

Computing Thresholds of Linguistic Saliency

Siaw-Fong Chung* Kathleen Ahrens* Chung-Ping Cheng**
Chu-Ren Huang# Petr Šimon##

*National Taiwan University, **National Chengchi University, #Academia Sinica,

##Academia Sinica and National Tsing-Hua University

Abstract

Sketch Engine (Kilgarriff and Tugwell, 2001) provides a quantifier value called ‘saliency.’ Saliency is robust and powerful in displaying collocations in descending importance as well as according to grammatical relations. However, the system is still unable to determine (like a human) the line where any list members above that is salient and below not salient linguistically. We propose and test several computational methods to automatic determination of linguistic saliency. The findings of this proposal will not only contribute to the building of lexical resources but will also contribute to linguistic analysis based on empirical data where results are usually presented in the form of a listing in descending order of importance. In addition, this proposal also suggests modification to the present corpora interface so that the presentation of results can be made more meaningful to the users by identifying list members that are salient linguistically.

1.0 Introduction¹

All lexical resources, at the point of their design, will take into consideration whether the resources are useful to a target group. For example, WordNet (Fellbaum, 1998) was originally designed for the use of psychologists but they were later used extensively by computational linguists. Similarly, corpora such as British National Corpus (BNC), the Academia Sinica Corpus of Mandarin Chinese (Chen et al., 1996) and the Gigaword corpus were also designed for the use of target groups such as lexicographers, linguists, language teachers, language learners, etc. These corpora usually provide some forms of statistical analyses so that users will be able to summarize their research results quickly. For example, many corpora provide collocational measures such as Mutual Information values (Church and Hanks, 1989) so that collocated words can be sorted according to their significance of co-occurrences. To date, the Sketch Engine (Kilgarriff and Tugwell, 2001) is a powerful resource which displays search summary in collocated patterns as well as according to grammatical relations. However, like many other resources, the Sketch Engine is unable to determine which of the results in the list are salient linguistically.

Therefore, when provided with search summary in lists of collocation, most linguists report the top “few,” based on their preferences. Some linguists report the top one or two and keep the rest in appendixes. In fact, the current search summary from corpora or lexical resources does not give enough information regarding which of the top words are significantly different from the bottom words. In this paper, a research question is asked, i.e., whether or not one can select top rankings from linguistic results using principled measures. This selection of top rankings is useful because it will provide an automatic identification of significant linguistic results from listings of linguistic data. This also involves deciding the significant results

¹ We would like to thank Professor Shu-Chuan Tseng for her comments on this paper.

which are likely to be most prototypically used in a certain linguistic environment (Rosch and Mervis, 1975). In this paper, we propose three methods in which threshold of linguistic saliency can be extracted.

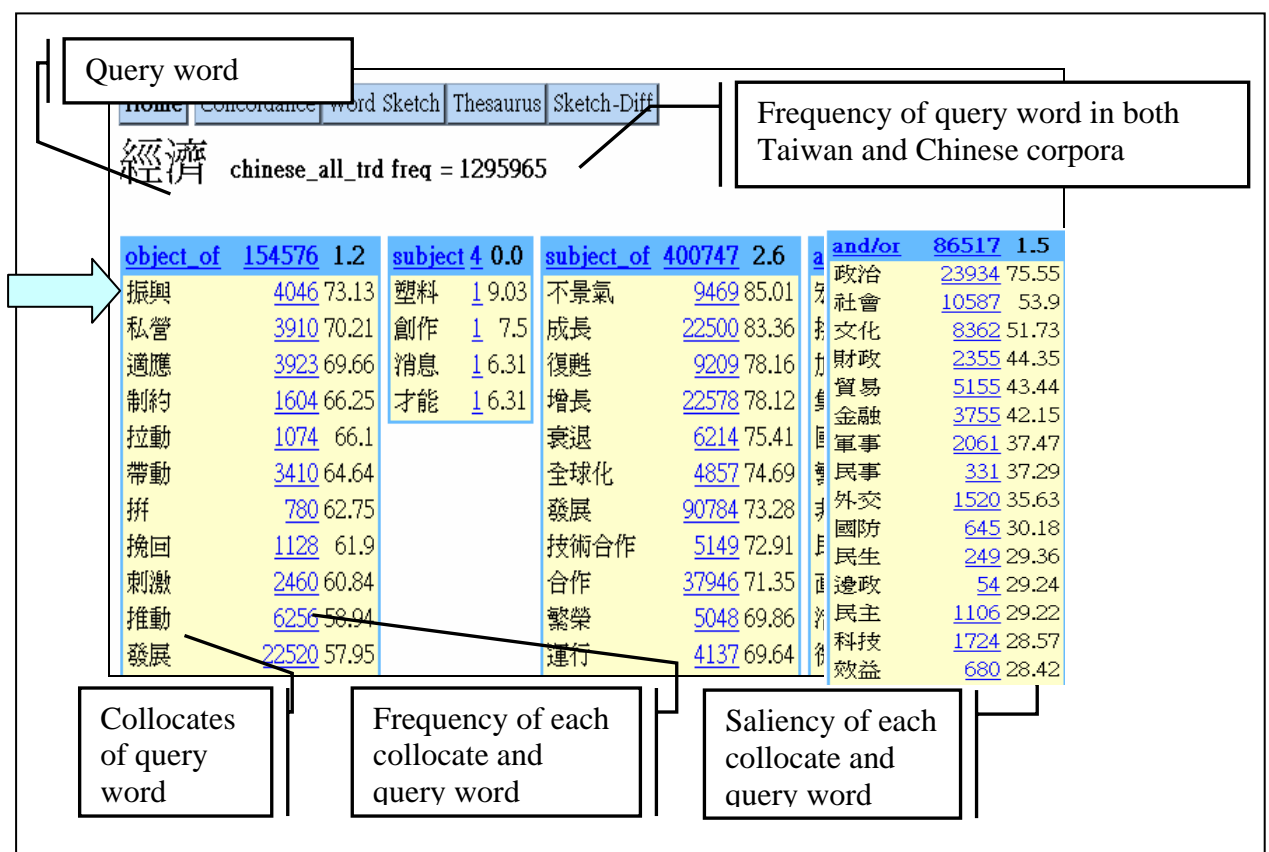
In the following section, data presentation in exemplified corpora and lexical resources is first discussed.

2.0 Data Presentation in the Sketch Engine

Sketch Engine is a system that provides the collocations of words according to grammatical relations. It has been used to analyze large scale corpora data such as British National Corpus (BNC) and the Chinese Gigaword corpus. The Chinese Sketch Engine was created by Kilgarriff, Huang, Rychly et al. (2005) and it has the same function as the English Sketch Engine, which also arranges collocates for Chinese query words in grammatical relations. For example, when a query word searched in Sketch Engine, the system will return with the collocates for this query word which are arranged in grammatical relations such as ‘objects of the query word,’ ‘subjects of the query word,’ ‘modifiers of the query word,’ etc.

The following Figure 1 shows an example of the search result for 經濟 *jing1ji4* ‘economy’ in the Chinese Sketch Engine.

Figure 1: Collocates for the Query Word 經濟 *jing1ji4* ‘Economy’ in the Chinese Sketch Engine



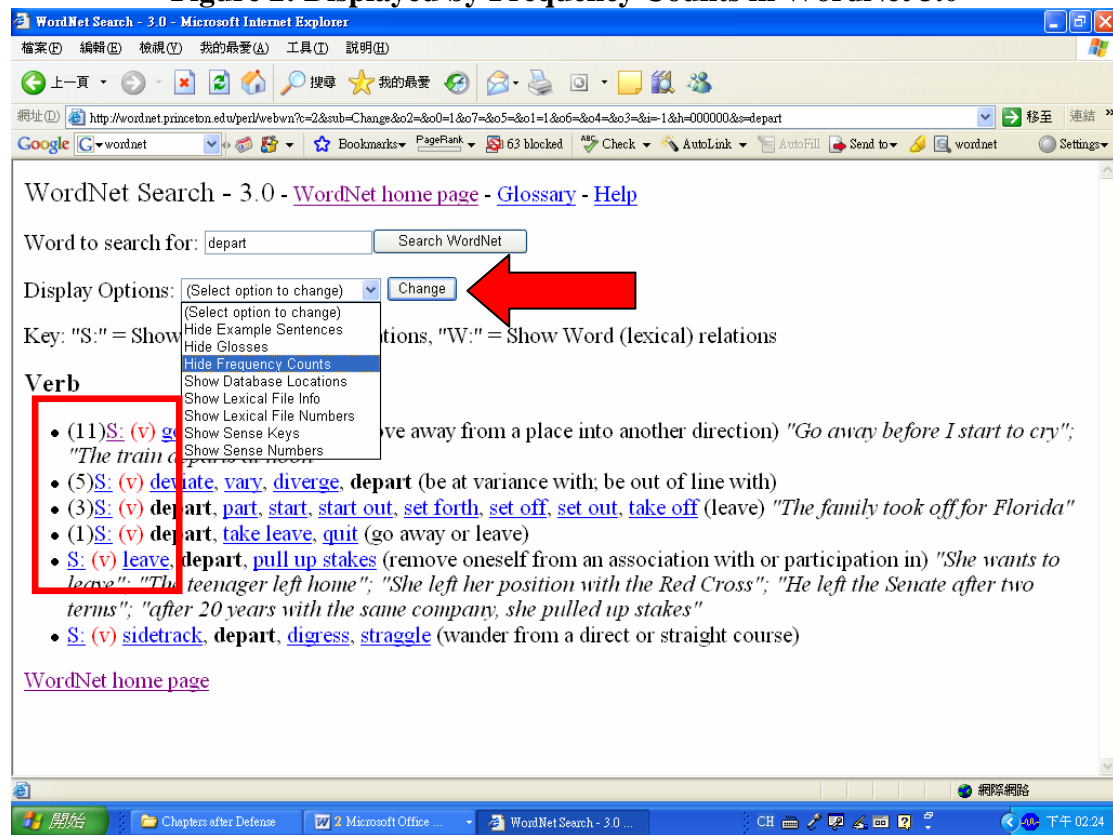
In Figure 1, the query word and its frequency in the entire Gigaword corpus are shown. The frequency for pair of collocates such as 經濟 *jing1ji4* ‘economy’ and 振興 *zheng4xing4* ‘to give life to’ under the ‘object-of’ relation (arrow in Figure 1) is given. In this case, it is 4,046 (in the second column for each relation), indicating that

經濟 *jing1ji4* ‘economy’ appears as the ‘object of’ the verb 振興 *zheng4xing4* ‘to give life to’ 4046 times in the whole Gigaword corpus.

In addition to frequency, Sketch Engine provides an additional score for the ranking of saliency of collocates. This is because Kilgarriff and Tugwell (2001) suggest that frequency alone may not be a reliable score because frequency of the collocates are relative to the number of both words in the whole corpus. Therefore, they suggest using a more reliable account to standardize all frequencies for the collocated based on the overall performance of the collocates in a particular condition.

The presentation of saliency in Sketch Engine is robust and useful. However, it does not inform which of the collocates in each relation are linguistically salient. Similarly, when we look at the arrangement of senses in WordNet (available online at <http://wordnet.princeton.edu/>), we see results shown in Figure 2 below.

Figure 2: Displayed by Frequency Counts in WordNet 3.0



WordNet can display the search results based on “high frequency count” (see Figure 2). This frequency count is the ordering of the most frequent sense to the least frequent sense (Tengi, 1999) that is computed using a semantic concordance created by Landes, Leacock and Tengi (1999) based on two corpora – the Brown corpus and Stephen Crane’s novella entitled *The Red Badge of Courage*.²

From Figure 2, one can see that the sense frequencies for ‘depart’ are 11, 5, 3 and 1. We can see that there is a bigger gap between the frequency of the first sense (11) and the frequency of the second sense (5). Based on this gap, we may say that the first sense is more often used than the second one. It is also possible to say that the first sense is more prototypical than the other senses. Therefore, there is possibly a threshold after the first sense to make the first sense more distinctive in use than the

² Only senses that were found in the two corpora can be shown their frequency counts in brackets.

others. Therefore, this paper suggests that there should be some objective methods which can help determine the threshold of linguistic saliency as such. This paper suggests three methods to find out how many of the top few results are considered significant. These methods are elaborated below.

3.0 Methods One and Two

Methods One and Two are based on the characteristics of the distributional listings, which usually follow Zipf's law (Zipf, 1932). Zipf's law states that the most frequent value is most likely to be twice as much as the second most frequent value. For example, when a sample size is large enough, the result of a frequency listing is likely to be in a distributional pattern. For instance, for the metaphorical expression of 起飛 *qi3fei1* 'takeoff' in (1) below, its collocates from the Sketch Engine are presented in Figure 3 below.

- (1) 但 在 台灣 經濟 起飛 後 (Central News Agency of Taiwan)
dan4 zai4 tai2wan1 jing1ji4 qi3fei2 hou4
but at Taiwan economy takeoff after
“But after the economy of Taiwan takeoffs...”

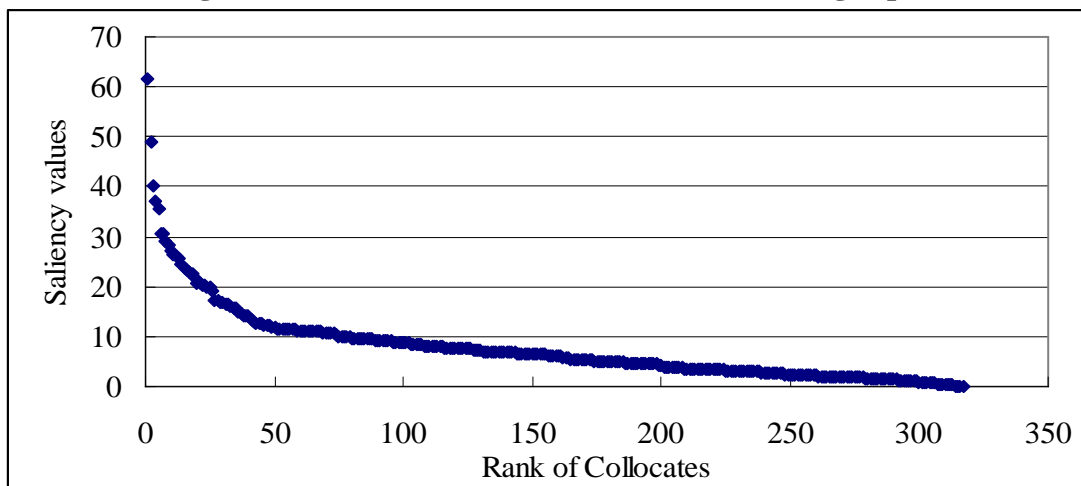
The collocates for 起飛 *qi3fei1* 'takeoff' which appears at the similar grammatical relation with 經濟 *jing1ji4* 'economy' (the 'subject' relation) can be seen in Figure 3 (such as 飛機 *fei1ji1* 'airplane,' 班機 *ban1ji1* 'flight,' 跑道 *pao3dao4* 'path' as well as 經濟 *jing1ji4* 'economy'). We can see that in Figure 3, the saliency values of the collocates are arranged in descending order (from 55.67, 48.31, 38.64...until the lowest value which will be zero).

Figure 3: Collocates of ‘Subjects’ of 起飛 *qi3feil* ‘takeoff’ in the CNA in the Sketch Engine

起飛		chinese_all_trd:taiwan-only freq = 16705	
subject	2208	19.6	
飛機	514	55.67	
班機	225	48.31	
跑道	70	38.64	
經濟	576	31.78	
夢想	27	30.68	
客機	51	30.4	
滑行道	7	28.31	
航機	14	25.19	
專機	25	24.33	
航空母艦	15	22.34	
小時	32	22.24	
軍機	15	21.75	
戰機	25	21.4	
包機	14	21.22	
直昇機	17	20.6	
航艦	8	20.37	
班次	13	20.07	
航班	14	18.98	
直升機	15	18.38	
協和機	3	17.38	
基督城	4	17.13	
運輸機	8	17.11	
志航基地	3	16.29	
甲板	5	15.54	
機	14	14.79	
貨機	5	14.74	
佳山基地	2	14.71	
才能	27	14.35	
回程	4	14.05	

Most frequency list follows the pattern of the Zipf’s law, where the top few are usually very high and the values will decrease until a state where changes become minimum. For example, for the saliency list in Figure 3, when plotted in graph, the representation can be seen in Figure 4 below. In Figure 4, the x-axis is the ‘Chinese subject’ and the y-axis is the ‘saliency’ (Figure 2 uses the rank of the Chinese word to represent the Chinese character – rank 1, 2, 3...). All these Chinese words are the collocates of 起飛 *qi3feil* ‘takeoff’.

Figure 4: Pattern of Distributional Data following Zipf's Law

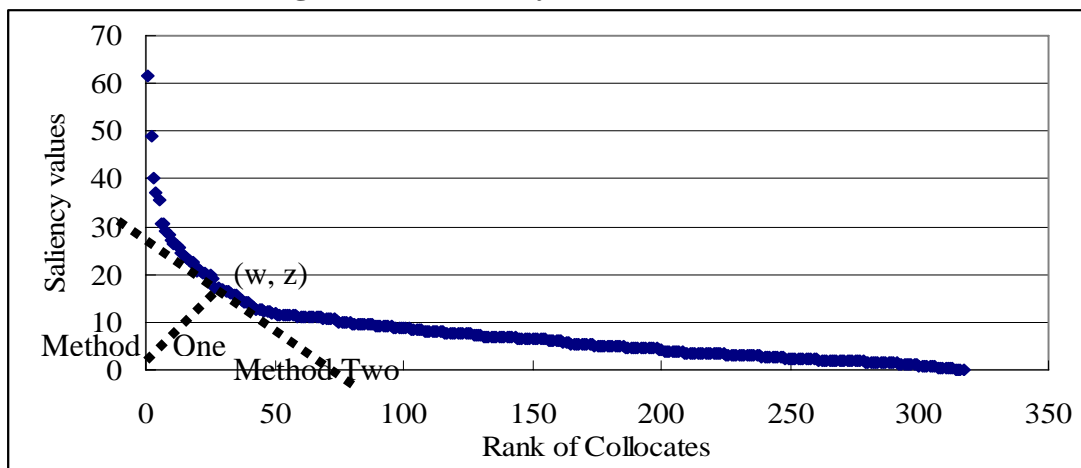


The function for the type of graph in Figure 4 is such that in (2), where any point in the graph will be $(x, f(x))$. x is the rank of Chinese subjects on the x-axis and $f(x)$ is the function to calculate the value on the y-axis.

$$(2) f(x) = b(x^a)$$

Using this formula, Methods One and Two will find a point that separates any distributional listing into two lists, i.e., significant and insignificant lists. The purpose of doing this is to find out which among the list should be considered significant and which to be insignificant.

Figure 5: Three Ways to find Threshold Values



Methods One and Two are based on the assumption that there is a point where the curve changes the most when it goes down the y-axis to the x-axis. Method One calculates the position of (w, z) where it is of shortest distance from $(0, 0)$. This is because when every line departs from the starting point of $(0, 0)$, there will be a line that is the shortest distance from the curve. The point where this line touches the curve is the point where the curve changes the most from the y-axis to the x-axis.

Method Two calculates the most slanted slope between the x-axis and the y-axis. When the slope is most slanted, the possibility is high that the curve changes the most

at a certain point (w, z). This is because the higher the curve on the y-axis, the more vertical the slope will be. Moreover, the further the curve moves away from (0, 0) on the x-axis, the more horizontal the slope will be. Therefore, the most slanted slope between the vertical and horizontal will be the possible threshold representing where the curve has changed the most.

The formula for the two methods are shown in (3) below. In these two formula, a and b are the variables in the function of the nonlinear regression $y = b(x^a)$ while i is the threshold value and n is the total number of collocates in the relation.

$$(3) \text{ Method One: } i = \left[((-ab^2)^{\frac{1}{2-2a}}) \right]$$

$$\text{Method Two: } i = \left[(-ab)^{\frac{1}{1-a}} \right]$$

These are the introduction to Methods One and Two, Method Three is elaborated below.

4.0 Method Three

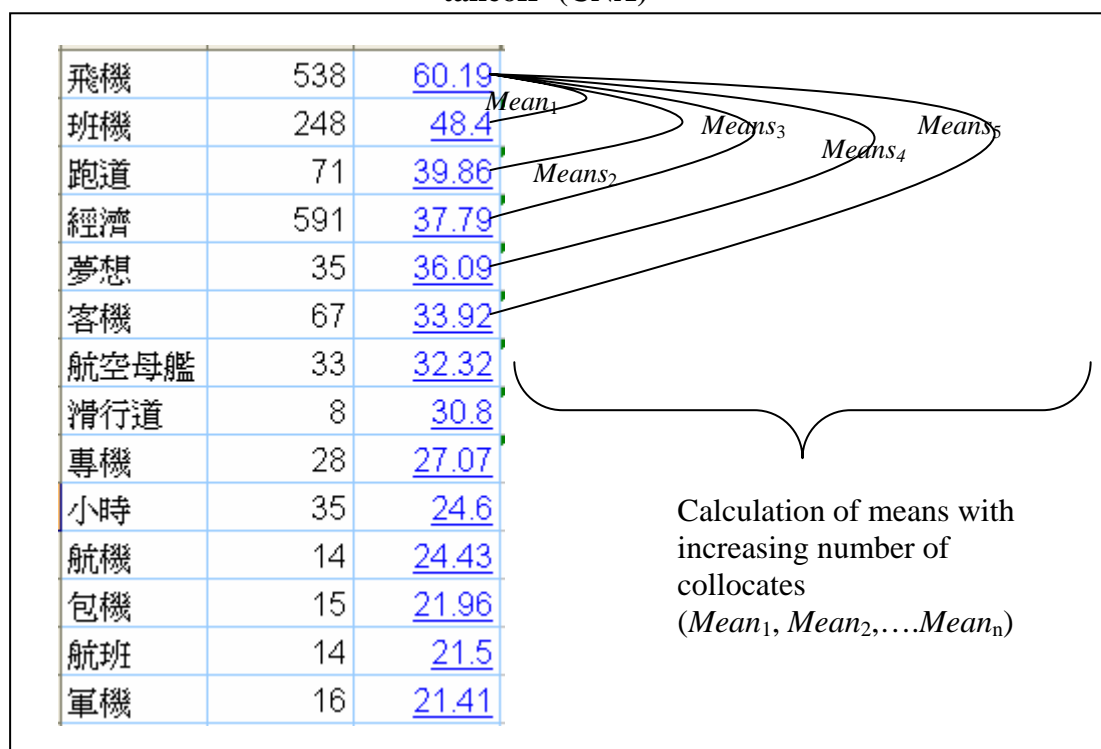
Method Three is called ‘mean of means’ where series of means will be calculated. For example, for the saliency list in Figure 3, the first mean is the mean of collocates one (55.67) and two (43.81); the second mean is the mean of collocates one (55.67), two (43.81), and three (38.64), i.e., add a new collocate every time. When all means have been calculated for all collocates, an overall mean is obtained from all the means (thus, mean of means). This overall mean will be used as a threshold value for the cut line, formulated below.

(4) Threshold=

$$\frac{Mean_1(Saliency_1, Saliency_2) + Mean_2(Saliency_1, Saliency_2, Saliency_3) + \dots + Mean_n(Saliency_{(n-2)}, Saliency_{(n-1)}, Saliency_n)}{n-1}$$

When all these means are laid out, the cutting line is above the threshold value. The computation of mean of means is shown in Figure 6 below.

Figure 6: Computing ‘Means’ for the Collocates of ‘Subjects’ of 起飛 *qi3fei1* ‘takeoff’ (CNA)



From Figure 6, we can see that a series of means is produced by increasing the number of collocates each time in the calculation. In the following section, we will discuss the overall results for the three methods.

5.0 Results

For both Methods One and Two, normalization is used because the ranking in the x-axis (1,2,3...) is not comparable to the y-axis (between 0 to about 50).³ The results for Methods One and Two are shown in Table 1 below for three metaphorical expressions, i.e., 成長 *cheng2zhang3* ‘grow/growth,’ 起飛 *qi3fei1* ‘takeoff’ and 癱瘓 *tan1huan4* ‘paralytic.’ In this table, the first column shows the metaphorical expressions, followed by the total collocates each grammatical relation possesses. “Pseudo-R-square” in column four shows the percentages of the curve that fit the non-linear regression (or in colloquial term, “curve fitting”). For example, the first relation

$$^3 (6) \text{ Axis-y: } \frac{\text{Collocate}_{\text{Rank } 1 \dots \text{Rank } n}}{\text{Rank}_n}$$

$$\text{Axis-x: } \frac{\text{Saliency}_{1 \dots n}}{\text{Sum}(\text{Saliency}_1, \text{Saliency}_2, \dots, \text{Saliency}_n)}$$

For the axis-y (saliency values), each collocate from rank 1 to *n* will be divided by rank from highest to lowest. For example, if a Chinese word has 200 collocates in a particular relation, the normalization will divide collocates ranked 1 to 200 with 200 (thus, $\frac{1}{200}, \frac{2}{200}, \dots, \frac{200}{200}$). Therefore, the output of the

axis-y is a list of numbers ranging from 0 to 1. As for axis-x, each saliency value will be divided by the sum of all 200 saliency values. The output of the axis-x is also displayed on a scale ranging from 0 to 1 (which is also the percentage of the saliency values).

(subject) of 成長 *cheng2zhang3* ‘grow/growth’ shows a “curve fitting” of 91%. The results for Methods One and Two are given in columns four and five.

Table 1: Calculation of Threshold Values Using Methods One and Two (CNA)

‘Types of Metaphorical Expressions’	Relations	Total Collocates	Pseudo-R-square	Method One	Method Two
成長 <i>cheng2zhang3</i> ‘grow/growth’	Subject	1490	0.906935	5.472613	4.211427
起飛 <i>qi3fei1</i> ‘takeoff’	Subject	268	0.933048	3.630461	2.926560
癱瘓 <i>tan1huan4</i> ‘paralytic’	Subject	276	0.935357	4.384251	3.748123
	Modifies	221	0.967868	3.787687	3.173571

The ‘subject’ relation of 成長 *cheng2zhang3* ‘grow/growth’ shows to have threshold values above collocate number 5 in Method One and collocate number 4 in Method Two. Similar results can be seen in the examples of 起飛 *qi3fei1* ‘takeoff’ and 癱瘓 *tan1huan4* ‘paralytic’ in Table 1 above.

As for Method Three, the results are shown in Table 2 below. The threshold is marked by a dotted line across the table after collocate number 97. This method locates the cut-off collocate at number 89, roughly one third down, from a total 268 collocates. The threshold value is given as ‘mean of means’ at the bottom, which is the mean value for all the means in the last column.

Table 2: Mean of Means: ‘Subject’ 起飛 *qi3fei1* ‘Takeoff’ (CNA)⁴

Collocate Number	Chinese Collocates	English Gloss	Frequency	Saliency	Means
1	飛機 <i>fei1ji1</i>	airplane	538	60.19	---
2	班機 <i>ban1ji1</i>	airliner	248	48.40	54.30
3	跑道 <i>pao3dao4</i>	runway	71	39.86	49.48
4	經濟 <i>jing1ji4</i>	economy	591	37.79	46.56
5	夢想 <i>meng4xiang3</i>	dream	35	36.09	44.47
6	客機 <i>ke4ji1</i>	passenger plane	67	33.92	42.71
7	航空母艦 <i>hang2kung1 mu3jian4</i>	aircraft carrier	33	32.32	41.22
8	滑行道 <i>hua2xing2dao4</i>	taxiway	8	30.8	39.92
9	專機 <i>zhuan1ji1</i>	special plane	28	27.07	38.49
10	小時 <i>xiao3shi2</i>	hour	35	24.6	37.10
11	航機 <i>hang2ji1</i>	flight	14	24.43	35.95
12	包機 <i>bao1ji1</i>	charter plane	15	21.96	34.79
13	航班 <i>hang2ban1</i>	flight	14	21.5	33.76
14	軍機 <i>jun1ji1</i>	military plane	16	21.41	32.88
15	戰機 <i>zhan4ji1</i>	fighter plane	26	21.19	32.10
16	直昇機 <i>zhi2shen1ji1</i>	helicopter	18	20.41	31.37
17	班次 <i>ban1ci4</i>	flight order	13	19.93	27.85
.....
87	駕駛員 <i>jia4shi3yuan2</i>	driver	3	7.94	15.28
88	特號 <i>te4hao4</i>	special umber	1	7.85	15.20
89	秋門 <i>ciu1men2</i>	a state in Siberia	1	7.83	15.11
90	產業 <i>chan3ye4</i>	Industry	15	7.82	15.03
91	雙機 <i>shuang1ji1</i>	dual machines	1	7.78	14.95
92	爸爸節 <i>ba1ba1jie2</i>	father’s day	1	7.66	14.87
.....
.....
267	能力 <i>neng2li4</i>	capability	1	0.04	7.45
268	目標 <i>mu4biao1</i>	goal	1	0.03	7.42
Mean of Means (Threshold)					15.03

Therefore, from the results, we can see that three different methods provide different threshold values. These methods are useful depending on the purpose of the research. For example, Methods One and Two can be applied to calculating smaller sampling of thresholds (about top 1 to 6) but Method Three allows the calculation of larger sampling of thresholds. For different purposes of linguistic research, these three methods provide choices as to how to select top results using principled methodology.

⁴ A small number of words in Sketch Engine are wrongly tagged. For example, 秋門 *ciu1men2* is a location where the airplane takeoffs but it is wrongly tagged. These errors are due to the problems of Sketch Engine but they will be removed automatically during clustering because they may not fall in any clusters within the list of collocates.

6.0 Conclusion

The above shows three methods which can help linguists to work further with their empirical linguistic data which are of distributional pattern. The reason why this proposal emphasizes finding significant list is because most empirical studies do not know where to stop listing results from listings such as frequency and saliency lists. Most studies tend to list the top few and the number of the top few depends on the choice of the researchers. If there are criterion-based methods to find out the thresholds for the linguistic listings, subjectivity will be reduced in terms of choosing which top few words to be selected. Furthermore, most lexical resources provide wordlists according to different criteria such as frequency, Mutual Information values, collocation, saliency values, etc. None has suggested which of the top few listed should be looked at. This paper, therefore, deals with the general problems of these listings and suggests three possible ways to solve the problem. Future work suggests incorporation of the calculation of threshold values in lexical resources such as Sinica Corpus, the English and Chinese Sketch Engine, etc. This proposed idea should have great contribution to computational linguists, researchers needing statistical ways to analyze linguistic data, and researchers who need to run psycholinguistic experiments related to word meaning. Currently, these three methods are being evaluated objectively. By doing so, we are not only able to validate the results of computational approach using human judgment; we can also evaluate which of the methods perform better in matching human's evaluation.

References

- Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. "Sinica Corpus: Design Methodology for Balanced Corpora." In B.-S. Park and J.B. Kim (Eds). *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*. Seoul:Kyung Hee University. pp.167-176.
- Church, Kenneth W. and Patrick Hanks. 1989. "Word Association Norms, Mutual Information and Lexicography." In the *Proceedings of the 27th Annual Meeting of ACL*, Vancouver. pp. 76-83
- Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Kilgarriff, Adam and David Tugwell. 2001. "WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography." In the *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse, July: 32-38.
- Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, David Tugwell. 2005. Chinese word sketches. In the *Proceedings of Asialex*, Singapore.
- Landes, Shari, Claudia Leacock, and Randee I . Tengi. 1999. "Building Semantic Concordance." In Christiane Fellbaum. (Ed.). *WordNet: An Electronic Lexical Database*. MIT: Cambridge, Mass. and London, England. pp. 199-216.
- Rosch, Eleanor and Caroline B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology*, 7. pp. 573-605.
- Tengi, Randee I. 1999. "Design and Implementation of the WordNet Lexical Database and Searching Software." In Christiane Fellbaum. (Ed.). *WordNet: An Electronic Lexical Database*. MIT: Cambridge, Mass. and London, England. pp. 105-127.
- Zipf, George Kingsley. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge (Mass.).

[3,086 words]