言語処理学会第 14 回大会チュートリアル資料，40-50．iskwshin@gmail.com 2008/03/17

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―

石川　慎一郎
神戸大学　国際コミュニケーションセンター／国際文化学研究科
E-mail:　iskwshin@kobe-u.ac.jp

１．語から句へ

言語研究において，語はしばしば分析の基礎単位となる。特定の語の頻度を数えたり，異なるコーパス間で語の頻度を比較して特徴度を探ったりするなど，語を中心にして幅広い研究が展開されている。とくに英語の場合，語は左右のスペースによって物理的に切り分けられているため，機械処理や自動処理との相性も良い。

しかし，最近の研究では，語という単位で言語を見ることの限界が次第に意識されるようになってきた。語彙意味論学者である Stubbs（2002）が，著書において「語から句へ」という立場を打ち出したのもこうした流れの中にある。語単位の限界としてはさまざまな点が想起できるが，意味・用法判別ができないことと，複数語からなる語結合を扱えないことの 2 点が特に重要な問題である。前者は語より小さい単位に関わり，後者は語より大きい単位に関わる。

まず，前者について考えてみよう。あるテクストで minute という語が一定の頻度で出現したとする。しかし，その中には次のような事例が混在して含まれているかもしれない。

（１）three minutes（3 分）
（２）have a minute（ちょっと時間がある）
（３）take minutes of the meeting　（会議録を取る）
（４）minute grains of salt（細かい塩の粒）
（５）a minute checkup（詳細な検査）

上記の 5 例において minute の意味や品詞はまちまちであり，われわれの通常の感覚では，これらを 1 つに括り，minute の頻度として要約してしまうことには抵抗を覚える。つまり，語という単位を取る以上，語の下位区分としての意味や用法は扱えないことになる。

次に後者について考えてみよう。本稿で扱おうとするコロケーションは後者に関わるものである。

（６）out of blue （突然に）
（７）kick the bucket（死ぬ）

上記の場合，out や blue という語と「突然に」という意味の間には直接的な関係はない。同じように，kick や bucket と「死ぬ」という意味も乖離している。つまり，out of blue の生起を out と of と blue の生起と読み替えたり，kick the bucket の生起を kick と the と bucket の生起と読み替えたりすることは本質的に不可能ということになる。しかし，語という単位にこだわる限り，語を超えるこうした結合を適切に扱うことはできない。

ここで，「語から句へ」という視点の転換を行うことは，語に関わる諸問題を解決する上で有益である。たとえば，minute という語は単独で存在しているのではなく，three minutes，minute grains，minute checkup などの句の形で存在しているのだと考え，同様に，out of blue や kick the bucket というかたまりが独立した句の形で存在しているのだと考えれば，さまざまな語結合をはるかに適切に扱うことが可能になる。

２．句と共起

しかし，言語分析という観点から見ると，句という概念は客観的な処理の枠組みに乗りにくいものである。*Longman Dictionary of Contemporary English* を見ると，句（phrase）は「複数の語が結合して特定の意味を持つもので，とくに数語の単位で意味をよく言い表すもの」と定義されている。つまり，「特定の意味を持つ」ことが句の成立要件ということになるが，意味は高度に抽象的な概念であって計量化がしにくい。

だとすれば，言語作用の結果として生成された意味に裏打ちされた句ではなく，むしろ言語作用を引き起こす原因態としての現象，つまり，複数の語が並んでいるという現象そのものを研究対象にしたほうが合理的な分析が可能になるように思える。

単語共起	→	共起形	→	意味
（現象／原因）	→	（表現）	→	（結果）

図 1. 意味生成のモデル

ここで，複数の語が並ぶ現象を指して広く共起（co-occurrence）と呼ぶ。共起という概念を採用することで，意味の問題をいったん括弧に入れて分析を始めることができる。

３．コロケーションとはなにか

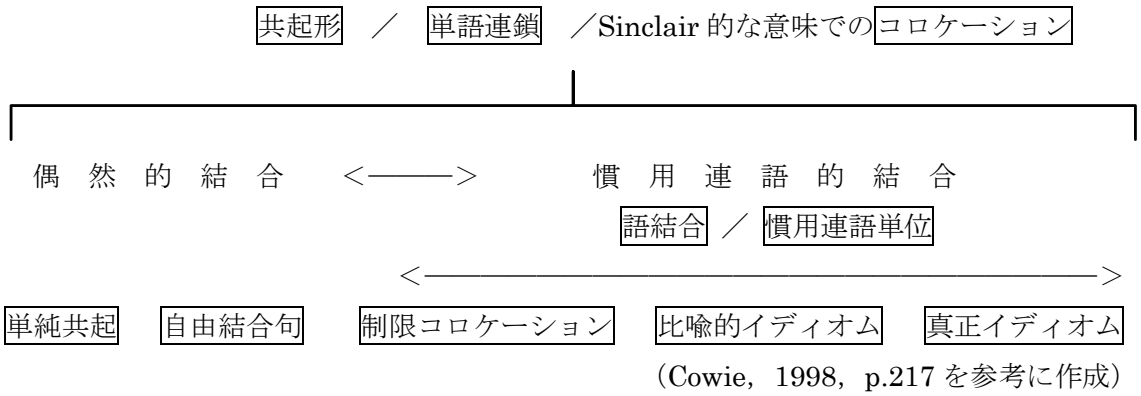
共起現象の実際の現れとしてコロケーション（collocation）がある。Sinclair（1991）も言うように，広義のコロケーションとは語と語が作るあらゆる結合態のことであるが，た

とえば is a pen という 3 語連鎖をコロケーションと呼べるかどうかは疑問も残る。この意味で，Firth（1957）で示された慣用的共起という考え方がコロケーションの議論の出発点になるであろう。

Kjellmer（1991）は，コロケーションの検出にあたって，連続性・頻度・統語関係という 3 つの尺度を考えた。語と語が連続して共起し，その共起形が高い頻度で出現し，かつ文法的にまとまりのあるかたまりとなっている（is a pen はまとまりのあるかたまりではない）場合にコロケーションであるとしたのである。

ただし，Kjellmer の基準を満たしているものの中にも，直観的に言って，「意味のある結合」と「意味のない結合」とが混在している。たとえば，global warming（地球温暖化）という結合はある種の専門用語で特別なものであると思えるが，spring warming（春になって暖かくなること）ならばさほど特別な結合とは思えない。では，直観が教える「意味がある結合」と「意味がない結合」の差はどのように同定できるのであろうか。

ここで重要なことは，コロケーションというのは 1 か 0 かの 2 値的概念ではなく，ゆるやかな層的・段階的概念となっている点である。コロケーションの段階性についてはさまざまな研究者がモデルを提唱しているが，Cowie（1998）の枠組みを大幅に拡張すると，およそ次のような概念図が得られるであろう。図において，左側の方向に行けばいくほど，語の結びつきは自由で交代可能で，共起形としての統語的・意味的自立性は低い。いっぽう，右側の方向に行けばいくほど，語の結びつきは制約され，語要素の交代が許されなくなり，かつ独立した統語的地位と意味を獲得するようになる。なお，Cowie 自身は単純共起や自由結合句については必ずしもはっきり言及していないため，下記には論者の解釈が多分に含まれている。



（Cowie，1998，p.217 を参考に作成）

図 2. 共起モデル

連続する語と語が共起してできた共起形（単語連鎖，ないし Sinclair 的な意味でのもっとも広義のコロケーション）にはいくつかのタイプがある。まず，is a pen のように文法的なまとまりをもたないものを単純共起とする。次に，go to the store のように文法的なまと

まりは持つものの，それ自身で特別な意味を持たず，文法規則からそのつど創造的に組み合わせて作られた言語表現の 1 つにすぎないものを自由結合句（あるいは自由コロケーション）と呼ぶ。

制限コロケーション（これを狭義の「コロケーション」と呼ぶ研究者も多い）とは，イディオムと自由結合句の間に存在するものである。Cowie（1998）は meet the demand（要求を満たす）という例をあげている。この場合，demand の意味は単独の場合と同じであるが，meet は demand によって慣用連語（phraseology）的に拘束されており，その意味は比喩的に変化している。つまり，制限コロケーションとは，意味変化を伴わない自由結合句の特徴と意味変化を伴うイディオムの特徴を両方持ち合わせたものと言える。

制限コロケーションにおける意味が，より元の意味から乖離・自立し，文脈が不透明になると，イディオムになる。このうち，真正イディオムとは spill the bean（秘密を漏らす）のように，個々の語から共起形の意味がまったく理解されないものである。いっぽう，比喩的イディオムとは，blow off steam（鬱憤を晴らす）のように，新しい自立的意味を獲得してはいるものの，その意味が「蒸気を噴出する」という個々の語の意味の総和から比喩的に拡張されたことが自明であるようなものを言う。

このように，コロケーション，句，共起，イディオムといった概念は時にオーバーラップしており，研究者による用語の定義の不一致も相まって議論はいたずらに難解になりがちであるが，ここでは上記のモデルに基づき，コロケーションを制限コロケーションとして扱いたい。すなわち，１）語と語の連続した結びつきであり，２）一定の文法的まとまりをもち，３）その結びつきは偶然性・新規性・創造性を超えた強度を持つが，４）結びつき自身がまったく新しい意味を担うイディオムとは異なり，結果として５）自由結合句とイディオムの中間領域に存在するゆるやか語結合に注目することになる。

４．コロケーションの検出

以上で見たようにコロケーションとは絶対的な概念ではなく，自由結合句とイディオムの間に帯的に存在する概念である。コロケーションとしての強度が弱く自由結合句に近いコロケーションもあれば，逆に強度が強くイディオムに近いコロケーションもありうる。では，コロケーションの強度を計量し，制限コロケーションを検出するにはどのようにすればよいのであろうか。

コーパス言語学でコロケーションを扱う場合，客観的処理を重視する視点から，しばしば頻度データに着目する。既に述べたように，global warming と spring warming の間には質的な差異が直観レベルで感知されるわけであるが，この差異をコーパス言語学では頻度から考えようとするのである。

本稿では，以下，コロケーション研究に広く用いられるダイス係数，t スコア，相互情報量という 3 種類の指標について簡単に見てゆくこととする。なお，コロケーション強度を

示す統計値には他にも各種あるので，詳しくは石川（2006），石川（近刊）ほかを参照されたい。

ところで，共起形を構成する中心語としての warming と，その共起語である global ないし spring のコロケーションを考える場合，ふつうは，global warming または spring warming という共起頻度（共起形頻度）のみに注目しがちである。

しかし，これだけではコロケーション強度を正確に測ることはできない。あわせて，中心語の単独頻度，共起語の単独頻度，共起頻度，それにそれらを含むコーパス全体の総語数を見る必要がある。

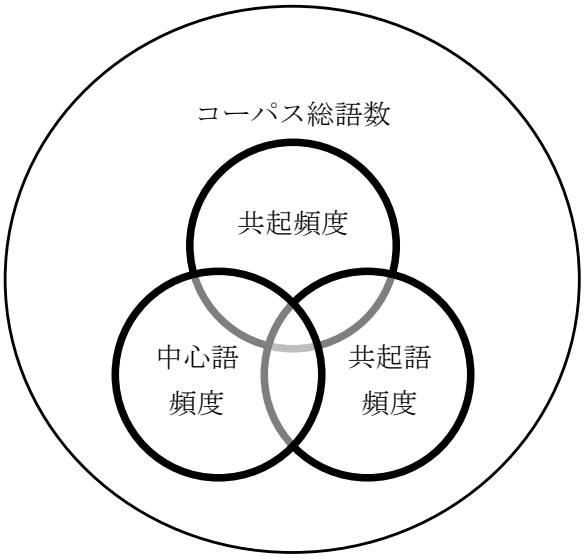


図 3. コロケーション強度計算のイメージ

ここで，コーパス総語数を見る必然性について疑問が生じるかもしれないが，頻度とは，そもそもすべて特定の母集団の中で決定されるものである。たとえば，漠然と the の頻度を言うことはできない。1 億語のコーパスの中での the の頻度，100 万語のコーパスの中での the の頻度というように，すべての頻度は母集団の全体に占める比率として解釈されるのである。このため，共起頻度・中心語頻度・共起語頻度と同時に，コーパス総語数を見る必要が生じるのである。

４．１．ダイス係数

ダイス係数は，情報理論やウェブサイトからの関連語自動抽出のアルゴリズムなどとしても広く使用されている。ダイス係数の計算にはコーパス総語数は必要ではなく，中心語頻度と共起語頻度の関係だけで 2 語のコロケーション強度を計測してゆく。

ダイス係数は，共起頻度を中心語頻度と共起語頻度の和で割って 2 倍した値で，式は次のようになる。

$$D = 2 \times \frac{\text{共起頻度}}{\text{中心語頻度} + \text{共起語頻度}}$$

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。ダイス係数は常に正の値を取るが，頻度が低い場合は著しく小さい値を返す可能性がある。

４．２．　t スコア

t スコアは，コーパス総語数を考慮し，全体における個々の語の出現比率を比較する。実際には，語の頻度情報を重視するため，高頻度語が高く評価されがちで，頻繁に用いられる一般性の高いコロケーションの評価に適している。

t スコアでは，多くの有意差検定手法と同様，期待値と実測値の差の大きさが問題になる。計算式は下記のとおりである。

$$t = \left(\text{共起頻度} - \frac{\text{中心語頻度} \times \text{共起語頻度}}{\text{コーパス総語数}}\right) \div \sqrt{\text{共起頻度}}$$

一般に，t スコアは 2 以上の場合に意味のある組み合わせであるとされる（Hunston, 2002, p. 72）。

４．３．相互情報量

相互情報量（mutual information score；略して MI Score とも呼ばれる）は，ある語が共起相手の語の情報をどの程度持っているかを示す指標である（Oakes, 1998, pp. 63-65）。その語が出れば，共起相手は自動的に決まる，というような場合に相互情報量は高くなる。相互情報量は，概念的には，2 語の頻度の実測値と期待値の比を対数に誘導したもので，式は下記のようになる。

$$I = \log_2 \frac{\text{共起頻度} \times \text{コーパス総語数}}{\text{中心語頻度} \times \text{共起語頻度}}$$

底を 2 とする対数を取ることで，頻度情報は圧縮され，高頻度であることのウェイトは小さくなる。このため，相互情報量では，頻度は低いが特殊な結びつきをしているコロケーションがうまく検出できるとされている。

５．コロケーション検出の実験

それでは，以上の手法を使うことで，果たして意味のある制約コロケーションがコーパスから検出できるかどうか実験してみることとしよう。ここでは，1 億語の大型コーパスである British National Corpus を用い，「～warming」という形のさまざまなコロケーションの強度について考えてみたい。

BNC に付与されたタグ情報を利用し，名詞の warming の直前に出現する共起語を検索すると全部で 43 語が得られるが，ここでは，頻度 2 以上の 10 語を検証対象とする。この中には the や a などの冠詞，また，that なども含まれているため，いわゆる単純共起のようなものも混在しているが，すべて一律に測定を行う。分析の基礎になるデータは下記の通りである。

表 1. X + warming のコロケーション頻度

	共起頻度	中心語頻度	共起語頻度	コーパス総語数
global	599	1090	3524	111173004
the	21	1090	6045331	111173004
a	9	1090	2138370	111173004
stratospheric	5	1090	93	111173004
that	3	1090	1085726	111173004
further	2	1090	35786	111173004
greenhouse	2	1090	997	111173004
rapid	2	1090	3552	111173004
spring	2	1090	5822	111173004

それぞれのコロケーションについて，各指標値を求めた結果は次の通りである。

表 2. コロケーション強度指標値

	共起頻度	ダイス	tスコア	相互情報量
global	599	0.26	24.47	14.08
the	21	0.00001	-8.35	-1.5
a	9	0.00001	-3.99	-1.22
stratospheric	5	0.0085	2.24	12.4
that	3	0.00001	-4.41	-1.83
further	2	0.00011	1.17	2.51
greenhouse	2	0.0019	1.41	7.68
rapid	2	0.00086	1.39	5.84
spring	2	0.00058	1.37	5.13

共起頻度によると，global に続くコロケーションは，the，a，stratospheric，that の順

になっていたわけであるが，統計的指標を使えば，the や a や that などの共起形の順位はいずれも最も低くなる。とくに，コーパス総語数を考慮している t スコアや相互情報量は，頻度を総語数に対する比率でとらえることになるため，これらの機能語との共起形にはすべて負の値が返されており，直観が示唆する「意味のある結合」と「意味のない結合」が客観的に判別できていることになる。

また，どの統計指標によっても，10 種類の共起形の中でもっともコロケーション強度が強いものは global warming で，2 番目が stratospheric warming（成層圏温度上昇），3 番目が greenhouse warming（温室効果による温暖化）となる。これらはいずれも自然科学分野のタームとなっており，特定の意味のある結びつきをした例である。一方，値が低い further warming（さらなる温度上昇），rapid warming（急激な温度上昇），spring warming などは，通常の統語規則の運用から生まれた表現であり，コロケーションというよりも自由結合句的な性質が強い。

以上のことから，3 つの統計値は，単純共起的な事例をはじくのみならず，形容詞（ないし名詞の形容詞的用法）と名詞のさまざまな組み合わせの中でも，「意味のある結合」と「意味のない結合」を判定する機能を果たしていると言える。

５．おわりに

本稿では，句・コロケーション・自由結合句・イディオムなどの関連する概念を整理したあと，コロケーション強度を測る指標として，コーパス研究においてもっともよく使われるダイス係数，t スコア，相互情報量の 3 つの統計値について概観した。

BNC から得られたサンプル・データを使った実験の結果，単純頻度では機能語などを含む共起形が上位に位置づけられるが，統計値を使用することで，「意味のある結合」と「意味のない結合」を一定の精度で判別できる可能性が示唆された。

冒頭でも述べたように，コロケーション研究については，概念のオーバーラップが特に問題となる。Cowie（1998）は，コロケーションの上位概念として慣用連語という言葉を用いた上で，次のように指摘している（引用は拙訳による）。

> 慣用連語を範疇に分類することは，周知のように困難である。というのも，慣用連語には，統語論・語用論・文体論・意味論に至るまで，多様な変数が常に関わってくるからである。言語分析者にとって重要な問題とは，まず，慣用連語という範疇の始点を定め，その中で，より構成要素が不変的で意味が不透明な項目と，より構成要素の組み換えが容易で意味が透明な項目とを区分することである。以上の問題を解決する上で必要なことは，適切な基準を選定し，慣用連語の分類範疇の枠組みを構築することである。（p.210）

Cowie の研究においては，統計的アプローチはほとんど言及されていなかったが，本稿で紹介したような統計的指標を用いることで，関連する諸概念間の曖昧なボーダーラインを改めて正確に引きなおすことができるかもしれない。

今後，情報系・工学系の研究者と言語系の研究者の協働が盛んになり，直観と客観を融合した新しいタイプのコロケーション研究が広がってゆくことを期待したい。

参考文献

Barnbrook, G. (1996). *Language and computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

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.

中條清美・内山将夫・長谷川修治 (2005). 統計的指標を利用した時事英語資料の特徴語選定に関する研究. 『英語コーパス研究』, *12*, 19-35.

Cowie, A. P. (1998). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press.

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

Firth. J. R. (1957). Modes of meaning. In J. R. Firth. *Papers in Linguistics. 1934-1951* (pp. 190-215). London: Oxford University Press

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

石川慎一郎 (2006). 言語コーパスからのコロケーション検出の手法：基礎的統計値について. 『統計数理研究所共同研究レポート』, *190*, 1-14.

石川慎一郎 (近刊). 『データとしてのテクスト』(仮題). 東京：大修館書店.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kjellmer, G. (1991). A min of phrases. In K. Ajimer & B. Altenberg (Eds.), *English corpus linguistics; Studies in honour of Jan Svartvik* (pp. 111-127). London: Longman.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Pearson Education.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. London: Language Teaching Publications.

松野和子・杉浦正利 (2004). コロケーションの定義：コロケーションの概念と判定基準に

関する考察.『なぜ英語母語話者は英語学習者が話すのを聞いてすぐに母語話者ではないとわかるのか：平成 13 年度～15 年度科学研究費補助金基盤研究(C)(2)研究成果報告書』, 79-96.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

齋藤俊雄・中村純作・赤野一郎(編) (2005).『英語コーパス言語学：基礎と実践』（改訂新版）. 東京：研究社出版.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1-24). Amsterdam: Benjamins.

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The PSTI report*. London: Continuum.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Malden, MA: Blackwell Publishing.

杉浦正利 (2001). 「コーパス解析手法」 http://sugiura3.gsid.nagoya-u.ac.jp/project/ouyougengogaku/

Sugiura, M. (2002). Collocational knowledge of L2 learners of English: A case study of Japanese learners. In T. Saito, J. Nakamura, & S. Yamazaki (Eds.), *English corpus linguistics in Japan* (pp. 303-323). Amsterdam: Rodopi.

上田尚一 (2003). 『質的データの解析：調査情報の読み方』. 東京：朝倉書店.