Appropriateness of Log-r for calculating strength of association : Comparison with MI, LLR using Japanese and English bigram data

結合の強度を測る指標としての Log-r の有用性：日・英語のバイグラムデータに基づく MI，LLR などとの比較

藤村 逸子（名古屋大学）†
青木 繁伸（群馬大学名誉教授）

Appropriateness of Log-r for calculating strength of association: Comparison with MI, LLR using Japanese and English bigram data

Itsuko Fujimura (Nagoya University)
Shigenobu Aoki (Gunma University)

要旨

２語からなるコロケーションは一般に共起頻度と２語の結合力によって特徴づけられる。本研究は，結合力の指標として Fujimura & Aoki (2016)において提案した Log-r を，同じ目的の指標として言及されることの多い MI，LLR，t-score，Dice，Jaccard と比較し，簡素な指標である Log-r の有用性を主張する。データは『現代日本語書き言葉均衡コーパス』と英語の大規模新聞コーパスから網羅的に採取した多量のバイグラムを用いる。横軸にバイグラムの共起頻度をとり，縦軸に各指標値をとった散布図を作成して各指標の特徴を視覚的に描き，散布図間の比較によって指標間の差異を明示する。

１．はじめに

大規模コーパスに基づく言語研究のひとつとしてコロケーションの研究が盛んに行われている。コロケーションは語と語の慣用的な結合と定義されるが，それには種々 のタイプのものが含まれる。それぞれのタイプを特徴づける基本的な特性として言及されることが多いのは，連語の粗頻度と，連語を構成する単語間の結合の強度の 2 つであ る（Ellis 2012; Gries 2012; Wray 2012）。粗頻度はわかりやすい特性であるが，結合の強度は名称もさまざまであり統一的に扱われてはいない。また，その指標（およびその計算式）としては MI（Mutual Information）（Church & Hanks 1990）に言及されることが多い(Ellis 2012; Evert 2009; Gries 2012; Hunston 2002)が，一方で種々の指標（および計算式）が提案され（Pecina 2010; 相澤・内山 2011），研究はいまだに途上にあると言える(Bybee 2010; Evert 2009; Gries 2013)。

日本語には「半信-半疑」，「徹頭-徹尾」，「金科-玉条」，「有象-無象」，「換骨-奪胎」，「夫唱-婦随」，「官尊-民卑」のようなコロケーションが存在する。これらがどれも 1 語性の強い連語であることは直感的に感じられる。また，本研究のコーパスの『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）においては，これらの連語の構成形態素の一方は必ず他方と共起し，その他の形態素とは共起しない（「半信」は「半疑」とのみ共起する。「半疑」も「半信」とのみ共起する。他も同様。）。しかし，結合の強度を測るはずの上記の指標によってこれらの連語を計測すると，直感や事実に反して，その値はこれらの連語間で同一とは限らない。

† fujimura@nagoya-u.jp

言うまでもなく，現象を計測するための指標の特徴が曖昧であることは望ましくないが，現段階において，それぞれの指標の特徴に関する明示的な説明はなされていないのが状況である。

我々は Fujimura & Aoki (2016) において，2 語連語（以下バイグラム）[1]の結合の強度をはかる簡素な指標として Log-r を提案し，英語とフランス語の大規模データをもとに MI と対照させて，言語現象としてのコロケーションを理解する上でのその有用性を主張した。本発表は主として日本語を扱い，Log-r を用いることによって日本語のコロケーションの記述に貢献できることを示す。また MI の他に，LLR(Log-Likelihood Ratio)，t-score, Dice, Jaccard とも比較してそれぞれの指標の特徴を明らかにし，バイグラムの結合の強度を測る指標としては Log-r が有用であることを明らかにする。

２．Log-r，対数，頻度と強度に基づく特徴づけ

本章では，Log-r を紹介する。また，Wray(2012)による連語の頻度と構成語の結合の強度に基づくコロケーションの特徴づけのモデルを示し，語彙の分布に関する研究における対数の価値を説明する。

２．１．　Log−r

2 語の結合の強さを示す指標として，2 変数（単語 x と単語 y）の属性相関を表すピアソンの積率相関係数(r)の常用対数を提案し，それを Log-r と名づける（Fujimura & Aoki 2016）。ピアソンの積率相関係数の定義式は(1)である。本研究では，ポワソン分布を仮定して，その近似式を用いる。Log-r はしたがって，(2)のように定義される。

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

$$\text{Log-r} = log_{10} \frac{f_{xy}}{\sqrt{f_x f_y}} \tag{2}$$

(f_{xy}：バイグラム xy の頻度，f_x：x の頻度，f_y：y の頻度)

Log-r は，2 語の結合度を測る指標としてすでに提案されている z スコア，カイ二乗値 (χ^2)，phi 係数，コサインと共通した性質をもっている（cf. Pecina 2010; 相澤・内山 2011）。すなわち Log-r は全く新規の指標というわけではない。

２．２．　連語の頻度と強度による特徴づけのモデル

図１は，連語の特徴をその頻度と構成要素間の強度に基づいて特徴づける Wray (2012)によるモデルである。横軸は頻度を表し，縦軸は強度を表している。縦軸はバイグラムの構成要素の結合度の強さ，すなわちバイグラムの１語性の度合いを表す。これはバイグラムの頻度とは異なる概念である。頻度の多さと結合度の強さは独立しているはずである。第１象限には頻度大かつ強度大，第２象限には頻度小かつ強度大，第３象限には頻度小かつ強度小，第４象限には頻度大かつ強度小の連語がプロットされる。それぞれの象限の典型的なバイ

1　ここで連語とは，その共起の慣用性に関わらず単に語の連続を指す。2 語連語(バイグラム)は 2 語の連続を指す。

グラムの例としては，第 1 象限には New York，White House などの高頻度の固有名詞，第 2 象限には低頻度で結合度の強固な bovine spongiform, lingua franca などのイディオム，第 4 象限には高頻度で強度は弱い of the, I am などのレキシカルバンドル，第 3 象限は pink roses や familiar enough などのその他の平凡な 2 語の連続をあげることができる。言うまでもなく，図 1 の縦軸と横軸は連続体をなしている。本研究では，このモデルにならい，頻度を横軸にとり，強度の指標を縦軸にとって，どの指標が現実の言語現象によりよく適合するかを検討する。

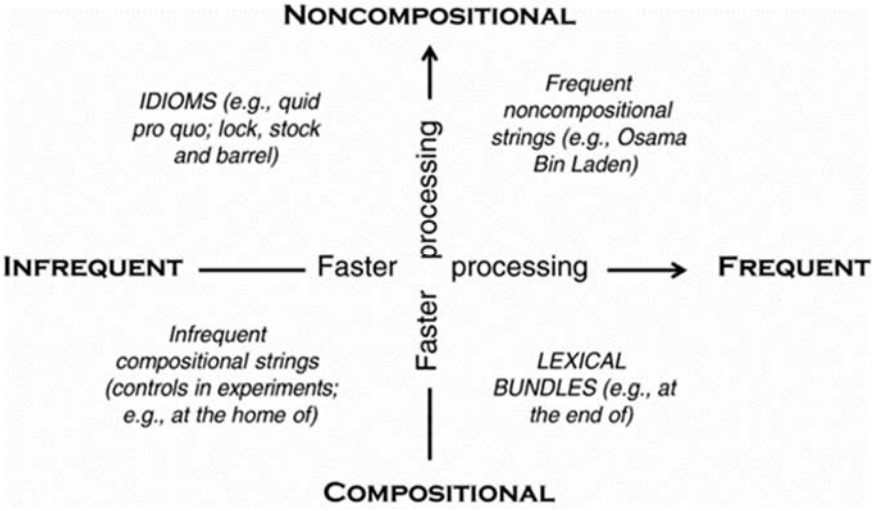

図 1　Wray(2012)によるコロケーションの特徴づけモデル

２．３．　対数

Log-r が r の常用対数をとっているのは，語の頻度分布が Zipf の法則(Zipf 1949)に従う極端に範囲の広い統計量だからである（概算では，単語のコーパス内での出現率（%）＝10/順位)。対数をとって比較すると数値の処理が容易になり，現象をグラフ化して視覚的に把握できるようになる(Baroni 2009; 青木 2009)。この点は，以下で検討する t-score，LLR，Dice，Jaccard においても同様である。MI は定義式においてすでに対数がとられているので，Log-r との比較が定義式のままで可能であるが，以上の 4 指標は通常の定義式のままでは Log-r との視覚的な比較はできない。したがって，LLR, Dice, Jaccard, t-score の検討はその常用対数値によって行う。また，横軸となる頻度それ自体もその常用対数値を基に考察する。

３．データ

データの詳細は表 1 のとおりである。

バイグラムデータの作成は日英両言語とも Unix 環境において, Unix コマンドと Perl スクリプトを使ったプログラミングによって行った。英語のデータの収集は，表 1 に挙げた新聞コーパスをもとに，形態素情報の付与は行わず，単語を単位として行った。こうしてすべての単語の頻度表とすべてのバイグラムの頻度表を取得した。日本語は，BCCWJ の短単位の TSV 形式データを用いた。バイグラムは形態素情報付きの出現形を単位として作成し，短単位形態素の頻度表とバイグラム頻度表を取得した。バイグラム頻度の低いものを削除した上で, 英語の 104 万件のバイグラムと日本語の 62 万件のバイグラムのそれぞれについて，

Microsoft Excel を用いて各指標の値を計算し，データベース化した。散布図の作成およびその他の統計処理は統計ソフトの Jmp ver.13 を用いて行った。

表１　使用コーパスとバイグラムデータ

	バイグラムの個数とコーパスの総語数・総形態素数	単位	テキストの種類	コーパスの名称と配布元
英語	・1，040，000 個（生起数 54 回以上） ・10 億語	・単語 ・形態素解析なし ・大文字・小文字は区別して扱われている	新聞	LDC ・North American News Text Corpus ・North American News Text Supplement
日本語	・615，000 個（生起数 10 回以上） ・1 億形態素	・短単位（レンマ化なし） ・UniDic による形態素解析情報付き	書籍全般，雑誌全般，新聞，白書，ブログ，ネット掲示板，教科書，法律	国立国語研究所 ・BCCWJ（DVD 版）

４．英語と日本語のバイグラムの Log-r と MI

４．１．Log-r と MI

表 2 は，英語と日本語データから取り出したバイグラムである。最上位の Log-r が 0 の場合は，バイグラムの構成要素が互いに排他的に共起する場合である。「半信半疑」においては，「半信」は常に「半疑」と結びつき，「半疑」も常に「半信」と結びつく。最下位の Log-r が-4 付近のものは，構成要素間の共起は何らかの偶然である場合である。この区間内は，結合度の強度に関する連続性が存在する。上位にあるほど２語の結合度は強く，Log-r も MI も同じ傾向を示しているように見える。MI の計算式は次の通りである。

$$\mathrm{MI} = \log_2 \frac{f_{xy}\, N}{f_x\, f_y} \qquad (3)$$

（Nはコーパスの総語数・総形態素数）

表 2 英語と日本語のバイグラム例の Log-r 値と MI 値[2]

結合の強度	英語			日本語		
	バイグラム例	Log-r	MI	バイグラム例	Log-r	MI
強	lingua franca	-0.01	22.5	半信 半疑	-0.00	19.2
⇕	apple pie	-1.01	13.8	有害 物質	-1.02	11.5
	medal winner	-2.00	8.0	環境 対策	-2.00	5.4
	earlier offer	-3.00	2.3	文化 意識	-3.02	2.4
弱	no there	-4.04	-3.4	利用 その	-4.03	-3.7

４．２．英語データ

しかし，図 2 と図 3 の散布図を見ると Log-r と MI には明確な違いがあることがわかる。

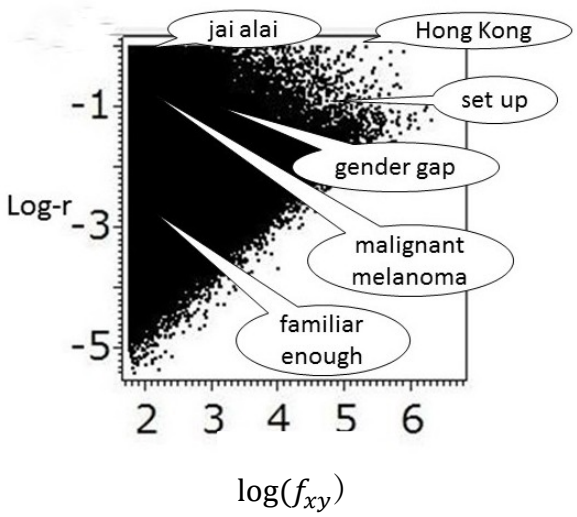

図 2　Log-r/log(f_{xy})の散布図(英語）[3]

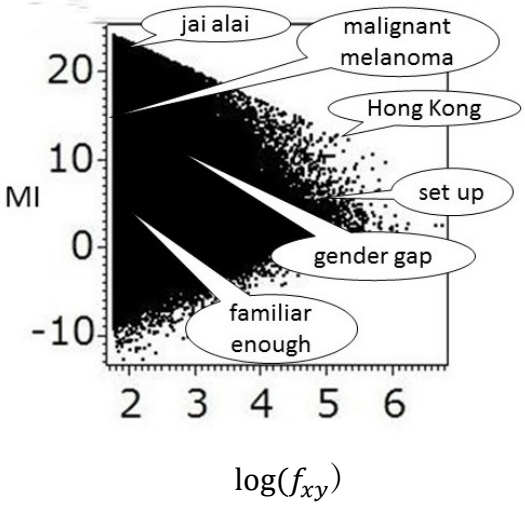

図 3　MI/log(f_{xy})の散布図(英語）

両図には約 100 万件のバイグラムがプロットされている。両図とも横軸はバイグラムの頻度（の常用対数）である。縦軸は図 2 ではバイグラムの Log-r であり，図 3 ではバイグラムの MI である。上掲の図 1 のモデルに従うと，縦軸には頻度とは独立したものとして強度がプロットされるべきである。図 2 ではそれが実現しているが，図 3 の上辺を見るとバイグラムの頻度が増加するにつれて MI 値は例外なく低下しており，２項目は独立しているとは言えない。たとえば散布図上に例として位置が示してある Hong Kong と jai alai (バスク地方のスポーツ)は，２語の結合の強度の観点においては，共通した特徴を持っている。それ

[2] MI 値は，コーパスの総語数・総形態素数 N にも左右される。N は英語データでは 10 億，日本語データでは 1 億であり，底を 2 とする N の対数値は，それぞれ 29.9 と 26.6 となる。本データの英語のバイグラムの MI 値は日本語のそれと比べてデフォルトで 3.3 大きい。この表において， MI 値を基準として日本語と英語のバイグラム（たとえば「半信半疑」と「lingua franca」）を比較するのは不可能と言ってよい。

[3] 図 2 と図 3 は，Fujimura & Aoki 2016 Fig2， Fig3 より転載。

ぞれの構成単語は他の要素とはほとんど共起せず，相互の結合は強固である[4]。すなわち，Hong の生起のうちの 97%が Kong と，Kong の生起のうちの 97%が Hong と共起している。同様に jai のうちの 91%が jai と，alai のうちの 98%が jai と共起している。図 2 においては Hong Kong と jai alai はグラフの最上部に等しくプロットされているのは理にかなっている。２つのバイグラム間で異なるのは頻度である。使用したコーパスにおいて Hong Kong は jai alai に比べて極めて高い頻度で出現している。図 2 において, Hong Kong の MI 値は malignant melanoma よりも低く gender gap のそれに近い。Hong Kong の 1 語性が強いことはデータから明らかであるので，この結果は言語使用の現実に反しているといえる。図 3 の上辺が右下がりの直線になっているのは現実の言語現象の反映ではなく, MI の計算式に由来している。強度の計測において頻度が影響を及ぼすのは，現象の正確な記述の目的には反すると考えられる[5]。なお，図 2 と図 3 において下辺が右上方向に切れ上がっているのは言語現象に対応している。バイグラムの頻度が上がるにつれて，構成語間の共起の偶然性は減じるからである。高頻度のレキシカルバンドルの結合度は必然的に高くなる。この点において，図 1 の Wray(2012)の正方形のモデルは言語現象の現実に対応してはいない。

４．３．日本語データ

次に，日本語データから作成したバイグラムを用いて，Log-r と $\log(f_{xy})$，MI と $\log(f_{xy})$の散布図を以下に示す。

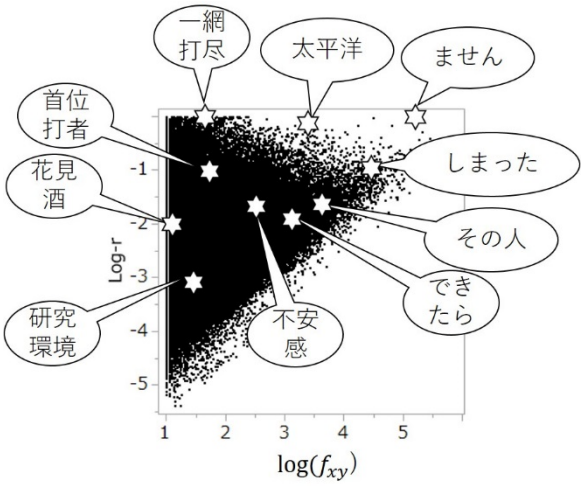

図 4　Log-r/$\log(f_{xy})$の散布図(日本語)

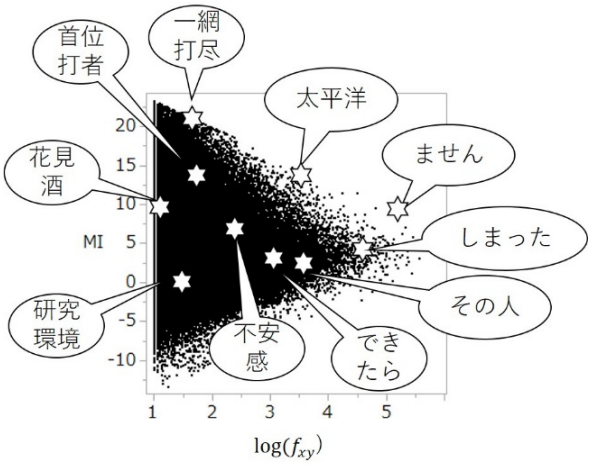

図 5　MI/$\log(f_{xy})$の散布図(日本語)

最初に言えるのは，Log-r を縦軸とする図 2 と図 4, MI を縦軸とする図 3 と図 5 の間で，散布図の形状が言語を超えて近似しているという点である。大規模なコーパス全体を網羅的に対象にする限り，言語を超えて Log-r と $\log(f_{xy})$による散布図の形状は同じであり，MI と $\log(f_{xy})$による散布図もまた同じである。この現象は，語彙の分布が言語を超えて Zipf の

[4] バイグラムの頻度とその構成要素の頻度は以下のとおりである。f(Hong Kong)：143，104，f(Hong):147，404，　f(Kong):147，852，　f(jai alai)：151，　f(jai):166，　f(alai):154.

[5] MI は，頻度は低いが，結合が強固なバイグラムの発見のための実用的ツールとしては役に立つと思われる。MI と Log-r の散布図を比べると両者は同一のグラフの変形であることがわかる。

法則に従うという問題に通じる[6]。換言すると，各指標の特徴を評価する際のサンプルとしては，コーパス全体を対象にする必要がある。

次に図 4 と図 5 とを比較すると，上述の図 2 と図 3 との比較と同様のことが観察できる。BCCWJ において「一網打尽」は「一網_名詞-普通名詞-一般」と「打尽_名詞-普通名詞-一般」のバイグラムと分析され，「ません」は「ませ-助動詞」と「ん-助動詞」のバイグラムと分析されているが，下に掲げる表 3 に記したように，「一網」はその 95%が「打尽」に後続され，「打尽」はその 98%が「一網」に前置されている。また，「ませ」はその 98%が「ん」に後続され，「ん」はその 90%が「ませ」に前置されている。「ません」は「一網打尽」に比べてわずかに結合の強度が弱いが，いずれにせよどちらも 1 語性の極めて強いバイグラムである。「一網打尽」と「ません」の極めて大きな差異はその出現頻度である。Log-r による散布図（図 4）はこの現実をよく表していると言える。

英語に関して述べたように，MI には頻度の低いバイグラムを高く評価し，頻度の高いバイグラムを低く評価するという数式上の特徴がある。「ません」の MI 値は「一網打尽」の MI 値と比べて大変低く，散布図上では「花見酒」や「首位打者」よりも下にある。「花見」が「酒」に後続されるのはその 2%に過ぎず，「酒」が「花見」に前置されるのはその 0.5%に過ぎない(表 3 参照)。共起の強度の測定である限り，90%以上の共起率の「ません」が 2 %以下の共起率の「花見酒」よりも下位にプロットされるのはあり得ないので，MI が測っているのは共起の強度ではないということになる。

直感的に，「一網打尽」や「花見酒」や「首位打者」は内容語的なバイグラムであるのに対して，「ません」や「しまった」は機能語的なバイグラムであると感じられる。しかし，機能語的か内容語的かの差異は結合の強度とは別の問題である。

本節では英語と日本語を対象に，Log-r と MI の比較を行った。ここで言えるのは，Log-r は結合の強度を頻度とは独立に測っているのに対して，MI は結合の強度と頻度を融合させた指標であり，図 1 における縦軸を構成するには適してはいないということである。

５．日本語バイグラムに基づく Log-r と t-score，LLR，Dice，Jaccard との比較

５．１．各指標とバイグラム例の値

本節では，日本語のバイグラムデータを用いて，Log-r とバイグラムの結合度を計測する指標として言及されることの多い他の指標とを比較する。取り上げるのは，t-score，LLR（Log-Likelihood Ratio），Dice，Jaccard の４つの指標である。それぞれの指標の計算は以下の式による(Petina 2010)。　なお，N はコーパスの総語数・総形態素数である。

- t-score

$$\left(f_{(xy)} - \frac{f_{(x)} \times f_{(y)}}{N}\right) \div \sqrt{f_{(xy)}} \tag{4}$$

- LLR (Log-Likelihood Ratio)

$$-2 \sum_{ij} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}} \tag{5}$$

[6] Fujimura & Aoki (2016)では，同様の比較を英語とフランス語間で行った。二次元的散布図の形状の差異は，英語と日本語間より，英語とフランス語の間の方が小さい。英仏語と日本語のデータを比べると，次の３つの顕著な差異がある。１）言語の特徴：英仏語は日英語より近い，２）処理の単位：英仏語では単語であり，日本語では短単位形態素である，３）テキストジャンル：英仏語は新聞のみであるが，日本語は種々のものを含む。

- Dice

$$\frac{2 \times f_{(xy)}}{f_{(x)}+f_{(y)}} \tag{6}$$

- Jaccard

$$\frac{f_{(xy)}}{f_{(xy)}+f_{(x)}+f_{(y)}} \tag{7}$$

上で述べたとおり，Log-r との視覚的な比較の目的のために，これらの 4 指標は常用対数に変換して用いる。

表 3 は, すでに図 4 と図 5 において例に挙げた 10 個のバイグラムのデータを掲げている。すなわち，表 3 には，それぞれのバイグラム頻度，構成形態素の頻度，対数化されたバイグラム頻度，Log-r 値，MI 値，対数化された t-score 値，対数化された LLR 値，対数化された Dice 値，対数化された Jaccard 値が示されている。

表 3　バイグラム例，それぞれの頻度と各構成形態素の頻度，および各指標値

	$f_{(xy)}$	$f_{(x)}$	$f_{(y)}$	$\log(f_{xy})$	Log-r	MI	t-score	LLR	Dice	Jaccard
一網-打尽	42	44	43	1.6	-0.02	21.1	0.81	3.11	-0.02	-0.49
ませ-ん	142,609	144,908	158,770	5.2	-0.03	9.3	2.58	6.31	-0.03	-0.50
太平-洋	2,635	2,961	3,934	3.4	-0.11	14.5	1.71	4.73	-0.12	-0.56
首位-打者	50	501	612	1.7	-1.04	14.0	0.85	2.94	-1.05	-1.37
不安-感	321	10361	18540	2.5	-1.64	7.4	1.25	3.43	-1.65	-1.96
その-人	5,734	390,373	150,791	3.8	-1.63	3.3	1.83	4.21	-1.67	-1.98
しまっ-た	28,516	37,113	2,692,208	4.5	-1.04	4.8	2.21	5.22	-1.68	-1.99
でき-たら	1,180	97,682	110,418	3.1	-1.94	3.5	1.49	3.55	-1.95	-2.25
花見-酒	12	626	2474	1.1	-2.02	9.6	0.54	2.13	-2.11	-2.41
研究-環境	34	39847	30,652	1.5	-3.01	1.5	0.57	1.42	-3.02	-3.32

以下では，4 節と同じく，60 万件のバイグラムによる散布図（各指標／$\log(f_{xy})$）と，その散布図上にプロットされた表 3 のバイグラムをもとに，各指標の特徴を明らかにする。

５．２．　t-score

t-score は MI とともに古くからコロケーションのための指標として言及されている。しかし，図 6 からわかるように，t-score はバイグラムの頻度にほぼ相関した値をとる。図 1 の Wray(2012)のモデルを想定して, t-score を結合の強度を測る指標として用いることは全く不適切であると言える。頻度によって値が変わるので，第 1 節に挙げた熟語の場合，BCCWJ において「半信-半疑」は「夫唱-婦随」より高頻度であるため t-score の値も高いが、別のコーパスにおいてもしも「夫唱-婦随」が「半信半疑」より頻度が高い場合には，両者の t-score の順序は逆転することになる。

５．３． LLR

G スコアとも呼ばれる LLR(Log-Likelihood Ratio)は，Dunning(1993) によりコロケーションの指標として導入された。ライプツィヒ大学の Wortschatz[7]にコロケーションの指標として実装されているなど，実際によく使用されている。図 7 からわかるようにこの指標も頻度との相関性が高い。従って，Wray(2012)のモデルの共起の強度の軸とするには不適切である。

t-score と LLR の値は log(f_{xy})が高いほど増加する。この点，MI とは正反対の指標である。いずれにせよ，これらの 3 指標は頻度から独立しては計測されないので，共起の強度を特徴付ける目的には適合しないと考えられる。

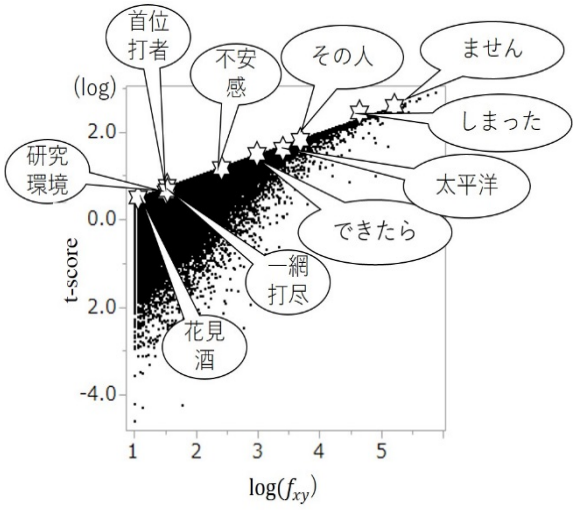

図 6　t-score/log(f_{xy})の散布図[8]

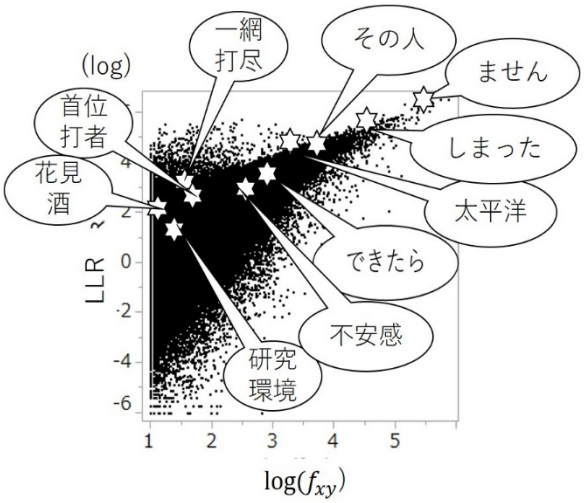

図 7　LLR/log(f_{xy})の散布図

５．４． Dice と Jaccard

最後に Dice と Jaccard を検討する。図 8 と図 9 からわかるように，Dice と Jaccard は大変よく似ている。例に挙げたバイグラムの中では「不安感」の位置が若干異なる以外は，順位に変わりはない。また，図 4 との比較からわかるように，この２つの指標は Log-r ともよく似ている。

平面的な散布図では，Dice，Jaccard, Log-r は見分けがつかないほどよく似ている。Dice と Jaccard よりも Log-r が結語の強度を測る指標として有用であることを主張するためには，これらの 3 指標の差異を別の角度から検討する必要がある。

7 http://corpora2.informatik.uni-leipzig.de/?dict=fra_mixed_2012

8 t-score の散布図において，高密度部分は上辺近辺に偏っている。LLR ではその偏りが緩和されている。

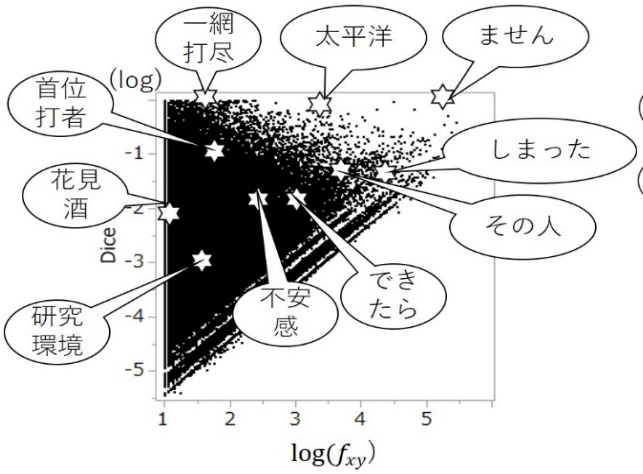

図 8　Dice /log(f_{xy})の散布図

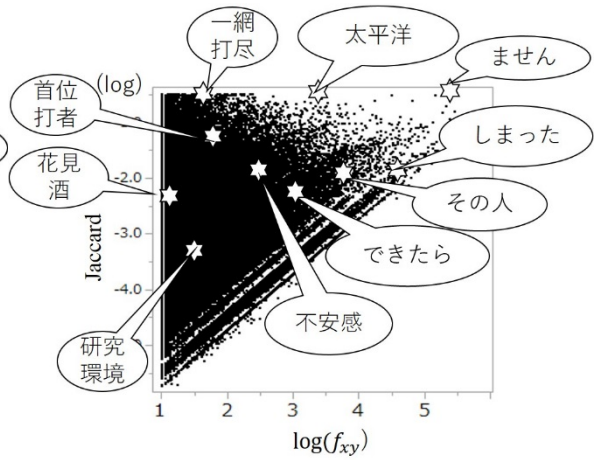

図 9　Jaccard /log(f_{xy})の散布図

６．三次元の散布図による Log-r，Dice，Jaccard の比較

Dice，Jaccard，Log-r の違いを明らかにするために，図 10，図 11，図 12 では，それぞれ図 8，図 9，図 4 と同一の散布図上に，JMP のノンパラメトリック密度推定を用いて，バイグラムの度数の密度に応じた 5%刻みの等高線を引いた。また，log(f_{xy})と各指標の頻度分布を散布図の上部と右にヒストグラムで示した。等高線は色別されており，赤色が最も密度が高く，寒色になるにつれて密度が低い。最後の等高線の外側は全体の 5%にあたるバイグラムが薄く分布する。図 10 の Dice と図 11 の Jaccard はほぼ同一の分布であるが，図 12 の Log-r は全く異なる。図 10 と図 11 では，最も密度が高いのは左端の最下部である。この位置はバイグラムの頻度と強度が最も弱いと想定される場合に当たり，典型的な出現は意味のない（多くの場合は何らかの誤りに基づく）偶然の共起である。このような共起が最も多いということは常識的に考えにくく，Dice や Jaccard は言語使用の現実を反映する指標ではないと考えられる。一方，Log-r による散布図（図 12）において密度が最も高いのは，左端最下部から少し上の位置である。この位置は図 1 の第 3 象限に当たり，Wray(2012)の言う infrequent compositional strings のための定位置である。このような特徴のバイグラムがコーパスにおいて多数存在することは Zipf の法則によって想定できる。図 12 を見ると，密度の分布はなだらかな連続を構成していて，これも Zipf の法則に適合している。

Fujimura & Aoki (2016)では，log(f_{xy})と Log-r による散布図を描くと，共起の頻度と強度に加えて，バイグラム構成単語の親密度（familiarity）[9]も測ることが可能となり，３つの観点からバイグラムの分類ができると主張した。Dice や Jaccard に基づくバイグラムの分布は Log-r による分布に比べて均整がとれておらず，語の親密度を加えたモデルを想定することは困難である。語の親密度を計算するとバイグラムの特徴がより精密に記述できるので，この観点からも Dice や Jaccard に比して Log-r の有用性は高いと言える。

[9] Log-r と log(f_{xy})の散布図を，左上を頂点とする二等辺三角形と見なした場合，最も語の親密度が低いのは左上の頂点である。最も親密度が高いのは下辺であり，機能語はこの場所に位置する。

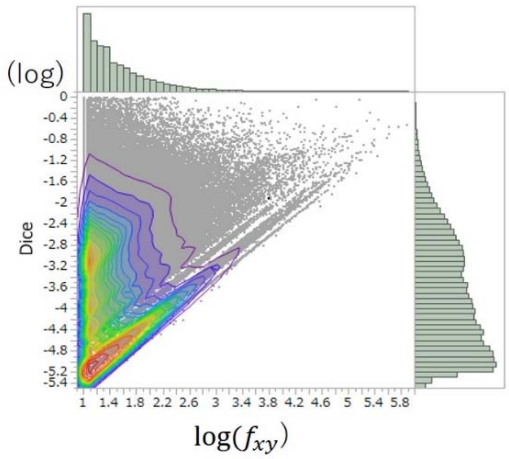

図 10　Dice /log(f_{xy})の三次元散布図

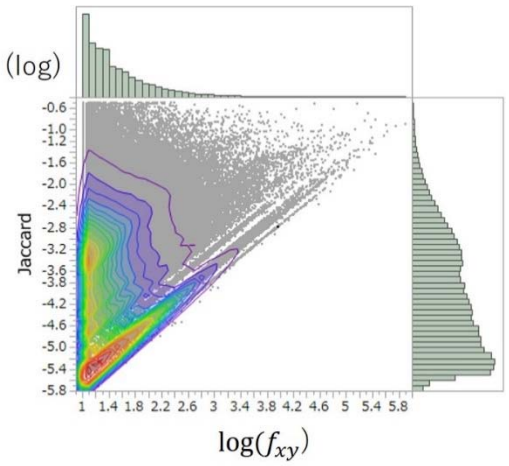

図 11　Jaccard /log(f_{xy})の三次元散布図

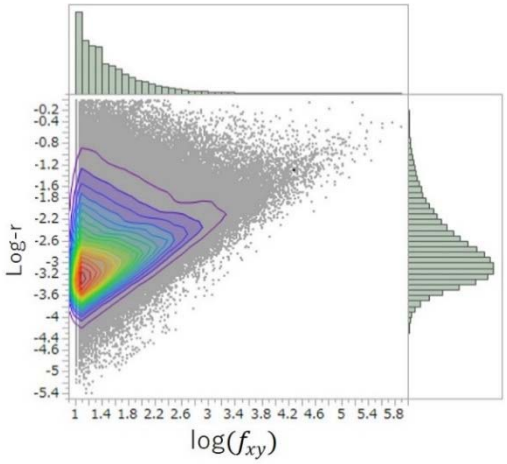

図 12　Log-r /log(f_{xy})の三次元散布図

7．おわりに

本稿では，日本語と英語の多量のデータを用いて，共起の強度を測る指標としての有用性の観点から筆者らが提案する Log-r と，コロケーションの指標として言及されることの多い MI, t-score, LLR, Dice, Jaccard とを比較した。その結果次のことがわかった。MI, t-score, LLR は頻度との相関が強く，強度の指標としては使えない。Dice と Jaccard は，一見 Log-r と近似してはいるが，現実の分布を表してはいない。Log-r は共起の強度のみを測る簡素な指標として最適である。Log-r を強度の指標として用い，頻度や密度などのその他の特徴を組み合わせることによって，バイグラムを多角的かつ正確に特徴付けることが可能になる。Log-r 以外の指標を用いる際には，その目的を明確に定める必要がある。

指標の特徴を評価するためには，本研究で行ったようにコーパス全体をサンプルとすることが必要である。バイグラムの特徴は一様ではないので，恣意的に選択したサンプルによる比較は避けるべきである。

Log-r を強度の指標として認定すると，頻度(log(f_{xy}))，強度(Log-r)，密度に加え，バイグラムの構成要素の親密度（familiarity）も散布図上にプロットでき，バイグラムの詳細な特徴付けに貢献できる (cf. Fujimura & Aoki (2016))。この作業は各形態素のレンマ形をもとに行う必要があるが，BCCWJ にはこのデータが備わっている。次の課題としたい。

謝　辞

本研究は科研費基盤(C)「大規模コーパスに基づく名詞と形容詞の使用パターンと構造化に関する日仏語対照研究」の助成による。英語バイグラムは，滝沢直宏教授（データ取得時は名古屋大学，現在は立命館大学）の提供によります。記して感謝いたします。

文　献

Baroni, M., (2009) Distributions in text, Lüdeling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, Berlin, pp. 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), pp. 22-29.

Dunning, T., E., (1993)　Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19(1), pp.61-74.

Ellis, N.C., (2012) Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear.　*Annual Review of Applied Linguistics*, 32, pp.17-44.

Evert, S., (2009) Corpora and collocations, Lüdeling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, pp.1212-1248.

Fujimura, I., & Aoki, S., (2016) A New Score to Characterise Collocations: Log-r in Comparison to Mutual Information, in *Europhras2015 Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives* pp. 271-282.

Gries, S. Th., (2012) Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36(3), pp.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, pp.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, pp.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, pp.231-254.

Zipf, G. K., (1949) *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

相澤彰子・内山清子（2011)「語の共起と類似性」松本裕治(編)『言語と情報科学』朝倉書店 pp.58-76.

青木繁伸(2009)『統計数字を読み解くセンス―当確はなぜすぐにわかるのか?』(DOJIN 選書 27) 化学同人.