International Journal of Education and Social Science; Vol. 6 No. 3; March 2019
ISSN 2410-5171 (Online), 2415-1246 (Print)
Published by Research Institute for Progression of Knowledge

Using iWeb Corpus to Explore the Collocation of *Statistics*

Chunmei Lu
School of English for International Business
Guangdong University of Foreign Studies
China

Abstract

Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation, which is widely used in different fields of studying, e.g. corpus linguistics. Based on a 14-billion-word iWeb Corpus, this paper aims at investigating the top 50 collocation of statistics by dividing them into six categories, including the category of Terms, Disciplines, Topics, Organizations, Spots and others. The result shows that: 1) in the category of Terms, statistics usually collocates with probability, descriptive, inferential and population; 2) in the category of Disciplines, four subjects including statistics, mathematics/math, economics and calculus frequently collocate with statistics; 3) in the category of Topics, crime, employment and justice are the three topics that mostly collocate with statistics in iWeb corpus; 4) in the category of Organizations, ABS, BLS, Statistics NZ are the three main departments; 5) in the category of Spots, Canada, USA, Australia and New Zealand are the four countries frequently collocate with statistics. Moreover, there's no relevance between the frequency of a collocate and its strength (measured by MI value).

Keywords: Statistics, Collocation, iWeb corpus

1. Introduction

Statistics is "a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation" (Dodge 2006), which is pervasively applied in different disciplines. Corpus linguistics is one of such branch, "a scientific method of language analysis which requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language" (Brezina 2018:2).

Studying collocation is one of the major branches in corpus linguistics. As Firth (1957: 6) puts it, "You shall know a lot about a word from the company it keeps." Collocation is thus "a way of understanding meanings and associations between words" (Baker 2006: 96). Node and collocates are the two key terms of collocation. Node refers to "a word that we want to search for and analyze" and collocates mean "words that co-occur with the node in a specifically defined span around the node, which we call the collocation window" (Brezina 2018:67).

In view of the significance of statistics and driven by the curiosity of finding out what is relevant to statistics, this paper thus aims to explore the collocation of *statistics* by using authentic data from a 14-billion-word corpus called iWeb, "which makes about 25 times as big as
COCA (560 million words) and about 140 times as big as the BNC (100 million words)" (https://corpus.byu.edu/iweb).

2. Related Studies

It was not until the twentieth century that statistics has come to be recognized as a separate discipline (Stigler 1986). "Making inductive inferences regarding various phenomena like social tension, frustration among educated youths based on evidences gathered is the role of statistics" (Mukherjee et al. 2018: 4), which pushes statistics into a widely applicable subject to different fields of studying, especially in social sciences, like linguistics.

Corpus linguistics is an applied linguistics combining statistics with the investigation of linguistic problems, which "actually depends on both quantitative and qualitative techniques" (Baker 2006: 2). More specifically, as Biber *et al*. (1998: 4) points out "Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors. However qualitative interpretation is also an essential step in any corpus-based analysis". Collocate is one of the three main function of corpus tool like AntConc and Wordsmith.

Lei & Liu (2018: 217) proposes that the various definitions of collocation may be grouped into two major meanings for the purposes of simplicity: 1) the countable use of the term designating habitual combinations of words as lexical items, such as look up in a dictionary; and 2) the uncountable use of the term referring to the linguistic property that some words tend to occur together but do not actually constitute lexical units (Sinclair 1991). Though researchers on collocation teaching and material developers prefer the first definition (Nesselhauf 2003; Siyanova & Schmitt 2008; Laufer & Waldman 2011), it's of great necessity to take the second definition into consideration for the purpose of investigating "the co-occurring strength or probability of words for understanding the relationships among topics and themes in various discourse analysis" (Liu & Lei 2018: 218), just as this paper adopts.

Collocation has triggered numerous studies: Based on the data from the EFL learner essay corpus and English native speaker corpus, Chen & Lin (2010) analyses the differences between EFL learners and native speakers in using colligations and collocations of the high-frequency word *good* and explores the problems in using this word of EFL learners to make suggestions for English language teaching and learning. Yamashita & Jiang (2010) investigates first language (L1) influence on the acquisition of second language (L2) collocations, finding out that both L1 congruency and L2 exposure affect the acquisition of L2 collocations with the availability of both maximizing this acquisition; it is difficult to acquire incongruent collocations even with a considerable amount of exposure to L2; and once stored in memory, L2 collocations are processed independently of L1. Possible differences in acquiring congruent and incongruent collocations are discussed. Starting from various defining approaches to collocation, Li (2017) discusses various approaches to collocation measurement and their corresponding problems as existent in corpus research, in an attempt to explore and evaluate both the application and significance of corpus analysis of Chinese learners' English collocations, It has been concluded that different approaches to defining collocation, being themselves indicative of distinctive perspectives as adopted in different fields, demonstrate a diversifying trend of collocation studies, and the measurement and retrieval of collocations cannot completely replace human decision analysis, since meaning production and negotiation do not purely conform to logic reasoning and probability statistics. Lei & Liu (2018) proposes a comprehensive and type-balanced academic English collocation list (AECL), which is based on a large corpus of academic English and was created to cover the types of collocations that will be most useful to ESL/EFL learners. AECL is the result of an innovative research-based procedure that involves a five-step selection method. A comparison of the collocations on AECL with those found in well-known collocation dictionaries of general English and on three existing academic English collocation lists indicates that AECL indeed contains mainly academic rather than general English collocations. In addition, AECL is more comprehensive with regard to the types of collocations that are relevant to learners.

Even though there are abundant studies on collocation, there's still a lack of researches using corpus methods to explore the collocation of *statistics*, which turns out to be the aim of this study.

3. Method

3.1 Corpus Data: iWeb

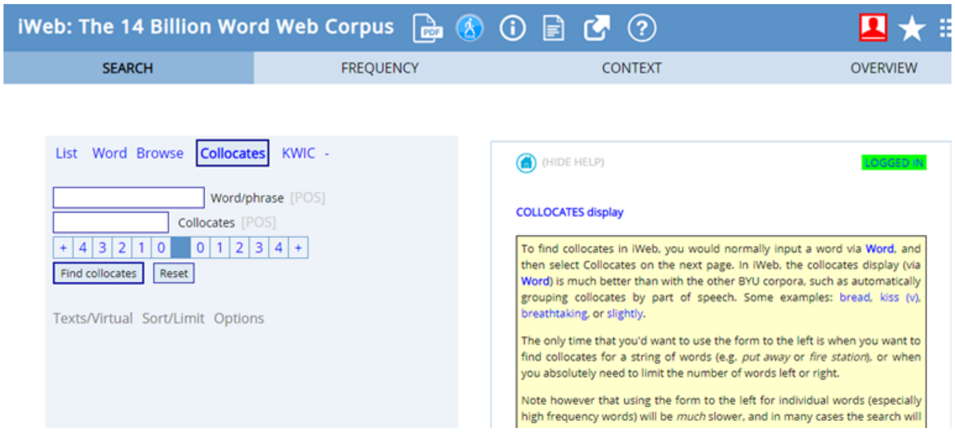
The iWeb corpus belongs to BYU corpora family and was released in May 2018. Compared to other corpora, iWeb corpus has something unique. Firstly, iWeb is about 14 billion words in size, which makes about 25 times as big as COCA (560 million words) and about 140 times as big as the BNC (100 million words). Secondly, virtual corpus for any topic (e.g. economics, statistics, accounting) could be created in just 4-5 seconds. Thirdly, with iWeb we can have access to the wide range of searches that we have for all of the other BYU corpora, including: words, phrases, substrings, lemmas, part of speech, synonyms, and customized wordlists. Fourthly, using iWeb language learners and teachers can browse through a list of the top 60,000 words (lemmas) in the corpus, and then to see an extremely wide range of information on each of these words.

In addition, it is different from any of the other BYU corpora in the attention that it gives to the top 60,000 words in the corpus, and the wide range of information for each word, including frequency information, definitions, synonyms, WordNet entries, related topics, concordances (new display in iWeb), clusters, websites that have the word as a "keyword", and KWIC/concordance lines (https://corpus.byu.edu/iweb).

3.2 Analytical Procedure

The present study follows three steps: search the collocations of *statistics* in iWeb corpus (Step 1); interpret the searching result of the collocation of *statistics* in iWeb corpus (Step 2), and conduct a top 50 frequency list of the collocations of *statistics* for further analysis (Step 3).

Firstly, to search the collocations of *statistics* in iWeb corpus, we need to enter the page of searching for collocation, click the "Collocates" button as Figure 1 indicates. "COLLOCATES display" refer to finding out "what words occur near other words, which provides great insight into meaning and usage" (https://corpus.byu.edu/iweb).



Word/phrase [POS]

Collocates [POS]

COLLOCATES display

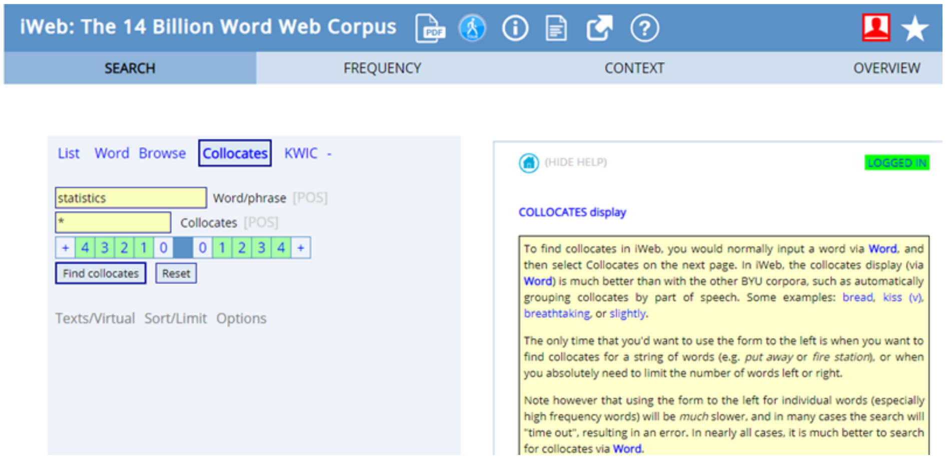
To find collocates in iWeb, you would normally input a word via Word, and then select Collocates on the next page. In iWeb, the collocates display (via Word) is much better than with the other BYU corpora, such as automatically grouping collocates by part of speech. Some examples: bread, kiss (v), breathtaking, or slightly.

The only time that you'd want to use the form to the left is when you want to find collocates for a string of words (e.g. *put away* or *fire station*), or when you absolutely need to limit the number of words left or right.

Note however that using the form to the left for individual words (especially high frequency words) will be *much* slower, and in many cases the search will

Figure 1. Enter the page of searching for collocation

After entering the page of searching for collocation, input *statistics* in the section of "Word/phrase". Then "select the 'span' (number of words to the left and the right) for the collocates. Use + to search more than four words to the left or right, and 0 to exclude the words to the left or right. If you don't select a span, it will default to 4 words left and 4 words right. The direction of the collocates and the length of the "span" between the "node word" and the collocates is quite important" (https://corpus.byu.edu/iweb). Figure 2 presents this process clearly, and this study employ the initial setting of the collocation span, that is 4 words left and 4 words right, which can find out more collocation compared to the span of 1-3 words left and 1-3 words right.



statistics Word/phrase [POS]

* Collocates [POS]

COLLOCATES display

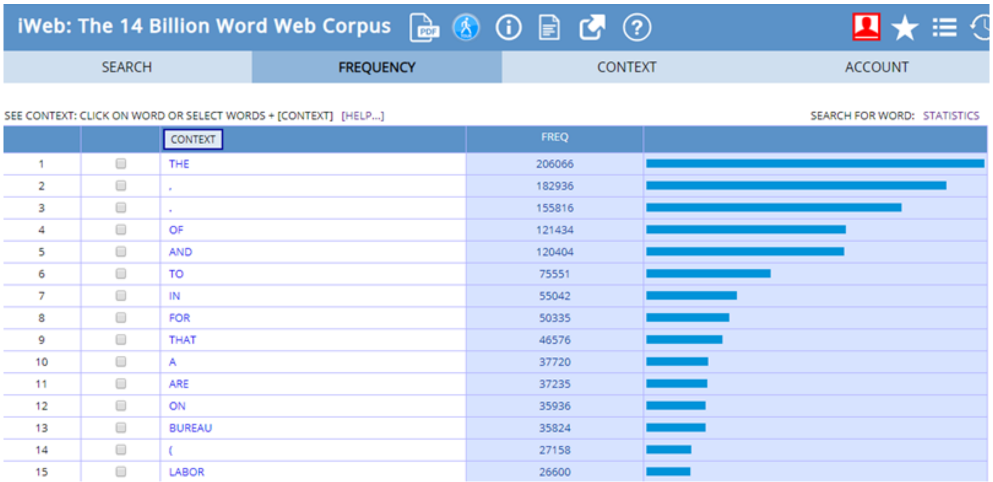
To find collocates in iWeb, you would normally input a word via Word, and then select Collocates on the next page. In iWeb, the collocates display (via Word) is much better than with the other BYU corpora, such as automatically grouping collocates by part of speech. Some examples: bread, kiss (v), breathtaking, or slightly.

The only time that you'd want to use the form to the left is when you want to find collocates for a string of words (e.g. *put away* or *fire station*), or when you absolutely need to limit the number of words left or right.

Note however that using the form to the left for individual words (especially high frequency words) will be *much* slower, and in many cases the search will "time out", resulting in an error. In nearly all cases, it is much better to search for collocates via Word.

Figure 2. Input *statistics* and searching for collocation

Next, click on the button of “Find collocates” and the results are shown as Figure 3 presents.



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

		CONTEXT	FREQ	
1	□	THE	206066	
2	□	,	182936	
3	□	.	155816	
4	□	OF	121434	
5	□	AND	120404	
6	□	TO	75551	
7	□	IN	55042	
8	□	FOR	50335	
9	□	THAT	46576	
10	□	A	37720	
11	□	ARE	37235	
12	□	ON	35936	
13	□	BUREAU	35824	
14	□	(	27158	
15	□	LABOR	26600	

Figure 3. Output of the collocation of *statistics*

Secondly, interpret the searching results of the collocation of *statistics* from Figure 3. The parameter shown on the table heading is “FREQ”, standing for the frequency of each collocation of *statistics*. The horizontal bar chart reveals the frequency of each collocate more vividly. Take the word “bureau” for instance. *Bureau* collocates with the node *statistics* for 35,824 times. What’s more, original context for each collocation of *statistics* can be viewed if clicking on the blue horizontal rectangular.

Thirdly, conduct a top 50 frequency list of the collocations of *statistics* for further analysis. From Figure 3 above, we can see many of the top collocates are functional words (e.g. the, of, to) which cannot find anything relevant to *statistics*, thus the top 50 frequency list only includes lexical collocates of *statistics*, i.e. “the nouns, verbs, adjectives and lexical adverbs” (Baker 2006: 54).

4. Results and Analysis

4.1 Frequency-based Collocates of *Statistics*

Table 1 displays the observed frequency of the top 50 collocates of s*tatistics* in iWeb corpus, the total frequency of which is 262, 941.

Table 1. The top 50 collocates of *statistics* in iWeb corpus

Rank	Collocates	Freq.	Rank	Collocates	Freq.
1	bureau	35824	26	facts	2857
2	labour	26600	27	population	2496
3	show	17448	28	applied	2449
4	national	17068	29	justice	2344
5	according	12802	30	indicate	2341
6	Canada	10672	31	figures	2217
7	U.S.	8959	32	economics	2160
8	office	8490	33	math	2105
9	statistics	7701	34	detailed	2099
10	crime	6619	35	survey	2034
11	vital	5631	36	census	2015
12	Australian	5630	37	showed	1951
13	mathematics	5614	38	trends	1674
14	reports	4987	39	compiled	1533
15	official	4922	40	inferential	1494
16	probability	4601	41	ONS	1487
17	descriptive	4318	42	reporting	1475
18	latest	4106	43	ABS	1451
19	BLS	4067	44	analytics	1439
20	usage	4000	45	collect	1403
21	analysis	3950	46	reveal	1395
22	released	3885	47	Zealand	1316
23	summary	3473	48	aggregate	1278
24	employment	3043	49	calculus	1252
25	reported	3024	50	collected	1242

From Table 1 above, "bureau" is the top collocates with *statistics* (n=35,824), and "labour" ranks the second (n=26,600). However, it would be much clear if divide the top 50 collocates of *statistics* into 6 main categories, including Statistical Terms, Disciplines, Topics, Organizations, Spots and Others, as Table 2 presents.

Table 2. Frequency of Each Collocation of *Statistics* in iWeb

Category	Frequency of Each Collocation of *Statistics*	Total frequency of each category	%
Statistical Terms	probability (4601); descriptive (4318); population (2496); inferential (1494)	12,909	4.9%
Disciplines	statistics (7701); mathematics (5614); economics (2160); math (2105); calculus (1252)	18,832	7.2%
Topics	crime (6619); employment (3043); justice (2344); survey (2034)	14,040	5.3%
Organizations	bureau (35824); labour (26600); national (17068); office (8490); official (4922); BLS (4067); census (2015); ONS (1487); ABS (1451)	101,924	38.8%
Spots	Canada (10672); U.S. (8959); Australian (5630); Zealand (1316)	26,577	10.1%
Others	show (17448); according (12802); vital (5631); reports (4987); latest (4106); usage (4000); analysis (3950); released (3885); summary (3473); reported (3024); facts (2857); applied (2449); indicate (2341); figures (2217); detailed (2099); showed (1951); trends (1674); compiled (1533); reporting (1475); analytics (1439); collect (1403); reveal (1395); aggregate (1278); collected (1242)	88,659	33.7%

The category of Organizations accounts for 38.8% among the six categories. The category of others ranks number two, involving general expressions to organize the discourse (e.g. according, released, reported) and specific expressions of statistics (e.g. analysis, figures, analytics). According to the category of Spot, Canada, U.S., Australia and New Zealand are the four countries that frequently mentioned in iWeb corpus. Disciplines of Statistics, mathematics/math, economics and calculus, the sub-category of math are closely linked to the node *statistics*. Crime, employment and justice are the three main topics that *statistics* concerns, contributing 7.2% to the total frequency of the top 50 collocates. When it comes to the Statistical Terms, probability, population, descriptive and inferential are the most frequent ones, the latter of which are the two main branches of *statistics*.

4.2 Categories of the Collocation of *Statistics* in iWeb Corpus

Six categories of the collocation of *statistics* in iWeb corpus would be discussed in this section, which involves Statistical Terms, Disciplines, Topics, Organizations, Spots and Others.

4.2.1 Statistical Terms

Defined as "the chance that a particular event will occur" (Groebner et al. 2009: 147), *probability* is the term that collocates most of the times (n=4,601) with *statistics* in the form of "*probability……statistics*". See example (1)-(3):

(1) Donald Trump becoming Republican nominee stands out he estimated only a 2% *probability*. Even though *statistics* are not about actualities but *probabilities*, subsequent events do not appear to be consistent. (iWeb_2017_necsi.edu)

(2) "……so *probability* is just like *statistics*? " "The truth is, they are related - but not as you imagined. (iWeb_2017_atarnotes.com)

(3) The statistical topics include *descriptive statistics*; hypothesis testing; *probability* distribution; Bayesian *statistics*; predictive modeling; and unsupervised learning. (iWeb_2017_bryant.edu)

Receiving the information of *statistics* are not about actualities but *probabilities* in example (1), we may begin pondering as example (2) indicates "so *probability* is just like *statistics*?" and iWeb corpus provides us with an answer like "they are related but not just like." However, probability and statistics have so many similarities that confuse novice learners. Example (3) offers us a hint that they're different by illustrating *probability* distribution is one of the statistical topics.

Descriptive (n=4318) and *inferential* (n=1494) refer to the second and fourth frequent collocate of statistics in the category of Statistical Terms, which always collocates with *statistics* in the form of *descriptive/inferential statistics*. View example (4)-(5):

(4) There are two broad categories of statistics: *descriptive statistics* and *inferential statistics*. *Descriptive statistics* are used to summarize the data and include things like average. (iWeb_2017_sciencebuddies.org)

(5) *Inferential statistics* is used to draw conclusions about a population by studying a sample. (iWeb_2017_andrews.edu)

From example (4) and (5), it's clear that descriptive and inferential statistics are two major branches of statistics and they deal with different kinds of statistical problems.

Defined as "the set of all objects or individuals of interest or the measurements obtained from all objects or individuals of interest" (Groebner et al. 2009: 14), *population* (n=2496) is the third frequent collocate of *statistics* in the category of Statistical Terms and appears with *statistics* in the form of "*population……statistics*" in iWeb corpus (example 6).

(6) Transylvania County's elderly make up 28.5 percent of the *population*, according to *statistics* from 2014, the most recent for the county. (iWeb_2017_blueridgenow.com)

Sample is the term relevant to population, but it's not included in the top 50 even top 100 collocates of *statistics* as *population* does.

4.2.2 Disciplines

Unexpectedly, *statistics* (n=7,701) is the collocate that mostly occurs with *statistics*. See example (7) and (8):

(7) Additionally, the US Bureau of Labor *Statistics*' Occupational Employment *Statistics* survey displays the median annual wages for high-tech jobs in the New York City metropolitan. (iWeb_2017_ computertrainingschools.com)
(8) Its all about *statistics*. By reviewing *statistics* relating to minimum credit scores and FHA loans, the government can see certain patterns…(iWeb_2017_ homebuyinginstitute.com)

It's impossible to distinguish the node and collocate when two of them are the same. As the examples above, two *statistics* in example (7) has turned into the name of an organization, while example (8) offers an occasional circumstance of collocation. It seems that *statistics* collocates with *statistics* just coincidently sometimes. Mathematics (math), economics and calculus are disciplines collocate frequently with statistics. See example (9)-(10):

(9) Courses that are most similar and helpful are: *mathematics* (algebra, *statistics*, *calculus*). (iWeb_2017_ foothill.edu)
(10) In a subject with high numerical content (e.g. *Economics*, *Mathematics*, *Statistics* or Geography) Excellent written/verbal English Good inter-personal skills and the ability to communicate technical… (iWeb_2017_zoek.uk)

Example (9) combs the relationship among *mathematics, statistics* and *calculus*. The latter of the two subjects actually underlie the discipline of *mathematics*. The reason why disciplines like *economics* often collocate with *statistics* is because they all prefer numerical content.

4.2.3 Topics

Crime (n=6619), *employment* (n=3043) and *justice* (n=2344) are the three collocates often co-occur with *statistics*, which have been categorized into Topics in this paper. View example (11)-(14).

(11) See the University Police website for more information on *crime* prevention, *crime statistics* and more. (iWeb_2017_ txstate.edu)
(12) The second step taken in NYC was the use of *crime statistics*. These *statistics* were real data and not just vendor supplied "fear, uncertainly and doubt" (iWeb_2017_ kraftkennedy.com)
(13) Labor force, *employment*, and unemployment *statistics* for persons with or without certifications and licenses credentials that demonstrate a level of skill or… (iWeb_2017_gpo.gov)
(14) No group in the criminal *justice* community doubts the *statistics*, the projections of prison population growth, or the reality of human rights abuses… (iWeb_2017_insightcrime.org)

From example (11)-(12), it's clear that *crime statistics* has nearly become a semi-fixed collocation in iweb corpus, indicating *crime* is one topic that *statistics* concentrates on. In example (13), it's not surprising that *statistics* involves both *employment* and unemployment is of vital importance, as the rate of unemployment has been a crucial indicator in economics, though *employment* collocates with *statistics* far more than that of unemployment. In example (14), *justice* does not collocate closely with *statistics* compared to *crime* and *employment*.

4.2.4 Organizations and Spots

The category of Organization and the category of Spots interact with each other. See example (15)-(18).

(15) According to the *Bureau* of *Labor* and *Statistics*. The best way to protect yourself? Arm yourself with a healthy level of… (iWeb_2017_veteransunited.com)
(16) According to the *Bureau* of *Labor* and *Statistics*, approximately 62.8 million Americans volunteered at least once last year. (iWeb_2017_ volunteerhub.com)
(17) In 2010, the latest year for which the *BLS* has released *statistics*, there were 63 workplace fatalities in a BLS-defined- real estate- industry subcategory. (iWeb_2017_inman.com)
(18) …jointly by a project team from the Australian Bureau of Statistics (*ABS*), Statistics New Zealand (*Statistics NZ*) and the Australian Government Department of Education… (iWeb_2017_abs.gov.au)

Bureau of Labor and Statistics (*BLS*) refers to the statistics department of USA, *ABS* refers to that of Australia and *Statistics NZ* goes to New Zealand's. Thus, it's not surprising to find that *USA*, *Australia* and *New Zealand* are the three countries or spots that collocate quite frequently with *statistics*.

From the eighteen examples above, iWeb corpus offers us a platform to understand important statistical terms, disciplines concerning statistics, topics, organizations and spots in an easy-acceptable way.

4.3 Collocation Measure of *Statistics*

Though we have gained many useful pieces of information after describing the top 50 collocates of *statistics*, this frequency-based method for collocation studying triggers some other problems. For instance, “they are so common that their regular co-incidence comes about by chance” or “misses word pairs which we might consider collocationally interesting, since strongly associated word pairs composed of words which are individually rare (*zero-sum game*, *abject poverty*) would not register at all” (Durrant & Doherty 2010: 129).

Thus, we need to go beyond frequency by pondering questions like: As Table 3 shows, *crime* collocates more often with *statistics* than *probability*, can we infer that *crime* is strongly associated with *statistics* than *probability*?

Table 3. Frequency of Each Collocation of *Statistics* in iWeb

node	collocates	frequency
statistics	crime	6619
statistics	probability	4601

To solve this problem, we need to calculate how strong these two collocations are, which we may turn to “mutual information” (MI) for help, the most commonly used statistical measure for this purpose and is “calculated by examining all of the places where two potential collocates occur in a text or corpus” (Baker 2006: 101). Corpus tools like AntConc and Wordsmith own statistical option of MI. A MI-score of 3 or higher can be taken to be significant” (Hunston, 2002: 71). MI score is a ratio of the observed frequency (fo) of the combination divided by the expected frequency (f e) of the combination: MI $=$ fo $/$ f e. And the formula of f e is:

$f\ e = (Target$ *word frequency*Collocate word frequency)/Total corpus size*

In table 3’s case, target word is *statistics* (n= 468,374). Frequency of *crime* and *probability* goes to 716,160 and 190,693. As iWeb corpus involves “95,000 websites and each websites has 145,000 words” (https://corpus.byu.edu/iweb), the total corpus size could be calculated as 13,775,000,000, thus:

1) *$f\ e$(crime…statistics) = (468374 * 716,160) / 13,775,000,000≈ 24.35*
2) *$f\ e$(probability…statistics) = (468374 * 190,693) /13,775,000,000≈ 6.48*
3) *MI (crime…statistics) = fo / $f\ e$ = 6,619 /24.35 = 271.83*
4) *MI (probability…statistics) = fo / $f\ e$ = 4,601 / 6.48 = 710.03*

Apparently, the value of MI (probability…statistics) is larger than MI (crime…statistics), revealing that due to *crime* collocates more often with *statistics* than *probability*, *crime* is less strongly associated with *statistics* than *probability*.

Then, another interesting question arises: do the case above is just a coincidence? More specifically, is there any relevance between the frequency and strength (MI value) of a collocate? To cope with this question, this paper firstly presents all the MI value of the top 50 collocates of *statistics* based on the data on iWeb corpus, as table 4 displays. For simple calculation, the MI value in table 4 has been log-processed.

Table 4. The top 50 collocates of *statistics* in iWeb corpus

Rank	Collocates	Freq.	MI	Rank	Collocates	Freq.	MI
1	bureau	35824	9.03	26	facts	2857	4.08
2	labour	26600	7.25	27	population	2496	3.04
3	show	17448	3.6	28	applied	2449	3.16
4	national	17068	4.08	29	justice	2344	3.27
5	according	12802	4.02	30	indicate	2341	4.2
6	Canada	10672	4.62	31	figures	2217	3.77
7	U.S.	8959	3.68	32	economics	2160	4.93
8	office	8490	3.28	33	math	2105	3.91
9	statistics	7701	5.95	34	detailed	2099	3.36
10	crime	6619	5.12	35	survey	2034	3.19
11	vital	5631	5.46	36	census	2015	5.71
12	Australian	5630	4.79	37	showed	1951	3.08
13	mathematics	5614	6.4	38	trends	1674	3.71
14	reports	4987	3.78	39	compiled	1533	5.14
15	official	4922	3.95	40	inferential	1494	10.54
16	probability	4601	6.5	41	ONS	1487	8.27
17	descriptive	4318	8.01	42	reporting	1475	3
18	latest	4106	3.36	43	ABS	1451	5.51
19	BLS	4067	9.35	44	analytics	1439	3.76
20	usage	4000	4.92	45	collect	1403	3.16
21	analysis	3950	3.46	46	reveal	1395	3.81
22	released	3885	3.31	47	Zealand	1316	3.3
23	summary	3473	4.62	48	aggregate	1278	5.19
24	employment	3043	3.86	49	calculus	1252	6.65
25	reported	3024	3.15	50	collected	1242	3.28

Based on the data above, then produce a scatter diagram with the aid of SPSS, a scatter diagram is shown in Figure 4.

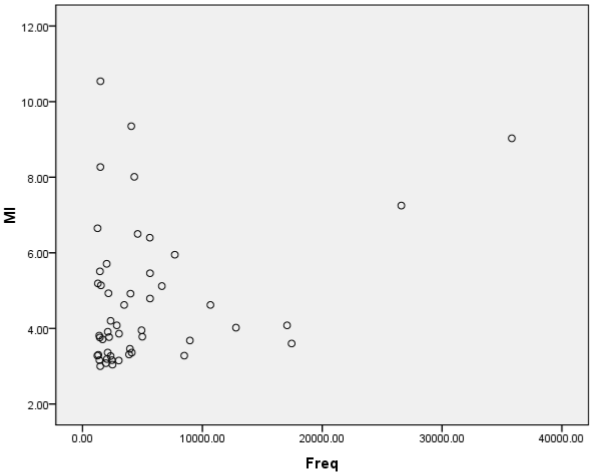


Figure 4. A Scatter Diagram of Freq. and MI

According to Figure 4, the scatter dots are out of order, from which we know that the frequency of a collocate lacks of relevance with its MI value, i.e. the strength of a collocate.

Nonetheless, "MI highlight rare exclusivity of the collocational relationship, favoring collocates which occur almost exclusively in the company of the node, even though this may be only once or twice in the entire corpus" (Brezina 2018: 70). Therefore other calculations have been suggested to take the frequency of collocates into account, e.g. the z-score (Berry-Rogghe 1973), log-likelihood (Dunning 1993), MI3 (Oakes 1998) and log-log (Kilgarriff & Tugwell 2001).

5. Conclusion

So far we have studied the collocation of *statistics* in iWeb corpus. A schema of *statistics* in iWeb corpus has been summarized in Figure 5.

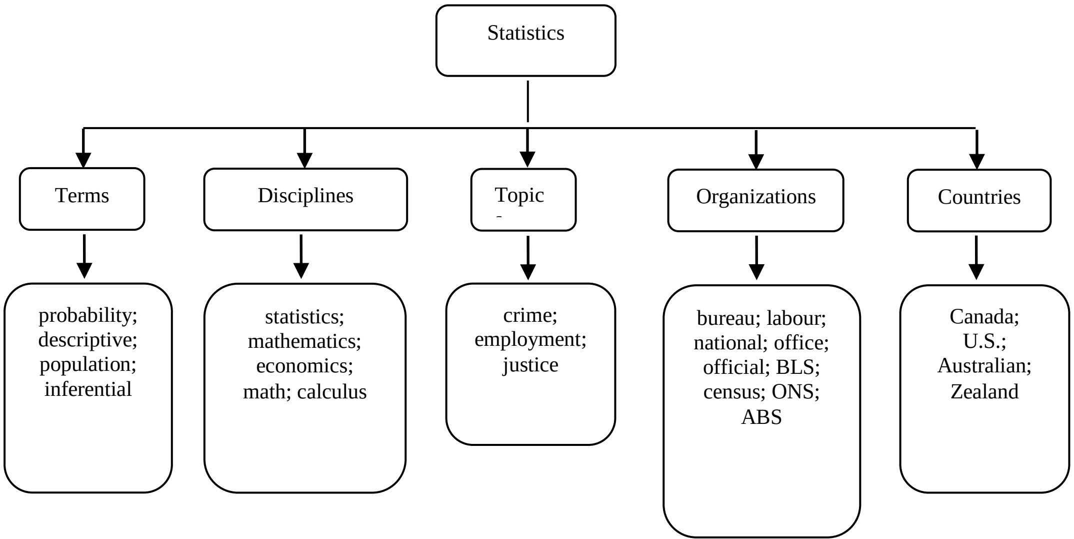

Statistics
Terms
Disciplines
Topic
Organizations
Countries
probability; descriptive; population; inferential
statistics; mathematics; economics; math; calculus
crime; employment; justice
bureau; labour; national; office; official; BLS; census; ONS; ABS
Canada; U.S.; Australian; Zealand

Figure 5. A Schema of *Statistics* in iWeb Corpus

In the category of Terms, *statistics* usually collocates with probability, descriptive, inferential and population. In the category of Disciplines, four subjects including statistics, mathematics/math, economics and calculus frequently collocate with statistics. Crime, employment and justice are the three topics that mostly concerned in iWeb corpus. In the category of Organizations, ABS, BLS, Statistics NZ are the three main departments. In the category of Spots, Canada, USA, Australia and New Zealand are the four countries frequently collocate with statistics. Moreover, the frequency of a collocate lacks of relevance with its MI value, i.e. the strength of a collocate.

Acknowledgement: This study is a part of the research project “Translating Metaphors in *the Wealth of Nations*: A Corpus-based Study (18GWCXXM－28)” funded by Guangdong University of Foreign Studies.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Berry-Rogghe, G. L. E. 1973. 'The computation of collocations and their relevance in lexical studies', in A. J. Aitken, R. Bailey and N. Hamilton-Smith (eds), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Biber et al. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Chen, J. S. & Lin, T. T. 2010. Explore the Colligations and Collocations of the high-frequency word *good* in the EFL Learner Essay Corpus.

Journal of Tianjin Foreign Studies University (1), 10-15.

Davies, M. 2017. *iWeb Corpus*. https://corpus.byu.edu/iweb/

Dodge, Y. 2006. *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74.

Firth, J. 1957. *Papers in linguistics*. Oxford: Oxford University Press.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Groebner et al. 2009. *Business Statistics: A Decision-Making Approach*. New York: Prentice Hall.

Kilgarriff, A. & Tugwell, D. 2011. WASP-Bench: an MT Leicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. *Proceedings of MT Summit VII*, 187–90.

Laufer, B. & Waldman, T. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2), 647–672.

Lei, L. & D. L. Liu. 2018. The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics* 23(2), 216–243.

Li, W. Z. 2017. A Study of Defining Approaches to Collocation, Measurement of Collocations and Collocation in Corpora of Chinese learners. *Foreign Language Education* (2): 70-74.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2), 223–242.

Oakes, M. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Siyanova, A. & Schmitt, N. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review* 64(3), 429–458.

Stigler, S. 1986.*The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge: The Belknap Press of Harvard University Press.

Yamashita, J. & Jiang, N. 2008. L1 Influence on the Acquisition of L2 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *Tesol Quarterly* 44(4), 647–668.