

ビジネス英語のコロケーション分析

「相互情報量」(MI) と「仕事量基準」(CC) による分析とその比較

国際ビジネスコミュニケーション学会第65回全国大会
2005年10月16日
中央大学多摩キャンパス

青山学院大学 英米文学科
染谷 泰正

someya@cl.aoyama.ac.jp

<http://www.cl.aoyama.ac.jp/~someya/>

発表要旨

- ✦ 「コロケーション」とは「語と語のシステマチックな結びつき」を指し、自然で流暢な発話を構成するための重要な要素である。しかし、外国語の場合、かなり上達したレベルの学習者であっても対象言語のコロケーションを母語話者のように使いこなすことは困難であり、そのために学習者の発話あるいは文章は、常に何がしかの「不自然さ」を伴うのが通例である (Pawley and Syder 1983; Kjellmer 1991; Shei and Pain 1999; 染谷 2001)。
- ✦ これは、いわゆる「ビジネス英語」についても同じである。本研究ではこの「コロケーション」に焦点を当て、その抽出方法について検討する。まず、100万語の「ビジネスレターコーパス」(Someya 1999) の中で見られるコロケーションを n-gram 分析の手法を使って「結合頻度」を基準に抽出し、次に、これを「相互情報量」(Mutual Information = MI) および「仕事量基準」(Cost Criteria = CC) という2つの統計手法を使って再分析し、その差を比較した。分析の結果、教育的見地から見て有意なコロケーションの抽出には CC の方が有効な手法であることが確認された。

Someya, Y. (2005). Collocational Analysis of Business English – A Comparison of Two Statistical Methods: Mutual Information (MI) and Cost Criteria (CC)

発表概要

1. コロケーションとは何か
 - 1.1 コロケーションの定義
 - 1.2 第二言語習得におけるコロケーションの重要性
2. ビジネス英語 (BE) のコロケーション
 - 2.1 「ビジネス・ジャーゴン」 (BJ) の位置づけ
 - 2.2 代表的ビジネスジャーゴンの使用分布比較
3. 大規模コーパスからのBE コロケーションの抽出
 - 3.1 コロケーションの統計的な抽出方法について
 - 3.2 相互情報量 (Mutual Information)
 - 3.3 仕事量基準 (Cost Criteria)
 - 3.4 分析結果と考察
4. まとめと今後の課題

1. コロケーションとは何か (1/2)

1.1 コロケーションの定義

- ⊕ 広義には「語と語の共起関係」(co-occurrence) をコロケーションと呼ぶが、Kjellmer (1991) や Lewis (1997) は単に語が共起しているだけではコロケーションとはみなさず、潜在的なコロケーション (potential collocation) にすぎないと主張している。
- ⊕ Kjellmer (1991) は、「繰り返し出現する語の連続 (recurring sequences)」で「文法的構造」(grammatical structure) をもったものをコロケーションと呼び、同様に、Lewis (1997) は (1) 頻度 (2) 統語的つながり (3) 連続性、の3つをコロケーションの要素としている。
- ⊕ Benson et al. (1997) は、その上で、コロケーションを (1) 文法的コロケーション (grammatical collocation) と (2) 語彙的コロケーション (lexical collocation) の2種類に分類し、これ以外の語連結を自由語彙連結 (free lexical combination) と定義している。(1) は主として「内容語 + 機能語」の組み合わせ (e.g. decide on, look forward to, inform (someone) of, would appreciate, etc.) を指し、(2) は「内容語 + 内容語」の組み合わせ (e.g. sincerely yours, thank you, make every effort, etc.) を指す。

1. コロケーションとは何か (2/2)

1.2 第二言語習得におけるコロケーションの重要性

- ✦ 近年、第二言語習得における語彙学習の重要性が改めて見直されてきているが、自然な言語運用のためには個々の語の意味を覚えるだけでなく、単語動詞の意味的なつながりを持つ連語関係 = コロケーションの意識的な学習が必要であることが指摘されている。

“A very important part of learning a new word is learning what words it goes with.” (Nation 1990)

- ✦ コロケーションは “one of the most powerful forces in making language coherent, fluent, comprehensible, and predictable” (Hill and Lewis 1997) であり、言語を運用する際に重要な役割を果たす。
- ✦ 言語使用者は「半ば規格化された語句」を使用し、すでに記憶している語句や定型表現、あるいはその一部の置き換えを行って文を生成している。

“A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choice, even though they might appear to be analysable into segments.” (Sinclair 1991)

2. ビジネス英語のコロケーション

2.1 「ビジネス・ジャーゴン」(BJ) の位置づけ

- ✦ ビジネス英語 (BE) における代表的なコロケーションとしては、いわゆる「ビジネスジャーゴン」(BJ) が挙げられる。
- ✦ BJ は、60年代から70年代において、BE を特徴付けるものとして盛んに研究が行われてきた。本学会の前身である「商業英語学会」の会報でも吉田 (1964) や碓井 (1973) などの研究* が報告されている。
- ✦ ただし、こうした BJ の大半は “obsolete” なものとして70年代以降にはほとんど見られなくなっている。

2.2 代表的ビジネスジャーゴンの使用分布比較



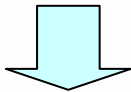
Click to View Excel Tables

* 吉田 (1964) 「Commercialese の特性について」日本商業英語学会年報 (pp.1-9)
碓井 (1973) 「欧米における Commercial Jargon の現況」日本商業英語学会年報 (pp.27-40)

3. BLC* にみられるビジネス英語のコロケーション

3.1 コロケーションの統計的な抽出方法について

- ⊕ いわゆる BJ の大半が学習者のデータ中以外にはほとんど見られないか、あるいは少なくとも英米の規範的な基準からすればすでに “obsolete” なものとしてその使用が戒められているとしても、われわれの直感が示唆するところによれば、ビジネス英語を特徴づける「コロケーション」は、いわゆるビジネスジャーゴンの枠組みを超えて、確かに存在するように思われる。
- ⊕ これはどのようなものか。また、どのように抽出すればよいか。



- ⊕ 本研究では、コロケーションの統計的な抽出方法として、「相互情報量」(Mutual Information) と「仕事量基準」(Cost Criteria) に注目し、この2つの手法を使って BLC からビジネス英語のコロケーションを抽出し、その結果を比較する。

* BLC = Business Letter Corpus. Someya (1999) において作成された英文ビジネスレターコーパスで、英米の参考書・文例集から収集したおよそ100万語の規範的な文例からなる。現在、このBLC を使ったオンライン・コンコーダナーが <http://ysomeya.hp.infoseek.co.jp/> で公開されている。

3.2 相互情報量 (Mutual Information) (1/3)

The mutual information between two words x and y is defined as follows (Brill 1993):

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

where $P(x)$ and $P(y)$ are word occurrence probabilities, which can be estimated from the number of occurrences of the words, $f(x)$ and $f(y)$, and the total number of words in the target corpus, N .

$$P(x) = \frac{f(x)}{N} \quad \text{and} \quad P(y) = \frac{f(y)}{N}$$

Source: Brill, Erick (1993). *A Corpus-Based Approach to Language Learning*. Ph.D Dissertation (pp.27-28). Originally proposed in Church and Hanks (1990).

(cont'd...)

3.2 相互情報量 (Mutual Information) (2/3)

$P(x,y)$, or the joint probability of x and y , can be estimated in a similar way.

$$P(x,y) = \frac{f(x,y)}{N}$$

where $f(x,y)$ is the number of occurrences of x followed by y .

Thus, the mutual information $I(x,y)$ compares the probability of observing the two words, x and y , occurring together in immediate sequence with the probabilities of observing the two words simply by chance.

One important drawback of mutual information is that MI values become unstable for low-frequency words.

(cont'd...)

3.2 相互情報量 (Mutual Information) (3/3)

⊕ 要約

相互情報量 (MI) は、2つの単語の共起関係(相互の結びつき = コロケーション)の強さを表す指標のひとつ。具体的には、ある単語ともうひとつの単語とが共起する確率と、それぞれが個別に生起する確率との比である。実際の計算では、正規化するためにコーパス全体の総単語数をかけ、その対数(底は2)をとる*。ただし、MI スコアは対象となる語の頻度が少ない場合は不安定になる欠点がある。

*対数計算をするのは、スケーリングのため。

Note: This simple method can only extract collocations of length two; however, Jelinek (1990) proposed a generalization of this method to include collocations of arbitrary length. Jelinek, F. (1990), "Self-organized language modeling for speech recognition." In Waibel, A., and Lee, K.F. (Eds.), *Readings in Speech Recognition*, pp. 450-506. Morgan Kaufmann Publishers.

(cont'd...)

3.3 仕事量基準 (Cost Criteria) (1/4)

- ⊕ The cost criteria, proposed by Kita et al. (1994), is based on the assumption that
 - (1) collocations are recurrent word sequences, and
 - (2) the recurrent property is captured by the absolute frequency of a word sequence.
- ⊕ A simply absolute frequency, however, is not appropriate as a measure for identifying or extracting collocations because the frequency of a sub-sequence is always higher than or equal to that of the original word sequence (e.g. would like would like to).
- ⊕ Instead, Kita et al. considers *a processing cost for a word sequence*, and introduces *cost criteria* which can qualitatively estimate the extent to which processing is *reduced* by considering a word sequence as one unit.

(cont'd...)

3.3 仕事量基準 (Cost Criteria) (2/4)

⊕ A *reduced cost* K for a word sequence is defined as:

$$K(a) = (|a| - 1) \times f(a) \quad \dots \quad (1)$$

where

... a word sequence

$|a|$... the length of a (= the number of words in a)

$f(a)$... the number of occurrences of a in a corpus.

⊕ $K(a)$ is interpreted as follows: Assume that, in the target corpus, there exists a word sequence a , which is composed of $|a|$ words and occurs $f(a)$ times. Also assume that the cost of processing one word is 1. Similarly, when processing a as a single unit, its processing cost is 1.

(cont'd...)

3.3 仕事量基準 (Cost Criteria) (3/4)

- ⊕ Here we assume that if a word sequence is processed one word at a time, its processing cost is proportional to its length.
- ⊕ Thus, the processing cost of α is equal to $|\alpha|$, and by considering β as one unit, the processing cost is reduced to $|\alpha| - 1$.
- ⊕ Since β appears $f(\beta)$ times, we can conclude that the total reduced cost become $(|\alpha| - 1) \times f(\beta)$, which is the definition of $K(\alpha, \beta)$.
- ⊕ However, β can always be a sub-set of α (e.g. $\beta = \text{would}$ like, $\alpha = \text{would like to}$), and $f(\beta) \leq f(\alpha)$, the actual reduced cost for α shall be defined as follows:

$$K(\alpha) = (|\alpha| - 1) \times (f(\alpha) - f(\beta)) \quad \dots\dots\dots (2)$$

(cont'd...)

3.3 仕事量基準 (Cost Criteria) (4/4)

⊕ Finally, we can extract collocation from a corpus by the following steps:

1. Calculate $K(\cdot)$ for each word sequence l^{-n} in a corpus.
2. Rank the word sequences l^{-n} in accordance with respective values $K(\cdot)$.
3. Extract higher rank word sequences as collocational candidates.
4. Re-calculate $K(\cdot)$ for each l^{-n} in the collocational candidates.*
5. Re-rank the the collocational candidates by their respective new values $K(\cdot)$ to get a final list of collocations.

(* To be more precise, by checking the sequence/sub-sequence relation between every two word sequences in the collocational candidates, modify the $K(\cdot)$ values according to above Equation 2.)

Source: Kita, K. et. al (1994). "A Comparative Study of Automatic Extraction of Collocations."
Journal of Natural Language Processing. Vol. 1, No. 1.

3.3 仕事量基準 (Cost Criteria) (4/4)

Φ 要約

仕事量基準 (CC) は、MI と同じく、2つの単語の共起関係(相互の結びつき = コロケーション)の強さを表す指標のひとつで、単語の処理コスト (processing cost) という考え方を導入して特定コロケーションの共起関係の強さを測る方法である。CC では、あるコロケーションの処理コスト = $K(\quad)$ を次のように解釈する。

1. あるコーパスの中に $\langle w_1, w_2 \rangle$ という連語があり、これが $|S|$ 語からなる連語で、その頻度を $f(\langle w_1, w_2 \rangle)$ とする。
2. このとき、 $\langle w_1, w_2 \rangle$ を構成する各語の処理コストをそれぞれ1とし、さらに $\langle w_1, w_2 \rangle$ をひとつのユニットとして処理するときのコストを1とする。
3. $\langle w_1, w_2 \rangle$ の各語を1つずつ処理したときの処理コストは $\langle w_1, w_2 \rangle$ の語数 ($|S|$) に比例し、かつ全体の処理コスト = $K(\langle w_1, w_2 \rangle)$ は $|S|$ に等しいとすると、 $\langle w_1, w_2 \rangle$ をひとつのユニットと考えた場合の処理コストは $|S| - 1$ となる。
4. $\langle w_1, w_2 \rangle$ は $f(\langle w_1, w_2 \rangle)$ 回生起することから、全体の処理コストは $(|S| - 1) \times f(\langle w_1, w_2 \rangle)$ によって求めることができる。
5. ただし、 $\langle w_1, w_2 \rangle$ は常に $\langle w_1, w_2 \rangle$ のサブセットであることから (例えば $\langle w_1, w_2 \rangle$: would like は $\langle w_1, w_2 \rangle$: would like to のサブセットである)、 $\langle w_1, w_2 \rangle$ の実際の処理コスト = $K(\langle w_1, w_2 \rangle)$ は $f(\langle w_1, w_2 \rangle)$ から $f(\langle w_1, w_2 \rangle)$ を引いたもの、つまり $(|S| - 1) \times (f(\langle w_1, w_2 \rangle) - f(\langle w_1, w_2 \rangle))$ となる。

3.4 分析結果と考察 (1/2)

⊕ Business Letter Corpus Collocation

1) Score method: **Mutual Information (MI)**

Initial Filter: MI (min. 0.5)

Full extract filter: MI

Span: 2-4 (2-4 grams を出力)

2) Score method: **Cost Criteria (CC)**

Initial Filter: T-Score (min. 0.5)

Full extract filter: CC

Span: 3-4 (3-4 grams を出力)



[Click to View Excel Table](#)

3.4 分析結果と考察 (2/2)

- ⊕ MI は “TOYOSAN MOTOR COMPANY LTD” や “The Wall Street Journal”、あるいは “(To) Whom It May Concern” や “be kept strictly confidential” といった、特定の文脈やトピックに依存した固有名詞や複合名詞句、および特定分野の専門用語に高いスコアを与える傾向が強い。
- ⊕ 一方、CC は “Thank you for” や “look forward to”、“I would like to”、あるいは “as soon as possible” や “in the future” といった、より日常的かつ定型的なフレーズを選好する傾向がある。
- ⊕ 以上のことから、Kita et al. (1994) が主張するとおり、語学学習のための基本的なコロケーションの抽出という観点から見た場合は、CC のほうが優れているという言うことができる。ただし、ESP 的見地からすれば、さらにその上で MI による task-dependent で text specific な複合名詞句や専門用語を加えてゆくことで、よりの確で包括的なコロケーションの一覧を作成することが可能になると考えられる。

4. まとめと今後の課題

- ⊕ 本研究では、特定分野(この場合はビジネス文書)の大規模コーパスから、ビジネス英語の「コロケーション」を自動的に抽出するための予備的調査として、「相互情報量」(MI) および「仕事量基準」(CC) という2つの統計的手法に注目し、それぞれの有効性および特性について検討した。
- ⊕ その結果、MI と CC はそれぞれ性格の異なったコロケーションを抽出することが確認された。語学学習のための基本的なコロケーション(主として文法的コロケーション)の抽出という観点から見た場合は CC のほうが優れており、一方、特定の文脈やトピックに依存した専門用語や慣用表現については MI のほうがより検出力が高い。
- ⊕ 今回の予備調査の結果は、上記2つの手法を適切に組み合わせることで、従来のビジネスジャーゴンの範疇を超えた、より有意義な、かつ現在のビジネス英語の使用実態に則した「コロケーション」のリストを作成することが可能であることを強く示唆するものであった。ただし、この2つの手法によって抽出したコロケーションリストの中には、有用性の低いジャンクデータや重複データも大量に混じっていることから、今後、これらのジャンクデータの出力を制御するためのアルゴリズムの開発が期待される。

参考文献

1. Benson, M., E. Benson and R. Ilson (1997). *The BBI Dictionary of English Word Combinations*. Amsterdam and Philadelphia: John Benjamins.
2. Brill, Erick (1993). *A Corpus-Based Approach to Language Learning*. Ph.D Dissertation (pp.27-28). Originally proposed in Church and Hanks (1990).
3. Hill, J. and Lewis, M. (eds.) (1997). *LTP Dictionary of Selected Collocations*. Hove: Language Teaching Publications.
4. Jelinek, F. (1990). Self-organized language modeling for speech recognition. In waibel, A., and Lee, K.F. (Eds.), *Readings in Speech Recognition*, pp. 450-506. Morgan Kaufmann Publishers.
5. Kita, K. et. al (1994). A Comparative Study of Automatic Extraction of Collocations. *Journal of Natural Language Processing*. Vol. 1, No.1.
6. Kjellmer, G. (1991). A mint of phrases. In K. Aijmer. and Altenberg, B. (eds.) *English Corpus Linguistics*. London: Longman.
7. Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Hove: Language Teaching Publications.
8. Lewis, M. (1997). *Implementing The Lexical Approach*. Hove: Language Teaching Publications.
9. Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. Boston: Heinle & Heinle Publishers.
10. Pawley, A. and Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency." In Jack Richards & Richard Schmidt (Eds.), *Language and Communication* (pp. 191-226). London: Longman.
11. Shei, C. and Pain, H. (1999). An ESL Writer's Collocational Aid, *Computer Assisted Language Learning*, Vol. 13, No. 2 (pp. 167-182).
12. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
13. Someya, Y. (1999). *A Corpus-based Study of Lexical and Grammatical Features of Written Business English*. Unpublshes MA Thesis submitted to the University of Tokyo.

参考文献

14. 碓井陽一(1973)「欧米における Commercial Jargon の現況」日本商業英語学会年報 (pp.27-40)
15. 染谷泰正 (2001)「日本人学習者の書いた英文ビジネス文書にみられるメタ・ディスコースマーカの分布と使用傾向について」日本商業英語学会関東支部例会 2001年1月13日
16. 吉田隆章(1964)「Commercialese の特性について」日本商業英語学会年報 (pp.1-9)