基本動詞のコロケーション難易度測定
—CEFR レベルに基づくテキストコーパスからのアプローチ—

内田　諭
九州大学大学院言語文化研究院

1. はじめに

本研究は、学習レベル別にカテゴリー化された英語テキストコーパスを基に、基本動詞のコロケーションの難易度測定について、探索的にその指標を探ることを目的としている。英単語の難易度は、JACET8000（相澤 et al. 2005）、SVL12000（アルク[1]）、CEFR-J[2]（投野 2013）など、多くのリストが存在するが、英単語のコロケーションの難易度を示すものは、管見の限り信頼性の高いリストは存在しないのが現状である。

複数単語のまとまり（multiword unit や chunk などと呼ばれることが多い）について、学習上の重要性を指摘する研究は多く存在する（Sinclair 1991, Hill 2000, Lewis 2000, Laufer & Waldman 2011, 堀 2011 など）。一方、コロケーションを含むそのような「まとまり」の学習上の難しさも指摘されている。Altenberg & Granger (2001)は、上級学習者であっても、make のような基本語のコロケーションの習得が難しいことを指摘し、望月（2007）は、特に日本人の学習者にとって、make のコロケーションの習得が容易ではないことを示している[3]。しかしながら、これらの研究は個別的な分析に留まっており、どのコロケーションがどの程度難しいかということについて体系的な分析は提示されていない。

コロケーションの難易度の測定の難しさの 1 つは、平易な単語同士の組み合わせが必ずしも難易度が低いとは限らないことである。たとえば、make は CEFR-J では A1 レベル、JACET8000 では上位 1000 語以内[4]、名詞の contact はそれぞれ A2 レベル、上位 1000 語以内であり、単語単体としての難易度は低いと言えるが、make contact （接触する、連絡する）という連語になると、直感的には最も簡単な学習レベルのレンジに位置するとは判断し難い。

本研究では、「コロケーションの難易度」を測定する足がかりを得るために、CEFR レベルに基づくテキストコーパスをインプットとし、対応分析を用いて基本動詞のコロケーションを難易度別にマッピングするということを試みる。CEFR は世界的に用いられている学習者レベルの判断指標であり、コロケーションの難易度を示す上でも強固な土台となることが期待できる。以下の議論では make をケーススタディとして取り上げ、レベルごとのコロケーションに意味的・構文的なパターンがないかを探

1 http://www.alc.co.jp/vocgram/article/svl/

2 CEFR（Common European Framework of Reference for Languages）を日本の英語教育に当てはまるように改良したもの。CEFR の A1～C2 の 6 段階の指標を、A・B レベルを中心に細分化している。詳しくは、投野（2013）を参照のこと。

3 望月（2007）によれば、日本人の学習者は創造を表す creative make を過剰使用し、軽動詞の make や「お金を稼ぐ」（make money）などの用法は過小使用するという。

4 JACET8000 は頻度順に単語をリストしており、make は 64 位、contact は 800 位である。

索的に検証する。また、その結果を受け、言語処理の分野の知見をコロケーションの難易度測定に応用する方向性を示す。

2. テキストコーパス

本研究で用いるデータは、CEFR を参照して編纂されたと考えられるテキストがベースとなっている。テキストの中にはレベルが複数に跨るものもあるが（例：B1~B2 レベルが対象のテキストなど）、それぞれのレベルの特徴を特定することを目的としているため、それらはコーパス化の対象とはしなかった。また、既存のテキストシリーズ等にカタログなどでの表示のため CEFR レベルを後から適応したと考えられるものについてもコーパス化の対象から除外した。このテキストコーパスの概要は表 1 の通りである。C2 レベルについては対象が 1 冊で語数も極端に少ないため、考察の対象外とした。

	採用冊数	総語数
A1	13	104,602
A2	21	262,335
B1	27	466,407
B2	24	563,016
C1	9	264,898
C2	1	28,607
Total	95	1,689,865

表 1　テキストコーパスの概要[5]

テキストコーパスは品詞タグ付けを行わず生データのまま make およびその活用形の後にくる単語（スパン 4）を集計し、頻度を相対化した上で、CEFR レベル別にクロス集計を行った。ただし、冠詞（a, an, the）や代名詞（I, he, she, this, that など）などの機能語については集計から省いた。

3. 対応分析の結果と考察

2 節で行ったクロス集計表に対して、make の共起語と CEFR レベルをマッピングするため対応分析を行った。対応分析には、統計ソフト R を用いた。MASS ライブラリの corresp 関数を使用し、描画には biplot 関数を用いた。結果は図 1 の通りである。

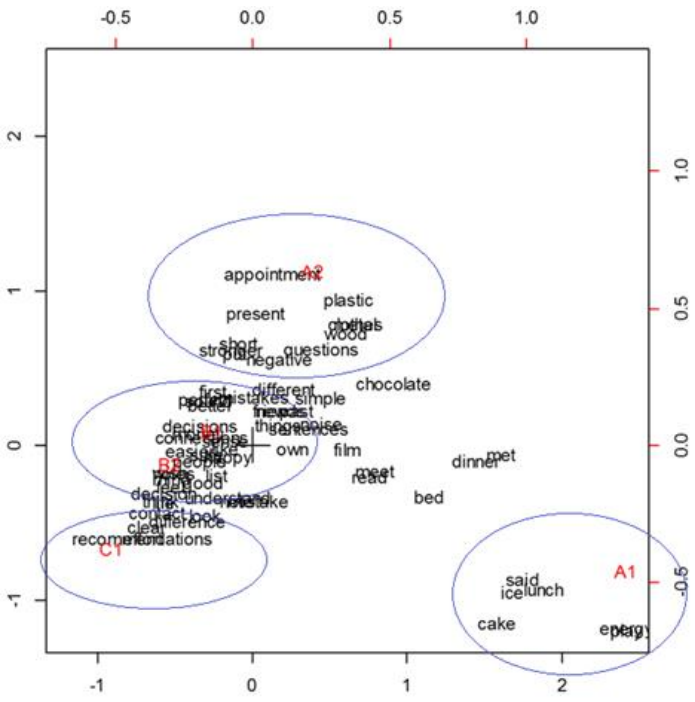

図 1　対応分析の結果

図中の円は、CEFR のそれぞれのレベルに該当すると考えられる語群を目視で囲ったものである。ラベルの重なりがある部分については同時に出力される固有値を参照し、対応を考察した。

[5] 本コーパスは作成段階であり、表中の数字は暫定値を示す。

3.1 A1 レベルのコロケーション

A1 のラベルに近い共起語に lunch (A1)[6], cake (A1), dinner (A1), ice (A1)などがある。これらは具体的なものであり、望月（2007）の分類に従うとこの make の用法は creative make であると言えるだろう。日本語では「作る」と訳すことができる最も基本的な用法で、学習者にとっても意味が理解しやすいものであると言える。

3.2 A2 レベルのコロケーション

次にA2 レベルで共起する単語について考察すると、present (A1↓), clothes (A1↓), friends (A1↓)などの具体的なものを表す単語のほか、appointment (A2), questions (A1↓), noise (A1↓)などの抽象的なものも連結する傾向があることが読み取れる。make の意味としては creative make として解釈できるものが多いが、コロケーションの内実が、具体物と抽象物が混ざっているという点でA1 レベルとは異なっている。

3.2 B レベルのコロケーション

B レベルのコロケーションについては、B1 および B2 のラベルが近接し、これらを明確に区別することは難しい。しかしながら、このレベルでの特徴として、次の 3 点が読み取れる。第 1 に、decisions (B1), connections (B1)など、より上位レベルの抽象名詞が観察されたことが挙げられる。ノードが基本語であっても共起語の難易度が上がれば、コロケーションとしての難易度も当然上がることになる。第 2 に、easier (A1↓), happy (A1↓), simple (A2↓), understand (A2↓)などの形容詞や動詞が多く共起していることから、このレベルから構文的な複雑性（特に SVOC の構文）が増していることが読み取れる。これらの単語は単体としては B レベル未満の単純なものであることに注意したい。最後に、sense (A2↓)、sure (A1↓)などが B レベルのラベルと近接することから、慣用表現が出現していることがわかる。これらの共起語についても B レベル未満のものが目立つことが指摘できる。

3.4 C1 レベルのコロケーション

C1 のラベルは B1・B2 の外周に当たる部分に配置され、はっきりと特徴を読み取ることが難しいが、recommendations (B2↓)のような比較的レベルが高いと考えられる語や clear (A2↓) (SVOC の構文)、difference (A1↓) (慣用的表現) などが見られた。特徴としては B レベルと類似しており、今回のデータでは C1 レベルにおける make の特性は指摘することが難しい。なお、冒頭の make (A1)と contact (A2)の組み合わせは B2 と C1 のラベルの間に見られ、テキストコーパスの分析からは B2 レベル以上とするのが妥当であることが読み取れる。

4. 難易度測定の指標

前節での議論の結果、CEFR レベルごとにコロケーションの特徴が異なり、一定のパターンが見られることがわかった。レベル別に見られるコロケーションのパターンとして、少なくとも(1)具体的なものを指す名詞のコロケーションは A レベルの特徴、(2) (a)比較的難易度の高い抽象的な名詞との組み合わせ、(b)構文的な複雑性、(c)慣用表現が B レベルの特徴であるということが言えるだろう。

コロケーションの難易度測定を機械的に行

[6] カッコ内は CEFR-J のワードリスト（投野 2013）でのレベルを示す。また、上下の矢印は当該レベルと単語のレベルにギャップがあることを示している。

う指標として、共起語自体の難易度に加えて、共起する名詞の具体性を示すものがあれば、レベルの基準の 1 つとなりうるだろう。詳細な議論は別の機会に譲るが、make 同様に基本動詞である have や get の分析でも具体性の高い名詞が A レベルの特徴として現れる傾向を示す。また、テキストコーパス中の CEFR レベルと、共起語単独の CEFR レベルの乖離が目立つのは、構文的な複雑性を伴う場合と慣用表現の場合である。前者については、品詞タグやパーザーなどを利用するなどして構文的な複雑性を数値化できれば、特に多様な用法を持つ基本動詞のコロケーションの難易度を測定することに繋がると考えられる。後者については辞書的なアプローチでイディオムを特定することで難易度の測定につなげることができるだろう。

5. まとめ

本研究では、基本動詞の make を例にとり、コロケーションの難易度測定について対応分析から探索的に基準となる指標を探った。予備的な調査ではあるが、共起語の具体性および慣用性、構文的複雑性などの候補を、コーパス分析に基づいて提示した。

コロケーションの難易度を明示的に測定できれば、教育的な示唆は大きい。そのためには大規模な学習レベル別のインプットおよびアウトプットコーパスの構築と、機械学習などの言語処理の手法を応用することが求められている。

［付記］

本研究は、JSPS 科学研究費補助金基盤研究 A「学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究」（研究課題番号：24242017）の助成を受けたものである。

［参考文献］

相澤一美 ・石川慎一郎・村田年 ［編著］ (2005). 『「大学英語教育学会基本語リスト」に基づく JACET8000 英単語』桐原書店.

Altenberg, B., & S. Granger (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics*. 22(2), 173-195.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (ed). *Teaching collocation: Further development in the lexical approach*. Heinle, Cengage Learning, 47-69.

堀正広 (2011). 『例題で学ぶ英語コロケーション』研究社.

Laufer, B. & T. Waldman (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*. 61(2), 647-672.

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (ed). *Teaching collocation: Further development in the lexical approach*. Heinle, Cengage Learning, 10-27.

望月通子. (2007). 「日本人大学生の EFL 学習者コーパスに見られる MAKE の使用」『外国語教育研究』14, 31-45.

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

投野由紀夫［編］(2013). 『CAN - DO リスト作成・活用　英語到達度指標 CEFR - J ガイドブック』大修館書店.