logDice 係数はどのような共起指標か

恒川, 元行
Faculty of Languages and Cultures, Kyushu University

https://doi.org/10.15017/4104141

出版情報：言語文化論究. 45, pp.35-44, 2020-10-30. 九州大学大学院言語文化研究院
バージョン：
権利関係：

logDice 係数はどのような共起指標か

恒　川　元　行

1．はじめに

logDice 係数は、Rychlý（2008）において初めて提案されて以降、約10年の間にコロケーション抽出の基本的指標の一つとして重要な地位を占めるようになっている。

たとえば、現代ドイツ語コーパス、ベルリン・ブランデンブルク科学アカデミーの DWDS-Wortprofil では、共起指標が logDice 係数、頻度の２つである。[1] また、ドイツ語ウェブコーパス DEWAC を公開している Sketch Engine は、2006年９月以降、[2] Word Sketches の「統計を logDice に変更した」（Lexical Computing Ltd. 2015:1）としている。

日本語コーパスでも、国立国語研究所・Lago 言語研究所の共同開発によるオンライン検索システム NINJAL-LWP for BCCWJ（NLB）が、頻度、MI スコアなどと並び、コロケーション抽出のための基本的指標のひとつとして logDice 係数を採用している（NLB ユーザマニュアル 2016:20）。

logDice 係数はこのように基本的な共起指標としてコーパス検索システムへの採用が目に留まるが、その定義式の背後にある考え方や特徴などへの言及は、不思議なほど見当たらない。恒川（2020:118f.【補足02】）で触れたように、李（2015:75）が MI スコアの特徴を紹介した「注６」の末尾で、「なお、MI スコアの場合、イディオムを発見する際に、役に立つとされますが、ログダイス（LogDice）は汎用的な連語パターンを発見する際に、役に立つとされています。」と付け足しのように説明しているくらいである。このほか、Rychlý（2008）の要旨を抽出・紹介した杉浦（2019）[3] があるが、これはあまりにも簡潔に過ぎ、背景も含めた logDice 定義式の理解のためには補足的説明がぜひとも欲しいところである。

そのため、本稿では恒川（2020:118f.【補足02】）では解明が不十分に終わったことを踏まえ、logDice 係数の提案論文 Rychlý（2008）に立ち戻り、定義式や数値の意味を再考し理解を深めようと試みた。[4] 以下はその報告である。

2．Dice 係数[5]

「ログダイスは、語と語の結びつきの強さを表すダイス係数を対数化したもの」（赤瀬川ほか、2016:70、NLB ユーザマニュアル2016:22；下線は恒川）と説明され、「ログ＋ダイス」という名称からもそのような理解が示唆される。実際にはしかし、logDice 係数は Dice 係数の単なる対数化ではなく、後述のように調整数14が加えられていることに注意が必要である（第４章参照）。

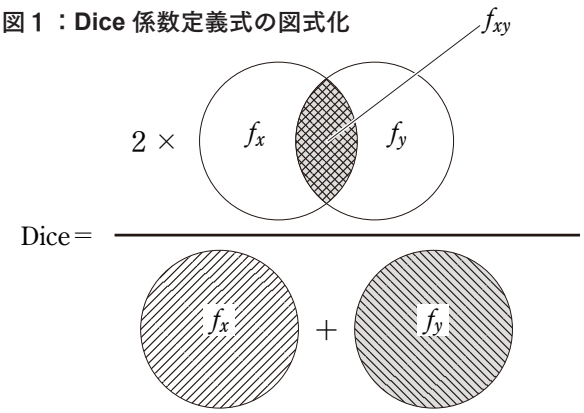
しかし、logDice 係数 が Dice 係数をその土台としていることは間違いなく、Dice 係数の特徴をそのまま引き継いでいる。そのため、ここではまず Dice 係数を概観しておきたい。

Dice 係数は次の式で定義される： $\text{Dice} = \frac{2f_{xy}}{f_x+f_y}$

すなわち、Dice 係数は「中心語 X と共起語 Y の共起頻度（f_{xy}）× 2」（=分子）を「X の出現頻度（f_x）と Y の出現頻度（f_y）の合計」（=分母）で割った値である。この定義式において分子の「中心語 X と共起語 Y の共起頻度（f_{xy}）」が × 2（2 倍）となっているのは、Dice 係数の最大値を 1 とするためと思われる。なぜなら、中心語 X と共起語 Y が相互に常時共起する場合（下の図の分子で 2 つの円が完全に重なる場合）があるとすると、その場合に Dice 係数は最大値となるが、$\frac{f_{xy}}{f_x+f_y}$ ではその値が0.5となってしまうからである。[6]

次の図は、この定義式を視覚的にわかりやすく図示したものである：[7]

図 1 ：Dice 係数定義式の図式化



前章で触れたように、DWDS-Wortprofil や NLB では logDice と並んで、単純頻度もまた共起指標として用いられている。頻度にはしかし、コーパスの規模に依存し、中心語や共起語自体が高頻度であれば偶然的共起の頻度も高くなり、規模の異なるコーパス間で数値の比較ができないという問題が伴っている（石川 2008b:108）。この問題は、Rychlý（2008:6）もそれまでの共起スコアに共通する問題点として指摘している。これに対し、Dice 係数は共起頻度を中心語の頻度と共起語の頻度との関係の中で評価する相対頻度であるため、この問題点が回避されており、コーパスの規模に依存しないという長所につながっている。[8]

石川（2008a:45）は、中條・内山（2004）に言及しつつ、Dice 係数を次のように評価している：「式からわかるように，ダイス係数は単純ではあるが，有用性・妥当性の高い指標で，特徴語検出における 9 種類の統計値の妥当性を比較した中條・内山（2004）によれば，ダイス係数の精度がもっとも良かったことが報告されている。」

Rychlý（2008:7ff.）自身も、the British National Corpus において 6 種の共起指標（T-score、MI-score、

MI^3-score、Minimum Sensitivity、MI log Freq、Dice）を用いた、英語動詞 *break* のコロケーション調査の結果を比較し（p.8 Table 2）、「Dice スコアが非常に良好なコロケーション候補を提示している」と述べている（p.9）。[9]

須永（2011:97）もまた、日本語中古語の「興あり」「甲斐なし」など、「名詞」と「あり／なし／よし／あし」の共起を調査した結果と、『日本国語大辞典』（『日国』）におけるそれらの立項状況を比較した研究において、Dice 係数の有効性を次のように評価している：「しかしダイス係数は，このように単純な割に，実用面でも比較的有用な指標として知られており，他の複雑な指標よりもかえって良好な結果を示すことも多い。実際，今回の調査範囲のうち，実例数も豊富で，『日国』見出し語数も多い「名詞＋なし」の場合を例に，ダイス係数，t スコア，MI スコアの 3 つの指標をもとにコロケーション強度を算出し，比較してみたところ，辞書立項の有無との相関を最も強く示したのはダイス係数であった。」

同様に恒川（2020:110ff.）も、logDice 係数とドイツ語コロケーション辞典における共起語採録状況との間に、緩やかな相関が見られることを指摘した。間接的ながら、これもまた Dice 係数の有効性を示す事実として見ることができる。

3．対数の利便性

Dice 係数のほか、logDice 係数に関わる重要な要素にはもうひとつ、対数がある。では、対数とはどのようなもので、対数を利用することにはどのような利点があるのだろうか？

対数は、航海術やそれを支えた天文学と深く関わり、「複雑なかけ算を、足し算に変換して楽に計算する」（『ニュートン』2019:62）ためには、「等比級数に等差級数を対応させたらよいことに気づいた」（小堀 1989:37）スコットランド人ジョン・ネイピアによって、17世紀初頭に発明された。これを利用すれば、たとえば 1 、10、100、1000、10000のような等比数列（それぞれ10倍で増加）を、0、1、2、3、4のような等差数列（それぞれ1ずつ増加）としてとらえることができる。したがって、たとえば1（=10^0）～100000000000000（=10^{14}；100兆）のような一方が急激に大きくなる数列を、0～14という日常的で扱いやすく、感覚的にもわかりやすい数列に変換して利用できるようになった（いずれも底10の場合；『ニュートン』2019:65）。

対数は、逆に一方が急激に小さくなる数列の変換にも用いられる。たとえば、上例の数を分母とし1を分子とする分数（$1/10^0$～$1/10^{14}$）の場合、その対数は0～ −14である。ちなみに、水溶液の酸性・アルカリ性を示す水素イオン指数（pH）は、これをプラス値に変換して取り出した数値（0～14）に他ならない。この他、調べてみると、音の強さを表すデシベル、音楽の音階、星の等級、地震のマグニチュードなど、科学と交わる日常の様々な分野において対数が関わっていることがわかる。これらの場合いずれも、等比的な数列が、対数の働きにより日常的な感覚にもなじみやすい等差的な数列に変換されている。[10]

4．logDice 係数

では、Dice 係数および対数は、logDice 係数にどのように関わっているのだろうか？

4.1 Dice 係数の「単なる」対数化か

logDice 係数は次の式で定義される： $\text{logDice} = 14 + \log_2 \frac{2f_{xy}}{f_x+f_y}$

すなわち、Dice 係数（$\frac{2f_{xy}}{f_x+f_y}$）を対数化した数値（$\log_2 \frac{2f_{xy}}{f_x+f_y}$）にさらに数14を加えたものが、logDice 係数である。上述のように、「ログ＋ダイス」という名称からも、また Rychlý（2008:9）自身がこの新しい共起指標をそう呼んでいることからも、logDice 係数は単純に「ダイス係数の対数化」と理解されがちである。実際にはしかし、対数化だけでなく、さらに数14が加えられていることに注意が必要である（4.3節以下を参照）。

4.2 Dice 係数の「小ささ」

第2章で見たように、Dice 係数は、コーパスにおける共起語抽出のための基本指標として一定の評価を得ている。他方ではしかし、弱点もいくつか抱えている。そのひとつが、分子（共起頻度 $f_{xy}\times2$）をより大きな数の分母（$f_x + f_y$）で割った値というその定義上、数値が小数点以下の小さな数（0 ≦ Dice 係数 ≦ 1）にならざるを得ないという点である。Rychlý（2008:9）は、これを Dice 係数の「唯一の問題」であると述べている。

表1：英語動詞 *break* の共起語
（Rychlý 2008:8: Table 2のうち Dice 係数［左表］と MI スコア［右表］）

	F_{xy}	Dice
down	2472	0.0449
silence	327	0.0267
into	1856	0.021
leg	304	0.0203
off	869	0.0201
barrier	207	0.0191
law	437	0.0174
up	1584	0.0158
heart	259	0.0155
neck	180	0.0148
news	236	0.0144
rule	292	0.0142
out	1141	0.0135
away from	202	0.0135
bone	151	0.013

	F_{xy}	MI-score
spell-wall	5	11.698
deadlock	84	10.559
hoodoo	3	10.43
scapulum	3	10.324
Yasa	7	10.266
intervenien	4	10.224
preparedness	21	10.183
stranglehold	18	10.177
logjam	3	10.131
irretrievably	12	10.043
Andernesse	3	10.043
irreparably	4	10.022
Thief	37	9.994
THIEf	4	9.902
non-work	3	9.809

実際、上でも触れた Rychlý（2008:8）の Table 2のうち Dice 係数の表（上表1の［左表］として掲出）を見ると、最大値でも0.0449（down）であり、最小値は0.013（bone）である。これらの値は数として小さく、同じ Table 2に含まれている MI スコアの数値（11.698～9.809：上表1の［右表］として掲出）と比べてもわかりにくいという印象を受ける。

4.3 対数化の利点と調整

しかし、「小ささ」それ自体は、特に大きな問題とは言えない。小ささだけの問題であれば、たと

えば「パーセント」のように100倍することでも、一定の適切な数値を得ることができるからである。Rychlý 自身は言及していないが、Dice 係数にはもうひとつの問題が関わっており、対数化の観点から見れば、こちらの方がより本質的ではないかと思われる。それはすなわち、Dice 係数が、一方が急激に小さくなる等比数列1.000000～0.000061をなしているという問題である（4.4節：表3「Dice 係数（小数）」の列参照）。しかし、このような数列こそまさに対数が得意とする対象であり、この問題は対数化によって容易に解決することができる。なぜなら、これにより相対的に変化のなだらかな等差数列0～ -14（底2；表3「Dice の対数」の列参照）を得ることができるからである。

他方しかし、Dice 係数の対数化によっては、もうひとつ別の問題が生じてしまう。すなわち、Dice 係数は値が1より小さく（0 ≦ Dice 係数 ≦ 1）、このような数値の場合、その対数がマイナスの数値（下表2の「Dice の対数」、また4.4節：表3中央「Dice の対数」の列も参照）になってしまうという問題である。したがって、このままではまだ問題の十分な解決にならないだけでなく、数値間の大小の違いなど、かえってよりわかりづらいものになってしまう。

そのため、対数化された Dice 係数にはもう一段の操作が必要で、それが数値14の付加による調整である。これにより初めて妥当なプラスの数値、すなわち logDice 係数0～14が得られ、それとともに数値の大小も明確にすることができる（下表2の「logDice」、また4.4節：表3右端「logDice」の列参照）。

表2：Dice 係数データ（表1の再掲）の対数、logDice 係数への変換

	F_{xy}	Dice		Dice の対数		logDice
down	2472	0.0449	→	-4.477	→	9.523
silence	327	0.0267	→	-5.227	→	8.773
into	1856	0.021	→	-5.573	→	8.427
leg	304	0.0203	→	-5.622	→	8.378
off	869	0.0201	→	-5.637	→	8.363
barrier	207	0.0191	→	-5.710	→	8.290
law	437	0.0174	→	-5.845	→	8.155
up	1584	0.0158	→	-5.984	→	8.016
heart	259	0.0155	→	-6.012	→	7.988
neck	180	0.0148	→	-6.078	→	7.922
news	236	0.0144	→	-6.118	→	7.882
rule	292	0.0142	→	-6.138	→	7.862
out	1141	0.0135	→	-6.211	→	7.789
away from	202	0.0135	→	-6.211	→	7.789
bone	151	0.013	→	-6.265	→	7.735

logDice 係数は結局、対数の利用と調整値14の付加という2段階の手続きにより、Dice 係数の有用性をそのまま引き継ぎつつ、数値としてのわかりやすさを改善した指標であると言えるだろう。

4.4 「＋14」の理由

では、調整値として15や16ではなく、なぜ14が付加されているのだろうか？　結論から言えば、これは、Rychlý 自身が、「Dice 係数の対数が −14であるときに logDice 係数が0になるように定義することが適切」と考えたからだと思われる。

下表３からわかるように、この「logDice 係数が０」とはすなわち、Dice 係数が1/16384の場合である。Dice 係数の定義式の分子は「共起（f_{xy}）×2」であるため、この1/16384は、中心語 X の出現（f_x）と共起語 Y の出現（f_y）の合計回数16384のときに共起（f_{xy}）自体は0.5回であることを意味している。もちろん、実際の共起は整数回でしかありえないため、これはf_x+f_y が32768回のときに共起（f_{xy}）が１回、あるいは65536回のときに２回、等々を意味する。

表３：Dice 係数と logDice 係数の対応関係[11]

Dice（分数表示）		（指数表示）		（小数表示）		Dice の対数		logDice
1 / 1	=	2^{0}	=	1.000000	→	0	→	14
1 / 2	=	2^{-1}	=	0.500000	→	−1	→	13
1 / 4	=	2^{-2}	=	0.250000	→	−2	→	12
1 / 8	=	2^{-3}	=	0.125000	→	−3	→	11
1 / 16	=	2^{-4}	=	0.062500	→	−4	→	10
1 / 32	=	2^{-5}	=	0.031250	→	−5	→	9
1 / 64	=	2^{-6}	=	0.015625	→	−6	→	8
1 / 128	=	2^{-7}	=	0.007813	→	−7	→	7
1 / 256	=	2^{-8}	=	0.003906	→	−8	→	6
1 / 512	=	2^{-9}	=	0.001953	→	−9	→	5
1 / 1024	=	2^{-10}	=	0.000977	→	−10	→	4
1 / 2048	=	2^{-11}	=	0.000488	→	−11	→	3
1 / 4096	=	2^{-12}	=	0.000244	→	−12	→	2
1 / 8192	=	2^{-13}	=	0.000122	→	−13	→	1
1 / 16384	=	2^{-14}	=	0.000061	→	−14	→	0
1 / 32768	=	2^{-15}	=	0.000031	→	−15	→	−1
1 / 65536	=	2^{-16}	=	0.000015	→	−16	→	−2
…	=	…	=	…	→	…	→	…

■表３の Dice 係数の「分数表示」、「指数表示」、「小数表示」は、表示形式が異なるだけで同じ数を表している；「Dice の対数」の底は２；logDice は「14＋ Dice の対数」

この数値1/16384（＝0.000061）の示す共起の可能性はすでに十分に小さい。しかし、中心語 X と共起語 Y との間に共起がある限り（f_{xy}≧1）、Dice 係数はさらにいくらでも小さくなりうる（表３最下行「1/32768、1/65536、…」参照）。そのような場合、Dice 係数は徐々に０に近づいていくが、しかし共起がある限り必ず0より大きく、0になることはない。

そのため、Rychlý は logDice 係数を定義するにあたり、共起頻度が十分小さく、事実上０と見なせるような Dice 係数としてこの1/16384（対数値 −14）を選択し、それ以下（−15、−16など）は統

計的に無視してよいと考えたのだと思われる。[12]

この Dice 係数の最小値1/16384に対する対数は、マイナスの数「−14」である。したがって、定義式に調整値としてプラスの数値「+14」を与えておけば、この場合に logDice 係数として「共起なし」を意味する数値「0」が得られることになる。

それでは、logDice 係数の 0 に関して、なぜ以上のような回り道的な操作が必要になるのか？　それは、対数が存在するのが定義上、正の実数に対してだけであり、数 0 には対数がありえないという本質的な理由からである。すなわち、本来、「共起なし（$f_{xy}=0$）」を表す logDice 係数 0 は、定義式（4.1節参照）に基づき Dice 係数から直接導き出せることが望ましい。しかし、「共起なし（$f_{xy}=0$）」の場合、その Dice 係数（$\frac{2\times0}{f_x+f_y}$）は 0 となり、0 には対数が存在しないため、定義式では logDice 係数を計算することができない。つまり、logDice 係数の定義式からは、「完全共起（係数14）」[13] の対極となる「完全に共起なし」を導き出すことができない。そのため、近似的に「ほぼ共起がない（$f_{xy}\fallingdotseq0$）」を意味する Dice 係数の最小値1/16384で代えることが必要になるのである。

4.5　数値差が表す関係

最後に、logDice 係数の数値差が表す関係について若干の補足をしておきたい。対数は等比数列を等差数列に変換する仕組みであるため、等差数列の数値差が表す意味は、対数化前の等比数列における数値間の関係に立ち戻って考えることが必要である。具体的に logDice 係数（等差数列）の場合を考えるならば、その数値差の意味は、前節の表 3 を右から左に逆にたどり、同じく等差数列である「Dice 係数の対数」を経由して、対応する等比数列である Dice 係数の数値間の関係を見ることによって明らかになる。

表 4 ：等差数列（logDice 係数）の数値差と等比数列（Dice 係数）の関係

数値差	logDice	Dice の対数		Dice	関係の比率
1	（例）10と9	−4 と −5	⇔	1/16と1/32	2（$=2^1$）倍
2	（例）10と8	−4 と −6	⇔	1/16と1/64	4（$=2^2$）倍
7	（例）10と3	−4 と −11	⇔	1/16と1/2048	128（$=2^7$）倍

上表 4 は、logDice 係数（等差数列）の数値差とそれに対応する Dice 係数（等比数列）の数値差の関係を、例示したものである。この表の内容は表 3 が示しているものと同じだが、わかりやすさため左右逆に、logDice 係数を起点とし、右に Dice 係数を配してある。この表から、前者の数値差1が後者では 2 倍（$=2^1$倍）、数値差2が4倍（$=2^2$倍）、数値差7が128倍（$=2^7$倍）というように、それぞれ「2 の‘数値差’乗倍」の関係になっていることがわかる。[14]

以上は、差が切りのよい整数の場合である。コーパスデータとして示される実際の logDice 係数は、DWDS-Wortprofil では小数点以下 1 位まで、国語研究所の NLB では同 2 位まで計算されており、小数点以下の数が含まれている。もちろん、このような場合でも、数値間の関係は上と同じ「2 の‘数値差’乗倍」である。たとえば、数値差0.1は約1.0718倍（$=2^{0.1}$倍）、数値差0.01は約1.00696倍（$=2^{0.01}$倍）を意味する。

ちなみに、logDice 係数（本来は Dice 係数）において等比数列を等差数列に変換するときの対数の底が2であるのは、この場合の数列が、等比数列ではあっても比較的変化の小さな数列1（$=2^{0}$）～1/16384（$=2^{-14}$）であるためである。これに対し、極めて大きく変化する等比数列、たとえば水素イオン指数（pH）の1（$=10^{0}$）～1/100兆（$=10^{-14}$）のような場合、同じく妥当な等差数列0～14を得るためには底に10を採用することが適切となる。[15]

5．まとめ

以上のように、logDice 係数は、Dice 係数の長所を残しつつ数値としての問題点を改善するために、対数の利用と調整数14の付加が行われている。その結果得られた数値0～14は妥当でわかりやすく、日常的な使い勝手も元の数列0～1/16384に比べはるかによい。logDice 係数は、このような一見シンプルであるが合理的な補正を加えることにより、Dice 係数の本質を最大限に引き出しつつ使い勝手をよくした共起指標となっている。

注

1　2009年当時の DWDS パネルでは、中核コーパスに関して他の複数の共起指標も利用可能であった（今道・恒川 2009:161、今道 2009:251f.「注 3 」）。現在の Wortprofil ではしかし、logDice 係数および頻度だけになっている。

2　2006年9月という日付は logDice 論文の発表年（2008年）に先行しており、つじつまが合わない。今道（2009:251f.）が「注 3 」の中で、Log-Likelihood ratio、t-score、MI-score のほか、Sketch Engine が「MI^{3}-score、MI-log-prod、minimum sensitivity などの指標も実装」と説明していることからも、思い違いである可能性が高い。いずれにせよ、現在では Sketch Engine も logDice を採用している（赤瀬川ほか 2016:70）。

3　この2019は、「更新履歴」の日付「2019/12/16」による。

4　執筆にあたっては鈴木孝典・元東海大学教授（アラビア科学史）に多大なご教示をいただいた。ここに記して感謝を申し上げたい。

5　Dice 係数は、もともと Dice（1945）において「異なる生物種の生態学的共起の測定指標」として提案されたものである。

6　この最大値を 1 に調整するという解釈（Dice = $\{f_{xy} / (f_x + f_y)\}\times 2$）だけでなく、「×2」には他の解釈もあり得る。たとえば、次の（1）の解釈は「共起頻度（f_{xy}）」が「X の出現頻度（f_x）」の一部でも「Y の出現頻度（f_y）」の一部でもあると考えるもの、また（2）は「X の出現頻度（f_x）と Y の出現頻度（f_y）」の平均を取ると考えるものである：
（1）Dice = $(f_{xy} + f_{xy}) / (f_x + f_y)$
（2）Dice = $f_{xy} / \{(f_x + f_y) / 2\}$

7　作図に際しては，㈱ Faber Company のサイト「MIERUCA-AI【技術解説】」の「集合の類似度（Jaccard 係数，Dice 係数，Simpson 係数）」（https://mieruca-ai.com/ai/jaccard_dice_simpson/）にある Dice 係数の図を参考にした。

8　この長所は、そのまま logDice 係数へと引き継がれている。Rychlý（2008:9）は logDice 係数の特徴・特色を 4 点にまとめており、これをその第 4 点として挙げている。

9 ただし、Rychlý（2008:8）の Table 2 に挙げられた 5 種類の表を見る限り、動詞 break の共起語の「良好な候補」を示していると考えられる指標は Dice 係数だけではない。Minimum Sensitivity、MI log Freq もまた、順位は別にして類似の共起語を抽出しており、前者との違いは15語中 3 語、後者とは 2 語に過ぎない。

10 『ニュートン』2019年 7 月号参照。水素イオン指数（pH）に関しては p.54f. 参照。水素イオン指数も logDice 係数も、数列が共に0〜14の範囲に設定されているのは、日常的な使い勝手との関連で示唆的であるように思われる。

11 本表 3 では、「指数」（累乗を表す右肩の数）が底 2 の整数倍になる場合のみを取り上げている。これはわかりやすさを優先したためで、実際の指数は「0 〜 −14」のいずれの実数でもありうる。Dice 係数 2^{-4}〜2^{-5}、すなわち logDice 係数10〜9の場合を例にとれば、整数間のより細かな数値は、たとえば次表 5 のようになる。

表５：表３（Dice 係数と logDice 係数の対応関係）の補足

Dice（分数表示）		（指数表示）		（小数表示）		Dice の対数		logDice
1 / 16	=	2^{-4}	=	0.062500	→	−4	→	10
1 / 17	=	$2^{-4.09}$	=	0.058824	→	−4.09	→	9.91
1 / 18	=	$2^{-4.17}$	=	0.055556	→	−4.17	→	9.83
1 / 19	=	$2^{-4.25}$	=	0.052632	→	−4.25	→	9.75
1 / 20	=	$2^{-4.32}$	=	0.050000	→	−4.32	→	9.68
1 / …	=	…	=	…	→	…	→	…
1 / 32	=	2^{-5}	=	0.031250	→	−5	→	9

12 この点に関し、Rychlý（2008:9）は次のような説明を与えている：「0は16000 X または 16000 Y あたりの XY の共起が1未満であることを意味する。数値がマイナスである場合、私たちは、XY コロケーションの統計的な意味がないと言うことができる。」（logDice 係数の 4 特徴・特色の 2 点目；16000は正確には16384 = 2^{14}）。

ここにある「共起が1未満」は、とりもなおさず共起（f_{xy}）が0.5回に相当する場合、すなわち Dice 係数が（事実上 0 とみなすことのできる）1/16384である場合のこと指していると思われる。

また、後半の「数値がマイナスである場合」とは、logDice 係数が -1（Dice 係数1/32768）、-2（同1/65536）などとなる場合のことであろう（表 3 最下行参照）。統計的に意味のある Dice 係数の下限値として1/16384を採用したのであるから、それ未満は当然「統計的に意味がない」ことになる。

13 Rychlý（2008:9）は、logDice 係数の 4 特徴・特色の 1 点目として、最大値14について次のように述べている：「理論的な最大値は14である。これは、X のすべての生起が Y と共起し、Y のすべての生起が X と共起する場合である。一般的には、値は10より小さい。」

14 これをさらに一般化すれば、等比数列を等差数列に変換するときの対数の底を a、等差数列の数値差を x とするとき、x が等比数列では a^x 倍の関係になる。なお、Rychlý（2008:9）は数値差を「プラス x ポイント」という言い方で表し、「プラス 1 ポイントが 2 倍」、「プラス7ポイントが約100倍」という 2 つの例を挙げている（logDice 係数の 4 特徴・特色の 3 点目；100倍は正確には128倍＝2^7倍）。

15 『ニュートン』（2019:54f.）参照。仮に前者が底10を、また後者が底 2 を採用したとすると、対

数化された等差数列はそれぞれ0～4.2、0～46.5となる。これらは0～14に比べあまり妥当な数列とは思われない。

参 考 文 献

赤瀬川史朗ほか（2016）：日本語コーパス活用入門．大修館．

中條清美・内山将夫（2004）：統計的指標を利用した特徴語抽出に関する研究．関東甲信越英語教育学会紀要18，99-108．

Dice, Lee R.（1945）: Measures of the Amount of Ecologic Association Between Species. in: Ecology 26/3, 1945, 297-302.
https://www.semanticscholar.org/paper/Measures-of-the-Amount-of-Ecologic-Association-Dice/23045299013e8738bc8eff73827ef8de256aef66

石川慎一郎（2008a）：コロケーションの強度をどう測るか── ダイス係数，t スコア，相互情報量を中心として──．言語処理学会第14 回大会チュートリアル資料，40-50．

石川慎一郎（2008b）：英語コーパスと言語教育．大修館，2008．

石塚（1952）：種類間の生態的結合の測り方（Rice（1945）論文の批判的紹介）．生態学会報1952年１巻４号，218-219．

今道晴彦（2009）：ドイツ語学習者のためのコロケーション抽出に向けて── 統計学的指標の有用性──．日本独文学会「ドイツ文学」138，250-271．

今道晴彦・恒川元行（2009）：DWDS コーパスの概要と利用法．九州大学大学院言語文化研究院研究会「言語科学」44，147-166．

小堀憲（1989）：ネイピア．「数学セミナー増刊　100人の数学者」．日本評論社，1989，37-39．

Lexical Computing Ltd.（2015）: Statistics used in the Sketch Englne. July 8, 2015.
https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf

『ニュートン』2019年７月号「数に強くなる　対数・指数・巨大な数」、24-69、ニュートンプレス．

NLB ユーザマニュアル　バージョン1.40（2016/12/12）
http://nlb.ninjal.ac.jp/site_media/pdf/NLB.manual.v.1.40.pdf

李在鎬（2015）：コーパス研究が切り開く新しい日本語教育．第17回 BATJ 大会基調講演，BATJ Journal No.16，63-76．

Rychlý, Pavel（2008）: A Lexicographer-Friendly Association Score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 6–9, 2008.
https://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf

杉浦正利（2019）：共起スコア：logDice.
http://sugiura-ken.org/wiki/wiki.cgi/exp?page=logDice

須永哲矢（2011）：コロケーション強度を用いた中古語の語認定．国立国語研究所論集２，91-106．

恒川元行（2020）：コロケーション辞典における名詞記述の比較と分析．井口靖ほか：ドイツ語基礎語彙のコロケーションに基づく意味分析とその独和辞典記述方法の検討── 理想の独和辞典を目指して──．科学研究費補助金（基礎研究（C）2016～2019年度 課題番号：16K02667）報告書．107-144．