A New Score to Characterise Collocations: Log-r in Comparison to Mutual Information

Itsuko Fujimura
Nagoya University (Japan)
fujimura@nagoya-u.jp

Shigenobu Aoki
Gunma University (Japan)
aoki@si.gunma-u.ac.jp

Keywords: Big Data, Association Measures, Data Visualisation, Typology of collocations, Zipf's law

Abstract

This paper proposes a new score named Log-r as a simple measure for calculating the strength of association between the constituent words of bigrams and argues that Log-r is more appropriate than Mutual Information for characterising collocation types. Arguments are based on the visualisation of one million English bigrams taken from a corpus of 1.1 billion words and of 0.4 million French bigrams taken from a corpus of 0.1 billion words. A three-dimensional analysis of each bigram will be made with its Log-r, its logarithmized frequency, and vocabulary level of its constituent words. Transparent typological study of collocations can be conducted using this procedure, which is only based on the frequency of words and bigrams, Pearson's *r* and Zipf's law.

1. INTRODUCTION

Linguistic research on collocations based on large-scale corpora is flourishing. A collocation is a string of two or more words that frequently co-occur.

There are various types of collocations, and there is a need to categorise and describe the characteristics of each. Currently, however, there is no generally accepted typological framework. Terms such as 'compound word', 'phrase', 'fixed expression', 'collocation', 'idiom', 'lexical bundle', and 'multi-word unit' are used without being given distinctive definitions.

The frequency of co-occurrence and the degree of association between the constitutive words have been recognised as basic properties characterising collocation types (Ellis, 2012; Evert, 2009; Wray, 2012). While frequency is easy to understand and measure, strength of association is more complex; it is referred to using different names ('degree of association', 'degree of compositionality', 'degree of fixedness', 'degree of coherence', etc.) and is not measured in a unified fashion. Furthermore, while Mutual Information (MI) is often used as a method to measure strength of association, various other methods have been proposed (Pecina, 2010), and research on this topic is still developing (Bybee, 2010; Evert, 2009; Gries, 2013).

This paper proposes a new score named Log-r as a simple measure for calculating the strength of association between the constituent elements of two-word collocations and argues that Log-r is more appropriate than MI for describing collocation types.

For the sake of simplicity, this paper addresses only bigrams –sequences of two words. A collocation is a bigram in which the words are more or less habitually associated; it includes all of the aforementioned terms ('compound word', 'lexical bundle', 'idiom', etc.).

1.1. A Frequency- and Strength-based Typological Model: Wray (2012)

There are various types of word strings that fall under the above-mentioned definition of 'collocation'. Wray (2012: 241) proposes the diagram in Figure 1 as a model for comprehensively expressing some of these.

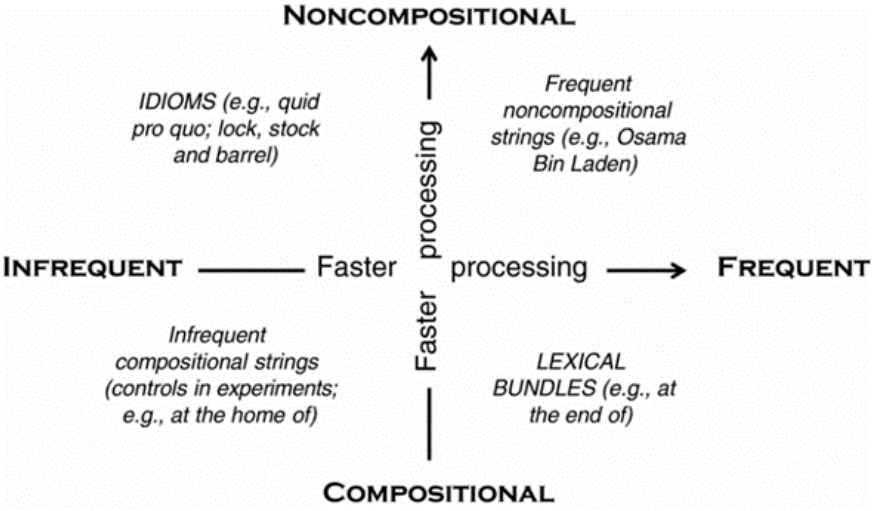

Fig 1: Wray's (2012) Typological Model of Collocations

The vertical axis in Figure 1 represents the level of compositionality, and the horizontal axis, the frequency. The compositionality is the same concept as the strength of association between constitutive elements. An exemplary type of word sequences is shown in each quadrant. Word sequences that often function as single words and appear frequently (e.g. *Osama Bin Laden*) are positioned in the top-right, first quadrant; word sequences that often function as single words but appear infrequently are in the top-left, second quadrant (idioms, e.g. *quid pro quo*); word sequences that generally do not

function as single words, appear infrequently, and cannot be called collocations are in the bottom-left, third quadrant (e.g. *at the home of*); and word sequences that appear frequently but generally do not function as single words are in the bottom-right, fourth quadrant (lexical bundles, e.g. *at the end of*).

Examples of bigrams of each exemplary type are as follows: Quadrant 1 – *White House*, *Hong Kong*; Quadrant 2 – *lingua franca, bovine spongiform*; Quadrant 3 – *pink roses, familiar enough*; Quadrant 4 – *I am, of the*.

1.2. Problems with the MI score

MI is one of the most frequently mentioned methods for measuring the strength of association between constituent elements. Church & Hanks (1990:23) propose MI as an 'association ratio', Ellis (2012:28) and Hunston (2002:71) introduces it as measure of the strength of association, and A Glossary of Corpus Linguistics describes it as follows:

> Mutual information: (…) In corpus linguistics it is often used as a measure of the strength of a collocation between two words. (Baker, Hardie, & McEnery, 2006: 120)

However, it is hard to say that MI can be relied upon as a measure of collocation strength. Firstly, based on English-language newspaper corpora introduced in Table 1 below, the MI value of *Hong Kong* (12.8) was lower than that of *Jacqueline Onassis* (13.9); also the MI value of *human rights* (11.0) was lower than that of *human societies* (12.3). Intuitively, one would think that the association between the two words in *Hong Kong* is stronger than that of those in *Jacqueline Onassis*, and that of those in *human rights* is stronger than that of those in *human societies*. Secondly, as the developers of MI have pointed out from the beginning, a problem with this measure is that low-frequency bigrams are overvalued (Church & Hanks, 1990:24) and it has become customary to exclude them from measurements. However, this feature has not yet been adequately explained. Thirdly, the fact that various measures are being proposed to calculate collocations (Pecina, 2010, Gries, 2012) and that there are considerable discussions around them (Evert, 2009; François & Manguin, 2006; Gries, 2013), indicates itself that MI does not satisfy the conditions for it to be admitted as an appropriate indicator of collocation strength.

In this study, we will propose the use of Log-r instead of MI to measure the strength of association between the constituent elements of bigrams. In terms of Wray's model (Figure 1 above), Log-r would be the vertical axis scale.

2. LOG-R PROPOSAL

2.1. Definition

As a measure expressing the strength of association between two words, we propose Log-r which is a common logarithm of the correlation coefficient *r* that expresses the attribute correlation of two variables (word x and word y). The Pearson's product-moment correlation coefficient is defined as:

$$r = \frac{cov_{xy}}{\sigma_x \sigma_y} \tag{1}$$

This study assumes a Poisson distribution and uses the approximation formula (2), where f_{xy} is the frequency of the successive words xy, and f_x and f_y are respectively the

frequency of word x and word y. A Poisson distribution can be assumed when large-scale data are used and the frequency of words xy is low.

$$r \fallingdotseq \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{2}$$

Log-r is therefore defined as:

$$\text{Log-r} = log_{10} \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{3}$$

2.2. Examples

Log-r's values are less than or equal to 0. Typical values and English and French examples of each are shown below.

- Log-r = 0, *r* = 1, e.g. lingua franca (en), statu quo (fr)

lingua franca: 100% of the occurrences of word x (*lingua*) are co-occurrent with 100% of the occurrences of word y (*franca*).

- Log-r = -1, *r* = 0.1, e.g. *apple pie* (en), *sud ouest* (fr)

apple pie: 10% of the occurrences of word x (*apple*) are co-occurrent with 10% of the occurrences of word y (*pie*).

- Log-r = -2, *r* = 0.01, e.g. *medal winner* (en), *gare SNCF* (fr)

medal winner: 1% of the occurrences of word x (*medal*) are co-occurrent with 1% of the occurrences of word y (*winner*).

- Log-r = -3, *r* = 0.001, e.g. *earlier offer* (en), *poids trop* (fr)

earlier offer: 0.1% of the occurrences of word x (*earlier*) are co-occurrent with 0.1% of the occurrences of word y (*offer*).

- Log-r = -4, *r* = 0.0001, e.g. *no there* (en), *pas il* (fr)

no there : 0.01% of the occurrences of word x (*no*) are co-occurrent with 0.01% of the occurrences of word y (*there*).

The Log-r value is 0 when two words are strongly associated and have come to function as a single word. Log-r is -4 when there is absolutely no habitual association between the two words. Between 0 and -4 there is a continuum of bigrams with different strengths of association.

2.3. Mathematical Characteristics of Log-r

The characteristics of Log-r can be summarised as follows. First, Log-r is a simple statistic, nothing more than a logarithm of Pearson's *r*. Like frequency of occurrence, it is robust and highly transparent. Second, Log-r is an appropriate statistic for linguistic phenomena. The frequency of words or bigrams is known to be an extremely wide-ranging statistic explained by Zipf's law (Baroni, 2009; Zpf, 1949). It is roughly calculated using the formula: 'occurrence rate (%) = 10/rank'. The most frequent word appears as 10% of the total number of words, the 10th most frequent word appears as 1% of the total, the 100th most frequent word appears as 0.1% of the total and so forth. Using a logarithm makes it easy to handle values and to grasp phenomena intuitively, by enabling us to visualise the phenomena, as in Wray's diagram shown in Figure 1. Third,

Log-r can measure all bigrams in a corpus, because the approximate value of r coming from formula (2) does not take a negative value, unlike the definitional value of r coming from formula (1). A logarithmic transformation is possible only for positive value. Fourth, Log-r is easy to calculate, because, compared to the definitional formula (1), the approximation formula (2) is simplified. However, there are restrictions when one uses this latter. Caution is necessary, as the difference between definitional formula-based values and approximation formula-based values grows greater as the value of f_{xy} increases and as the scale of the corpus decreases.

It should be finally noted that Log-r is not an entirely novel measure. Among the 82 measures introduced in Patina (2010), Pearson's chi-square test, z-score, Pearson, and Phi share fundamental characteristics with Log-r.

2.4. Comparison with MI

MI is frequently mentioned as a measure of a bigram's strength of association (Ellis, 2012; Evert, 2009; Gries, 2012; Hunston, 2002). The MI definitional formula is (4), and the MI approximation formula is (5). The latter is used in practice.

$$\mathrm{MI} = \log_2 \frac{P_{xy}}{P_x P_y} \tag{4}$$

$$\mathrm{MI} = \log_2 \frac{f_{xy}\, N}{f_x f_y} \tag{5}$$

The essential difference between the MI and Log-r calculation formulas is that in the former (5), the denominator is $f_x f_y$, whereas, in the latter (3), it is square rooted, $\sqrt{f_x f_y}$.

As the formulas clearly show, MI does not simply express the strength of association between x and y. Even if the total number of words in the corpus N is the same and the f_x: f_y: f_{xy} ratio is the same, the value of MI changes depending on the value of f_{xy}.[1] The greater f_{xy}, the smaller the value of MI, and the smaller f_{xy}, the greater the value of MI. On the other hand, in the case of Log-r, if the ratio f_x: f_y: f_{xy} is the same, the Log-r value stays the same regardless of the value of f_{xy}. We illustrate this clearly below on the basis of examples.

3. DATA

Details of the data used in this study can be found in Table 1. Our corpora were English- and French-language newspapers. From the main text of articles of these newspapers, we manually extracted 1.04 million English bigrams that appeared 54 times or more and 400,000 French bigrams that appeared 20 times or more. Then we calculated the occurrence frequency, Log-r, and MI of each. A morphological analysis was not carried out on the data; bigrams were presented in the form that they appeared in texts. We distinguished between upper and lowercase letters, but we did not take into account the presence or not of an apostrophe or hyphen between two words.

[1] The value of MI changes also depending on the size of the corpus, contra Hunston (2002 :73).

Table 1: Data

Language	No. of Bigrams	Total Number of Words	Newspaper	Corpus Distributor and Name
English	1.04 million (54 tokens or more)	1.1 billion words (Only main text of articles)	*L.A. Times-Washington Post* (1994-1998), *New York Times* (1994-1998), *Reuters Financial News* (1994-1996), *Reuters General News* (1994-1996), *Wall Street Journal* (1994-1996), *Associated Press Worldstream* (1994-1998)	LDC ・North American News Text Corpus ・North American News Text Supplement
French	400,000 (20 tokens or more)	118 million (Only main text of articles)	*Le Monde* (1988, 1994, 1996, 1999, 2000, 2006)	ELRA ・Le Monde

4. SCATTERPLOT-BASED COMPARISON OF LOG-R AND MI

4.1. One-dimensional Model

Table 2 shows the Log-r and MI of five English and French bigrams in our database. In these examples, Log-r and MI do not contradict each other in their assessment of the bigrams' strength of association.

Table 2: One-dimensional Display of Log-r and MI

Strength of association	English			French		
	bigram	Log-r	MI	bigram	Log-r	MI
+	*lingua franca*	-0.01	22.5	*statue quo*	-0.00	16.7
	apple pie	-1.01	13.8	*frère aîné*	-1.03	11.9
↕	*medal winner*	-2.00	8.0	*gare SNCF*	-2.04	8.1
	earlier offer	-3.00	2.3	*poids trop*	-3.01	2.3
-	*no there*	-4.04	-3.4	*pas il*	-4.01	-5.9

However, it is easy to find contradictions between the two measures. For example, in English, according to Log-r, *White House* (-0.23) is between *lingua franca* and *apple pie*, whereas according to MI, it is lower (11.1) than *apple pie*. Similarly, in French, according to Log-r, *sans doute* (-0.53) is between *statu quo* and *frère aîné*, whereas according to MI, it is lower (8.8) than *frère aîné*. In both instances, it is not easy to judge which measure is superior using a one-dimensional model.

4.2. Two-dimensional Model

A two-dimensional model brings into relief the differences between Log-r and MI. We placed Log-r and MI on the vertical axes of Figures 2 and 3, respectively, and put $\log(f_{xy})$ on the horizontal axis. We used logarithm for the frequency also, which allowed

us to create a diagram like that of Wray (2012) and to examine the phenomenon through a visual representation.

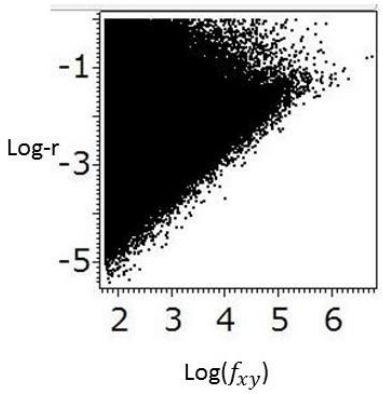

Fig 2: Log-r and Log(f_{xy})

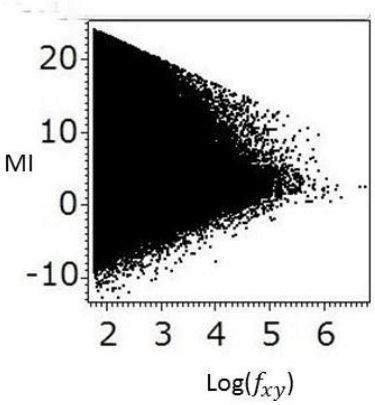

Fig 3: MI and Log(f_{xy})

It is clear that Figures 2 and 3 show variations of the same shape. In the Log-r-based Figure 2, the upper area extends horizontally. In the MI-based Figure 3, it descends toward the right, even though there is no reason to expect that bigrams with higher frequency will be less strongly associated and that bigrams with lower frequency will be more strongly associated. The MI formula results in an unnatural shape. Figure 3 is a transformation of Figure 2.

Although the bottom area of the scatter plot rises up to the right as frequency increases in both graphs, this is a natural increase that reflects actual language use. As the number of words in actual language use, as well as that in the corpus (sample) is limited, the coincidental co-occurrence of the bigrams' constituent elements decreases as their frequency increases (See Table 3 for details regarding this point.) In Figure 1, which we used as our model, the graph forms a square. However, in fact, the difference between *at the end of* in fourth quadrant and *at the home of* in third quadrant should be one of both frequency and strength of association.

4.3. Position of Examples

Next, we will analyse the bigrams positioned on the scatter plots.

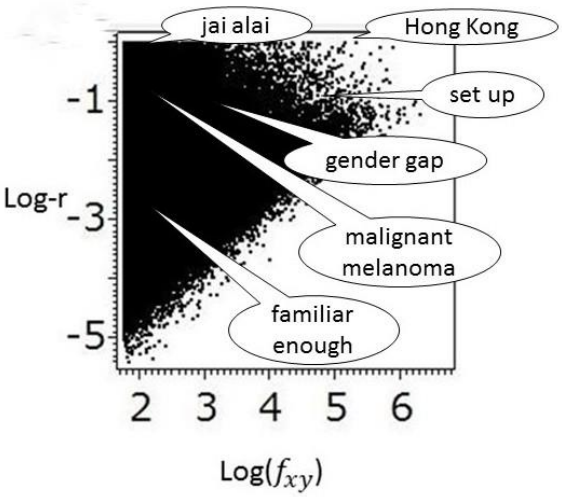

Fig 4: Log-r and Log(f_{xy}) with examples

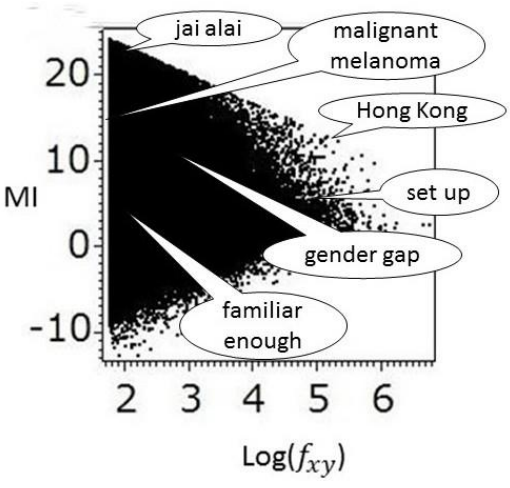

Fig 5: MI and Log(f_{xy}) with examples

A comparison of the placement of bigrams in Figures 4 and 5 reveals more clearly that the MI scatter plot (Figure 5) is a transformation of the Log-r scatter plot (Figure 4). Whereas in Figure 4, *jai alai* (ball game of Basque origin) and *Hong Kong* are located at the same high position on the y-axis, in Figure 5, the latter is considerably lower than the former. Intuitively, one would think that the two collocations differ in frequency but not of strength of association, and Figure 4 supports this. In the MI-based Figure 5, the placement of *Hong Kong* is below *malignant melanoma* and near *gender gap*. Thus, in the results based on MI, *Hong Kong* is undervalued. It can be seen from the graph shape that the cause of this undervaluation is its high frequency.

4.4. MI's Structural Problem

As we already mentioned, the developers of MI have acknowledged that it overvalues bigrams that have a low frequency.

> Since the association ratio becomes unstable when the counts are very small, we will not discuss word pairs with f(x, y) < 5. (...) For the remainder of this paper, we will adopt the simple but arbitrary threshold and ignore pairs with small counts. (Church & Hanks, 1990: 24)

However, as can be seen in Figure 5, MI has at the same time the structural problem that it undervalues high-frequency bigrams. This problem does not exist for Log-r, as its evaluation of bigrams is not influenced by frequency. Therefore, it appears that, compared to MI, Log-r is more appropriate for accurately describing the strength of association of bigrams.

4.5. The Universality of the Shape of Log-r and MI-based Scatter Plots

Figures 6 and 7 below present scatter plots of 400,000 French bigrams based on the data listed in Table 1. In Figure 6, Log-r and log(f_{xy}) are given on the vertical and horizontal axes, respectively; and in Figure 7, MI and log(f_{xy}) are on the vertical and horizontal axes, respectively. Despite differences in language, number of corpus words, number of bigrams, and the lower limit of the frequency of occurrence, the shapes of the Log-r scatter plots in Figures 6 and 2, as well as that of the MI scatter plots in Figures 7 and 3, are similar to the extent that it is difficult to tell them apart.

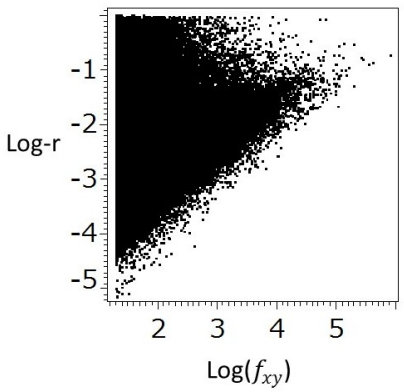

Fig 6: Log-r and log(f_{xy}) in French

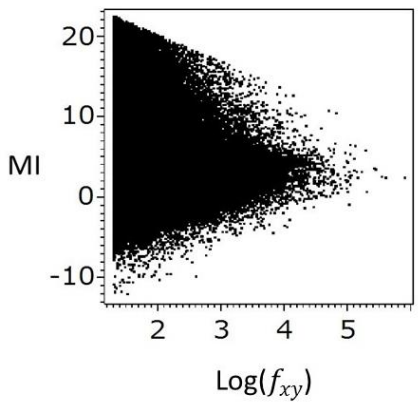

Fig 7: MI and log(f_{xy}) in French

If one uses bigrams extracted from the entirety of a naturally formed large-scale text, the shape of Log-r / log(f_{xy}) scatter plots will be the same, as will that of MI /

log(f_{xy}) scatter plots. They can be said to have a universality that transcends individual languages. In other words, the difference between Log-r and MI is universal.

Our analysis of the scatter plots made certain points clear: MI is a measure that merges strength of association and frequency such that, as frequency increases, the MI value decreases. On the other hand, Log-r is a simple measure that only reflects strength of association. By combining Log-r and the logarithm of frequency: log(f_{xy}), it is possible to show clearly the characteristics of bigrams.

5. INFERRING TYPES BASED ON SCATTER PLOT POSITION

Lastly, we will attempt to create a typology of bigrams considering their position in scatter plots. We will do so based on the strength of association between bigram words: Log-r, the logarithm of bigram frequency: log(f_{xy}) and, as well as the vocabulary level of the words constituting the bigram. Below, we will assume that the frequencies of word x and word y: f_x and f_y are the same for the sake of simplicity.

Generally, when given the total number of words in a corpus N and a word's frequency f_x in this corpus, following the approximate formula based on Zipf's law: 'rank $= 10/$occurrence rate (%)' (See 2.3 above), the frequency f_x can be converted into its frequency rank within the corpus. It can be assumed that in the case of a large corpus, a word's frequency rank will match its vocabulary level. These calculations enable one to grasp intuitively the word type listed in Table 3 like 'unknown', 'usual', 'functional'.

Table 3 shows the results of a simulation for each part of the scatter plot when the total number of words in the corpus N was assumed to be 1 billion. Figure 8 illustrates this as a scatter plot (as in Figures 2 and 4).

The arrow going from the top left to the bottom right in Figure 8 shows the vocabulary levels of the constituent words. Bigrams comprised of unfamiliar words are found at the top left, and highly familiar words are found around the bottom right. At the very top left are bigrams comprised of words at the 1 million word level (labelled 'unknown' in Table 3) that are decidedly atypical and unknown. At the bottom area appear bigrams comprised of words at the 10 or 100 word level (labelled 'hyper-functional' or 'functional').

According to the above-proposed criteria, as an illustrative example, we show in Table 4 the characteristics of the six English bigrams from Figure 4. The number of +'s indicates the level of each measurement. Both *Hong Kong* and *jai alai* have strongest association between their constituent words. The most significant difference between them consists of their frequency. Similarly, both *set up* and *malignant melanoma* have the same degree of strength of association. Their difference lies in their frequency and their constituent words' vocabulary level (Word type in Table 4). *Set up* is a frequent collocation constituted of 'functional' words, while *malignant melanoma* is an infrequent collocation constituted of 'rare' words.

Table 3: Simulation of Each Part of a Log-r / Log(f_{xy}) Scatter Plot

Log-r	Log(f_{xy})	f_x	Occurrence rate: f_x/N (%)	Rank/ Vocabulary level (x)	Word type (x)	cf. MI
0	2	100	0.00001	1,000,000	Unknown	23.3
-1	2	1000	0.0001	100,000	Rare	16.6
-2	2	10,000	0.001	10,000	Usual	10.0
-3	2	100,000	0.01	1,000	Basic	3.3
-4	2	1,000,000	0.1	100	Functional	-3.3
-5	2	10,000,000	1	10	Hyper-functional	-10.0
0	3	1,000	0.0001	100,000	Rare	19.9
-1	3	10,000	0.001	10,000	Usual	13.3
-2	3	100,000	0.01	1,000	Basic	6.6
-3	3	1,000,000	0.1	100	Functional	0.0
-4	3	10,000,000	1	10	Hyper-functional	-6.6
0	4	10,000	0.001	10,000	Usual	16.6
-1	4	100,000	0.01	1000	Basic	10.0
-2	4	1,000,000	0.1	100	Functional	3.3
-3	4	10,000,000	1	10	Hyper-functional	-3.3
0	5	100,000	0.01	1,000	Basic	13.3
-1	5	1,000,000	0.1	100	Functional	6.6
-2	5	10,000,000	1	10	Hyper-functional	0.0
0	6	1,000,000	0.1	100	Functional	10.0
-1	6	10,000,000	1	10	Hyper-functional	3.3

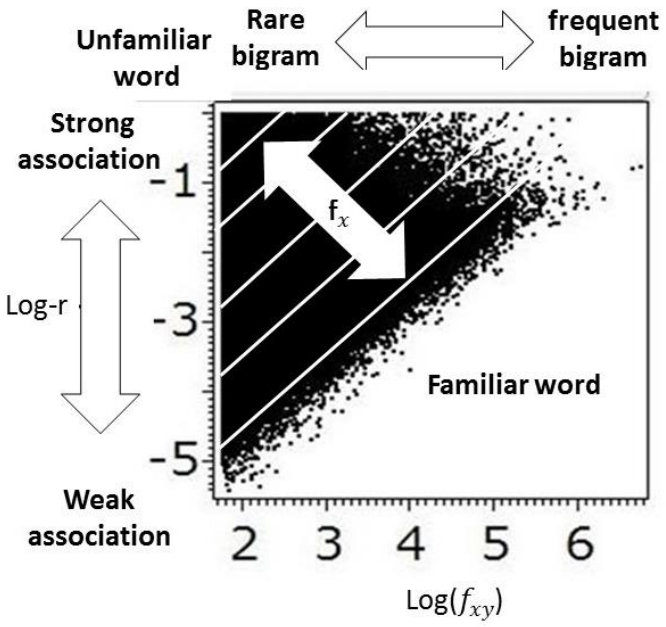

Fig 8: Parts of a Log-r / Log(f_{xy})-based Scatter Plot

Table 4: Characterisation of Six Bigrams

Bigram	Strength of association: Log-r		Frequency of bigram : Log(f_{xy})		Word type (x)	cf. MI
Hong Kong	+++++	-0.01	+++++	5.16	Basic	12.82
jai alai	+++++	-0.02	++	2.18	Unknown	22.63
set up	++++	-1.01	+++++	4.90	Functional	7.04
malignant melanoma	++++	-1.02	++	2.04	Rare	16.73
gender gap	++++	-1.20	+++	3.08	Usual	11.86
familiar enough	++	-2.97	++	2.05	Basic	3.47

We can continue to analyse bigrams from these three points of view and make a typological study of collocations. MI does not allow us to describe collocation types.

6. CONCLUSION

In this paper, we proposed a new measure, Log-r, as a straightforward measure for calculating the strength of association between bigram's constituent elements and showed that Log-r is more useful than MI for describing collocation types.

MI measures both frequency and strength of association at the same time, whereas Log-r measures only strength of association. Therefore, even if MI is a practical tool for finding collocations that are infrequent but are composed of strongly associated words, it is not capable of correctly evaluating the strength of association between two words.

By measuring degree of association of bigrams using Log-r, a simple statistic, and combining it with other simple statistics like frequency of occurrence and vocabulary level of constituent words based on Zipf's law, we can describe and explain, through visual representation, different collocation types including those that have been overlooked in the past.

It goes without saying that it is important to know the characteristics of a measure if it is used for measurement of an object of research. To make a comparison between various association measures including t-score and Log-Likelihood Ratio, in relation to the frequency of bigrams, it is necessary to work on the entirety of a naturally formed large-scale corpus of texts and not on a list of predetermined bigrams like adjective + noun.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 26370483. We thank Naohiro Takizawa, Ritsumeikan University (Japan), who helped us a lot to achieve this work.

REFERENCES

BAKER, P., HARDIE, A., & MCENERY, T., 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.

BARONI, M., 2009. Distributions in text, Lduelling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, Berlin, 803-822.

BYBEE, J., 2010. *Language, usage, and cognition*. Cambridge University Press.

CHURCH, K., & HANKS, P., 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1), 22-29.

ELLIS, N.C., 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, 32,17-44.
EVERT, S., 2009. Corpora and collocations, Lduelling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, 1212-1248.
FRANÇOIS, J., & MANGUIN, J.-L., 2006. *Dispute théologique, discussion oiseuse* et *conversation téléphonique*: Les collocations adjectivo-nominales au cœur du débat, *Langue Française*, 150, 50-66.
GRIES, S. Th., 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36(3), 477-510.
GRIES, S. Th., 2013. 50-something years of work on collocations What is or should be next..., *International Journal of Corpus Linguistics*, 18:1, 137-165.
HUNSTON, S., 2002. *Corpora in Applied Linguistics*, Cambridge University Press.
PECINA, P., 2010. Lexical association measures and collocation extraction, *Lang Resources & Evaluation*, 44, 137-158.
WRAY, A., 2012. What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play, *Annual Review of Applied Linguistics*, 32, 231-254.
ZIPF, G. K., 1949. *Human Behavior and the Principle of Least Effort*, Addison-Wesley.