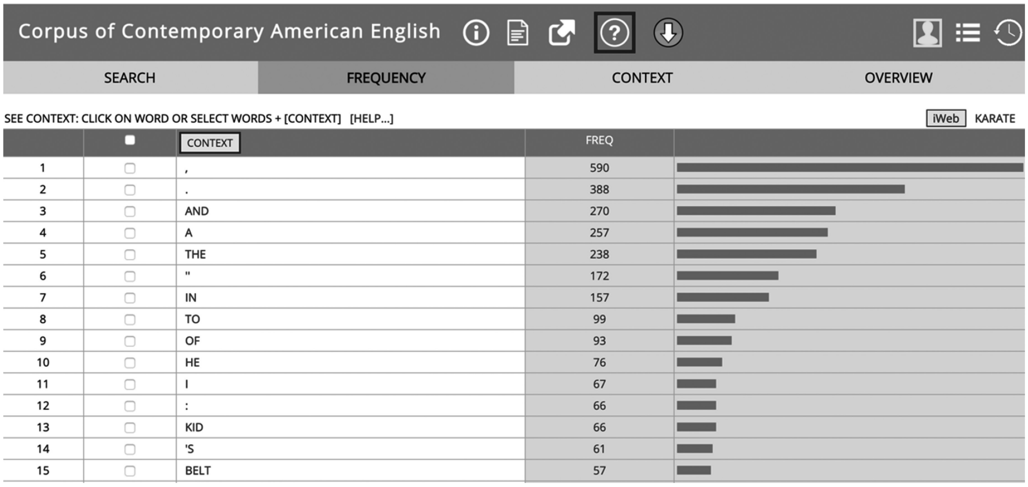
Abstract

This research investigates the issues and implications of using different association measures to derive a list of collocates for words in a web corpus. A case study was carried out on the search word 'zero' in the enTenTen12 corpus of web-based English, comparing the list of its collocates ranked by raw frequency with lists ranked by the traditional association measures of MI, t-score and log-likelihood, and then further comparing these with a list ranked by the more modern logDice statistic. It was found that the logDice measure was by far the most effective of the five at building up a semantic profile of the search word, carrying the implication that the corpus analyst needs to be aware of the various issues involved in the application of different association measures to the analyses of web corpora.

1. Introduction

An understanding of the behaviour of lexical items in natural language informs many areas of linguistic study, such as descriptions of their context-dependent frequencies for inclusion in corpus-informed dictionaries, analyses of their semantic meanings for better developing computer-based natural language processing, and evaluations of their sociolinguistic usage for helping second language learners understand how language is actually used. Whilst word behaviour can be investigated through drawing together people's intuitions, it can be explored on a far more immense scale, with similarly immense speed and reliability, through the systematic analysis of corpora (Hunston, 2002; Lindquist, 2009; McEnery & Wilson, 2001). One such method of researching word behaviour in natural language is to produce a frequency-ranked list of a word's collocates, which are the words with which it habitually co-occurs in a given linguistic context. These lists of collocates can then be used, amongst other purposes, to help sketch a preliminary semantic profile of the main word under investigation (Hunston, 2002; Kilgarriff, Rychly, Smrz, & Tugwell, 2004; Stubbs, 1996). In producing the lists, corpus software programs typically use one or more of a large number of statistical association measures with which to sort and rank a corpus of millions or billions of words, which would be all but impossible to do manually. A statistical association measure is defined as "a formula of an association score which indicates the amount of statistical association between two words" (Rychly, 2008, p. 6). It is then up to the user of the corpus software program to select the measure of statistical association most appropriate for their research purposes (Cheng, 2012).

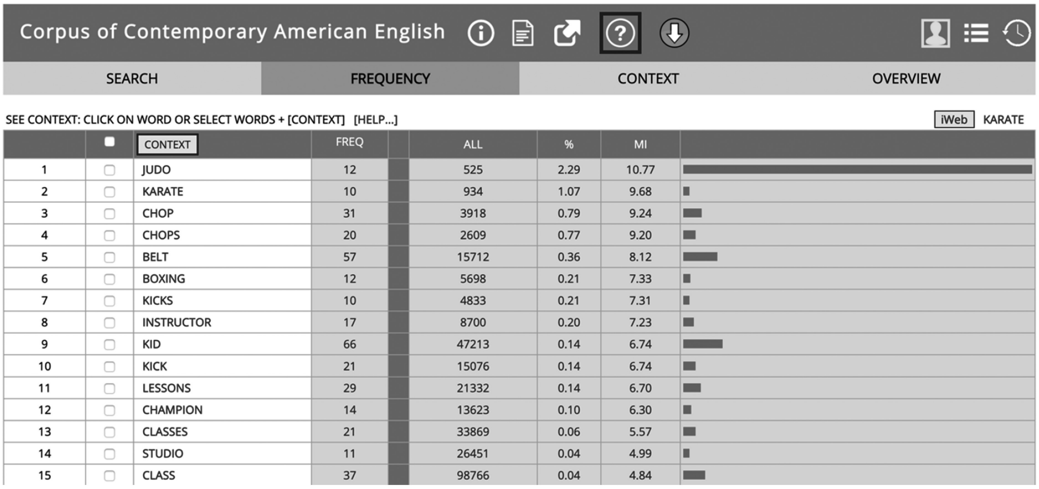
Over the relatively short history of modern (i.e. computer-based) corpus linguistics, which goes back to the 1960's with the publication of the Brown Corpus, a large number of association measures have been applied in the ranking of collocates. These measures have been used primarily because ranking collocates by raw frequency alone, whilst initially informative, does not express the strength of association and/or statistical significance between the node and collocate. This strength of association can be thought of as the amount of magnetic pull between words. As an example, Figure 1 below shows the top 15 collocates of the search word 'karate' in the Corpus of Contemporary American English (COCA), ranked by raw frequency. Ignoring the punctuation marks and function words higher up on the list, the first word of notable interest is 'kid', at rank 13, from the collocation 'The Karate Kid' (a popular 1980's movie). This shows that the word 'kid' is a frequent collocate of 'karate', but 'kid' is also likely to be a frequent collocate of many other words that are not particularly associated with 'karate', such as 'Sundance' (i.e. the nickname of the American outlaw Harry Longabaugh), 'little' and 'cute'.



SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

	■	CONTEXT	FREQ
1	□	,	590
2	□	.	388
3	□	AND	270
4	□	A	257
5	□	THE	238
6	□	"	172
7	□	IN	157
8	□	TO	99
9	□	OF	93
10	□	HE	76
11	□	I	67
12	□	:	66
13	□	KID	66
14	□	'S	61
15	□	BELT	57

Figure 1. The top 15 collocates of 'karate', ranked by raw frequency.



SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

	□	CONTEXT	FREQ	ALL	%	MI
1	□	JUDO	12	525	2.29	10.77
2	□	KARATE	10	934	1.07	9.68
3	□	CHOP	31	3918	0.79	9.24
4	□	CHOPS	20	2609	0.77	9.20
5	□	BELT	57	15712	0.36	8.12
6	□	BOXING	12	5698	0.21	7.33
7	□	KICKS	10	4833	0.21	7.31
8	□	INSTRUCTOR	17	8700	0.20	7.23
9	□	KID	66	47213	0.14	6.74
10	□	KICK	21	15076	0.14	6.74
11	□	LESSONS	29	21332	0.14	6.70
12	□	CHAMPION	14	13623	0.10	6.30
13	□	CLASSES	21	33869	0.06	5.57
14	□	STUDIO	11	26451	0.04	4.99
15	□	CLASS	37	98766	0.04	4.84

Figure 2. The top 15 collocates of 'karate', ranked by mutual information (MI) association measure.

To get a better sense of the collocates which are more strongly associated with (or magnetically attracted to) 'karate', it is useful to use an association measure. Figure 2 again shows the top 15 collocates of 'karate', but this time ranked with the 'mutual information' association measure (a description of this measure is given below). In this list, 'judo' is ranked at the top of the list and represents a collocate that naturally feels more strongly associated with 'karate' (i.e. both being martial arts), than does the collocate 'kid'.

Association measures work by comparing what has been observed about the co-occurrence of the node (i.e. the main search word) and a collocate, with what would be expected under the null hypothesis (i.e. the assumption that the node has no statistically significant influence over the words that surround it). Commonly-used measures of statistical association in the field of corpus linguistics are the MI, t-score and log-likelihood measures (Hunston, 2002; Lindquist, 2009; McEnery & Hardie, 2012). Despite their popularity, each measure has a weakness: MI tends to highlight collocates for which there is little evidence in the corpus, t-score tends to highlight function-word collocates, and log-likelihood produces a list of function words and punctuation, with lexical collocates appearing much lower down the ranked list. The weaknesses of these measures are typically acknowledged as an inconvenience of which the analyst needs to be aware (Lindquist, 2009; McEnery, Xiao, & Tono, 2006), and as a result many corpus analysis software programs such as AntConc and the COCA online interface allow the user to exclude function words and punctuation from the analyses. A more serious problem arises, however, when the association measures are applied to web corpora. With web corpora, these weaknesses are amplified because the massive size of the corpora increases the amount of rare words, which affects the MI measure, and increases the frequency of function words and punctuation, which affects the T-score and log-likelihood measures.

This research investigates the issues and implications of using different association measures to derive a list of collocates for words in a web corpus. Specifically, a case study was carried out on a search word in a large web corpus, comparing the list of its collocates ranked by raw frequency with lists ranked by the traditional association measures of MI, t-score and log-likelihood, and then further comparing these with a list ranked by the more modern logDice measure.

2. Methodology

The word ‘zero’ was chosen as the main search word for this case study for two main reasons. First of all, it is one of the most frequent words in English, meaning that tens of thousands of examples of this word in natural language usage can be processed by a computer to inform the collocation analysis. Secondly, because of the fact that ‘zero’ can be used as both a numeral (0) and a word (zero) it is likely to collocate with not only a wide variety of lexical words, such as ‘tolerance’, but also functional words (e.g. ‘and’) and punctuation (e.g. ‘.’). In this way, the various lists of collocates produced by the application of different association measures can be compared for how well they sort through the hundreds of thousands of collocational pairs to give a list which provides an overall semantic profile of the way in which ‘zero’ is used in the web corpus.

The Sketch Engine corpus query system (www.sketchengine.co.uk/) was used as the software with which to generate the various lists of collocation candidates. This software not only allows the application of various association measures in the construction of word lists, but also includes access to hundreds of language corpora, including the enTenTen12 corpus of around 10 billion words of web-based English. Five separate collocation lists were generated using the default setting of a contextual span of 5 words to the left and right of the search word. The five lists were generated by applying (1) raw frequency, (2) MI, (3) t-score, (4) log-likelihood, and (5) logDice to the ranking option. Because of the massive size of the corpus, the lists for each method of ranking included thousands of collocates. For this case study, only the top 20 collocates in each list were compared because this was considered sufficient to show the major differences between each list.

3. Results

As discussed in the introduction above, there are known issues with the formulaic workings of the MI, t-score, and log-likelihood association measures. In general, the weakness of MI is that it tends to highlight collocates which are rare in the corpus, whilst the weakness of the t-score and log-likelihood measures is that they tend to highlight function words and punctuation (McEnery & Hardie, 2012). These issues are then likely to get magnified when used in the analysis of web corpora, due to their massive amount of text combined with the very high likelihood of spelling, grammar, and formatting issues.

Table 1 shows the top 20 collocates for the English word 'zero', ranked initially by raw frequency and then by MI, t-score, log-likelihood and logDice. As can be expected, the raw frequency list gives little to work with when attempting to build up a semantic profile

Table 1. A comparison of the collocates produced by different ranking methods.

Rank	Raw Frequency		MI		t-score		log-likelihood		logDice	
1	.	189438	xpeople	15.311	.	395.043	.	641005.57	Ground	9.176
2	,	144808	dougpositive	15.044	,	334.836	,	402171.22	zero	8.782
3	the	133602	Paymydownpayment	14.851	the	314.867	to	345614.6	tolerance	8.713
4	to	111724	mask-charge	14.851	to	301.849	the	332816.48	Zero	7.841
5	and	85901	pointsaffiliations	14.851	and	257.555	is	235870.91	Double	7.653
6	a	77624	UnionsAbsolutely	14.851	a	249.234	and	225711.1	cost	7.251
7	of	75102	Kiyona	14.851	is	239.744	a	222246.29	emissions	7.235
8	is	66836	tazzari	14.851	of	237.233	of	181910.05	Waste	7.175
9	in	57581	budgetaug	14.851	in	214.61	in	162256.11	gravity	7.087
10	with	40860	degreesbelow	14.851	with	188.077	with	145225.34	ground	7.061
11	for	37856	Wait-State	14.851	for	174.854	Ground	131636.04	degrees	6.846
12	that	34553	predecession	14.851	on	169.934	cost	129623.48	reset	6.808
13	on	33961	Fome	14.819	that	163.719	zero	121395.94	carbon	6.783
14	have	27222	Tsukaima	14.805	have	153.514	on	113066.01	near	6.614
15	or	26143	lygerzero	14.787	or	150.576	for	108046.88	balance	6.612
16	you	25362	Craigslitst	14.752	(	147.93	have	95975.606	interest	6.601
17	be	25200	childrenreadingbookswithparents	14.714	at	147.68	tolerance	93109.086	below	6.552
18	(	24940	superif	14.699	)	144.059	or	92703.602	sum	6.531
19	at	24784	distributelab	14.681	be	143.394	at	92661.343	percent	6.527
20	The	24461	MeltDown	14.681	The	141.809	(	92233.513	mosque	6.493

of the word 'zero' in the enTenTen12 corpus, as it highlights only grammatical function words and punctuation. For the MI score, which is biased towards low-frequency words, of which there are many in web corpora, the list of collocates produced by this measure seems to be dominated by company names, web addresses and spelling mistakes. Indeed, the average frequency of these collocates is 24.6, which in a ten- billion-word corpus is extremely low. The t-score and log-likelihood measures are biased towards frequent words in the corpus and because the enTenTen12 corpus contains around ten-billion words, these scores produce lists similar to the raw frequency list in that they are dominated by function words and punctuation items. However, log-likelihood is slightly the better of the two as it includes several words which start to give a sense of how 'zero' is used in the corpus (Ground, cost, tolerance).

When the collocates are ranked by logDice, which is a variant form of the Dice score that fixes the issue of the scores being very low numbers (Rychly, 2008, p. 6), the list is markedly different. The logDice measure brings out collocates which give a clear overview of the variety of ways in which 'zero' is being used. In other words, the logDice association measure brings to the top of the list collocates which have a stronger, more magnetic relationship with the search word. A check of a sample of the concordance lines of each collocate showed that 'Ground' very often refers to 'Ground Zero' (which also explains the frequent occurrence of 'Zero' with a capitalized 'Z'), the collocates 'emissions' and 'waste' show that zero is being used to talk about the reduction and control of something, and 'cost', 'interest', 'tolerance' and 'gravity' reveal that zero relates generally to an absence of something. As such, with the lexical rather than grammatical collocates derived from the application of the logDice association measure, it is possible to begin building a semantic profile of the word 'zero', grouping the collocates into different categories, such as 'location', 'number', 'reduction', and 'absence'. Such categorical groupings are not possible with these top 20 collocates in each of the other lists, and instead would need much longer lists.

4. Conclusion

This small case study of the various collocates of the search word 'zero' generated by the application of different association measures sheds light on the issues involved when analysing large-scale web-based corpora. The massive number of words in large-scale web corpora, such as the enTenTen12, most of which are function words, causes the t-score and log-likelihood measures to rank function words and punctuation high on the collocate lists. And the fact that web corpora contain many slogans, company names, web addresses and spelling errors also means that the MI association measure is similarly problematic. In contrast, the results in Table 1 showed the logDice score to be an effective association measure at bringing high quality collocates to the top of collocation list. This statistic has other strengths as well in that it is not corpus- specific because it does not depend on corpus size, so a logDice score from one corpus can be compared to the score in another corpus of different size (Rychly, 2008, p. 8). And further, the theoretical maximum score of logDice is 14, which means it is much easier for the user to comprehend than some of the very large and very small numbers given by other measures (Rychly, 2008, p. 9).

The main implication of these results is that because corpus analysis software usually comes with default sorting and ranking options applied to the creation of word lists, with the option of selecting a different measure based on user preference, the user of the various corpus analysis software tools needs to be aware of the various issues involved in the application of different association measures. With this knowledge, they can then make an informed decision as to whether they keep the default option or make a new selection of association measure when running their analyses. Indeed, when analysing web-based corpora, it may even be that a new corpus analysis software tool needs to be selected if the current selection does not include a sufficient range of association measures.

References

Cheng, W. (2012). *Exploring corpus linguistics: Language in action.* Abingdon, England: Routledge.

Davies, Mark. (2008–). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.* Available online at https://www.english-corpora.org/coca/.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge, England: Cambridge University Press.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *The Proceedings of the Eleventh Euralex International Congress* (pp. 105–116). Lorient, France: Information Technology Research Institute. Retrieved from ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-08.pdf

Lindquist, H. (2009). *Corpus linguistics and the description of English.* Edinburgh, Scotland: Edinburgh University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics.* Cambridge, England: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus based language studies.* Abingdon, England: Routledge.

Rychly, P. (2008). A Lexicographer-Friendly Association Score. *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 6–9.

Stubbs, M. (1996). *Text and corpus analysis.* Oxford, England: Blackwell Publishers Ltd.