

CEFRレベルに基づいた教材コーパス：レベル別基準 特性の抽出に向けて

内田, 諭
九州大学大学院言語文化研究院

<https://hdl.handle.net/2324/1932355>

出版情報：英語コーパス研究. 22, pp.82-88, 2015-03. 英語コーパス学会
バージョン：
権利関係：

CEFR レベルに基づいた教材コース レベル別基準特性の抽出に向けて

内田 諭

1. はじめに

近年、言語教育の分野では学習評価指標として「言語を使って何ができるか」という機能的な側面が重視されるようになってきている。その流れの中で CEFR (Common European Framework of Reference for Languages; 詳細は2 節参照) が利用されることが多くなり、CEFR に準拠した教材の数も増えてきている。その一方で、CEFR の指標は、「できること」を記述した CAN-DO リストが中心となっているため、語彙や文法などの形式的な言語特徴(以降、基準特性と呼ぶ)と CEFR レベルとの関連については学術的な蓄積が十分にある状態ではない。CEFR と語彙や文法の関連を明らかにすることで、機能的で実践的な CAN-DO リストに基づいたシラバスと、伝統的な英語教授法をつなぐことが可能となり、教授法を段階的に転換していくための足がかりとなる可能性がある。その一つの方法として、CEFR レベルを変数に持つコースを作成し、そこから統計的に情報を抽出するということが有効であると考えられる。

本稿は、「CEFR レベル別の基準特性を明らかにする」ことを目的として現在構築を進めているコースを具体例として取り上げ、教材コースの作成方法と課題、およびその利用例を示すことを目的とする。特に、教材コースを構築する上で必要となる注意点や課題に焦点を当て、実際の作業で遭遇したことを中心に述べる。以下、2 節では CEFR の概要を示し、3 節で CEFR に基づいた教材コースの作成過程と課題について述べる。4 節は、教材コースの利用方法として、get, have, make の3 つの基本語を例に取り、そのコロケーションから CEFR レベル別の基準特性を考察する。5 節はまとめである。

2. CEFR の概要

CEFR はヨーロッパの複言語主義(plurilingualism)を背景に、相互に意思疎通

を図るための学習到達指標として、世界的に広く用いられている基準である。CEFR は言語学習の達成度を測定する上で行動指向アプローチ (cf. 投野, 2013) をとり、「言語を用いて何ができるか」ということが基準となる。CEFR の指標は、Global scale というレベル別の一般的な記述に加えて、Listening, Reading, Spoken Interaction, Spoken Production, Writing の5 つ技能に基づいた記述も存在する。また、CEFR のレベルは A1, A2, B1, B2, C1, C2 の6 段階であり、おおよそ A レベルが初級、B レベルが中級、C レベルが上級に相当する。記述はすべて CAN-DO statements の形で提供され、状況や条件などについての説明も含む。以下に B1 レベルの記述の例を示す。

[B1 レベルの global scale]

Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.

Can deal with most situations likely to arise while traveling in an area where the language is spoken.

[B1 レベルのライティングの指標]

I can write simple connected text on topics which are familiar or of personal interest.

I can write personal letters describing experiences and impressions.

CEFR の CAN-DO statements は、多様な言語が使用されているヨーロッパでのコミュニケーションの確立と改善に資するだけでなく、教育の場面では、教授者と学習者間で目的を共有する事が可能になり、学習の到達点が明確になるというメリットがある (cf. 投野, 2013)。日本でも関心が高く、文部科学省は2013年3月に「各中・高等学校の外国語教育における『CAN-DO リスト』の形での学習到達目標設定のための手引き」を発行し、教育現場への浸透を図っている。また、日本の教育現場の共通指標として CEFR をベースに構築された CEFR-J (投野, 2013) も公開されている。この指標は、主に初級の A レベルを細分化した点が特徴で、日本人の学習者の大半が A レベルであるという現状 (cf. Negishi, Takeda and Tono, 2012) に対応することを謳っている。このような特徴から、CEFR-J は、日本の英語教育の現場への応用がより実践的で平易にできるものとして注目されている。

3. 教材コーパスの作成について

CEFR が言語教育の文脈で注目を増すにつれて、海外の出版社を中心に、CEFR に準拠した英語教材が編纂されるようになってきている。本研究で用いるコーパスは、そのような教材を集め、電子的に集積したものである。学習者にとって重要なインプットである「教材」は、語彙や文法項目などの面で意図的な統制が行われており、汎用コーパスを作成する場合は異なる点に注意を要する。本節では、「教材コーパス」を作成する過程とそれぞれの段階で直面する課題について、実際の作成作業を通じて明らかになったことに基づいて述べる。

3.1 教材コーパスの目的

CEFR の指標は言語の機能面に重点をおいた実用的な指標であるが、CAN-DO statements の性質上、語彙や文法などの言語の形式的な特徴についてはほとんど触れられていない。2 節でみた例でも文脈や条件についての記述は見られるが、形式的な側面への言及がないことがわかる。これらをリンクさせる試みは English Profile のプロジェクト (<http://www.englishprofile.org/>) や CEFR-J のプロジェクト (語彙リストを提供) などで部分的に行われている。しかしながら、管見の限り、コーパスデータを用いた包括的なアプローチはこれまでには行われていない。

本研究で用いるコーパスは、CEFR を参照して編纂されたと考えられる教材をレベル別に収集し、レベルを分けている基準特性はどのようなものがあるかを探索的に明らかにすることを目的として構築が進められているものである。大量のデータを基に、統計的な手法も援用しつつ語彙や文法事項と CEFR レベルの対応付けが可能となれば(例えば「現在完了形」は B1 レベルで顕著に見られる、など)、data-driven なアプローチから抽出した基準として価値があるだけでなく、English Profile などで提示されている記述 (RLD: Reference Level Descriptions) を検証することができるという点でも有益である。

3.2 コーパス化の対象

本研究で収集対象とするテキストは、CEFR 準拠を謳うものである。出版社がどこまで厳密に CEFR に従っているかを知ることは難しいが、可能な限り良質なテキストを集めるため、CEFR の包括的な解説書である *Common European Frame-*

「シンポジウム」CEFR レベルに基づいた教材コーパス レベル別基準特性の抽出に向けて

work of Reference for Languages: Learning, Teaching, Assessment (Cambridge University Press) が発刊された2001年以降のものを中心とし、既存のテキストを「後付け」で CEFR レベルに当てはめたと考えられるものは除外した。また、CEFR のそれぞれのレベルを厳密に区別するため、レベルがまたがって提示されているもの(例: B1-B2が対象)もコーパス化の対象から外した。流通しているテキストはいずれも商業的な出版物であり、パイアスや制約がかかっている可能性があるが、少なくとも教材作成者が共有する CEFR レベルのイメージを具現化するという点で、教材を分析することは意義のあることだと考えられる(妥当性を確保するためには、抽出された基準特性を質的に検証する作業が必要であるがこの点は本稿の対象外とする)。

3.3 作成過程と各段階の課題

本節では、CEFR レベルを弁別する基準特性を抽出することを目的とした教材コーパスを具体例として取り上げ、その作成過程について、テキストの電子化、カテゴリー化、タグ付けの順に、それぞれのステップの概要と教材コーパスを作成する際に考慮すべき注意点や課題について述べる。

3.3.1 テキストの電子化

コンピューターを使った分析を行うためには、テキストを電子化する必要がある。教材の場合、出版社によっては電子媒体を提供していることもあるが、多くの場合は手作業で電子化を行うことになる。特にチームで行う場合、可能であれば同一のスキャナを、異なる場合もスキャナの読み取り設定を統一することに注意を払いたい。本研究では、すべてカラー読み取りの300dpiに統一した。この設定が読み取りの精度やファイルの大きさなどの点で最適だったからである。また、読み取りに用いる OCR ソフトも統一したほうがよい。本研究の場合、Abbyy FineReader 11 Professional Edition を使用した。

この段階における最大の問題点は読み取りの精度である。高性能なソフトであっても認識率は100%にはならず、ある程度目視で確認する作業が必要となる。特にレイアウトが複雑なページ、カラーの写真の上に文字がプリントされているなど、背景とのコントラストが低いページ、手書き風のフォントが使用されたページなどは読み取りの精度が低くなる可能性が高い。このようなページを中心に紙面と読み取り結果を地道に比較するというのが最も堅実な方法である。また、Microsoft Word などの文書作成ソフトにテキストをペーストし、スペル

チェックをすることも有効である。校閲の機能を使えば、上から順に確認していくことができる。ただし、固有名詞などの特殊なスペルの場合もエラーの候補として指摘されることがあるので、一律には修正候補を信頼することはできない。

3.3.2 カテゴリー化

次にコーパスの目的や使用方法を念頭に置いて電子データをカテゴリー化していく。本研究ではCEFRレベル別の基準特性の抽出が目的であるため、テキストのレベル別にサブコーパスを作成した。特に教材の場合、一般的な英語能力試験などのレベル指標が提示されていることが多く、これらの指標は分析の重要な変数になり得るため、タグなどをつけて情報を残しておくとういだろう。

教材が目的とする技能を基準にしたカテゴリー化も有効である。例えば、リーディングの教科書とライティングの教科書では書かれている英文の質や内容が異なることが多いので、技能タグをつけておくという方策が考えられる。また、4技能統合型の教材の場合は、それぞれのセクションごとに技能タグをつけることもできるだろう。本研究では、Reading, ReadingC(会話を読むタイプ), Writing, WritingV(書くための語彙や文例), Speaking(発話のモデル), SpeakingV(話すための語彙や文例), Listening(リスニング用モノログ), ListeningC(リスニング用ダイアログ), Conversation(RC・LCの判断が難しい場合などの曖昧タグ), Grammar(文法の説明文と用例の両方を含む), Vocabulary(語彙リスト) にカテゴリー化した。4節のケーススタディではこれらのカテゴリーは用いないが、他の研究において有益なものとなる可能性があるため、労力と時間の許す範囲でデータのカテゴリー化しておく。

この段階での難しさは、カテゴリー決定の揺れである。今回、CEFRレベルに関しては教材に提示されているものをそのまま用いることとしたが(ただしレベルがまたがる教材は除外)、技能系のタグに関してはセクションタイトルだけで判定することが難しいことがあり作業者によって判断が異なる場合も考えられる。作業指標を共有する、同一テキストを用いてアノテーターの訓練を行う、判定が難しい場合は保留する印を付ける等の工夫により、可能な限り揺れを抑えることが課題である。

3.3.3 品詞タグ付け

カテゴリー化を行った後、コーパスとして詳細な検索が可能となるように品詞タグなどの情報をテキストに付与する。機械的に行うことが一般的であるが、

教材コーパスの場合、処理に不適切な形式が含まれていることがあるので注意が必要である。例えば、単語のリストであれば、前後の文脈がないため、increaseのように動詞でも名詞でも使用できるものは、正確な品詞判定は困難である。そのほか、I'm sure that ...などのように不完全なフレーズ、練習問題のように空所があるもの、エラーを提示するための誤例などがあると、品詞タグの精度が大きく低下する恐れがある。

3.4 教材コーパスの概要

以上のようなプロセスを経て集められた教材コーパスは、表1の通りである。ここで注意したい点は、各カテゴリーのバランスである。均衡コーパスのように厳密に調整する必要はない場合が多いと考えられるが、極端に少ない部分は分析から除外するか、可能であれば資料を補強するなどの方策を取る必要がある。また、プレーンテキストの状態では検索や集計が容易ではないため、目的にあったコンコーダンサーで処理するなどの方法も検討する価値があるだろう。

表1 教材コーパスの各レベルの冊数と語数

レ ベ ル	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

4 . 教材コーパスの利用例：基本語のコロケーション

本節では、CEFR レベル別に構築した教材コーパスを使用して基本語のコロケーションからレベルごとの特徴を探るというケーススタディを提示する。表1から明らかなように、C2レベルはサイズが極端に小さいため今回の分析では除外した。対象となる基本語は、get, have, make の3つである。これらの動詞は、基本語であると同時に様々な構文をとるもので、コロケーションの意味的な特徴に加えて、構文的な特徴が抽出されることが期待できる（例えば make は「～を作る」という基本的な意味に加えて、SVOCの構文で使役動詞となる）。

コロケーションの学習上の重要性を指摘する研究(e.g., Sinclair, 1991, Hill, 2000, Lewis, 2000, Laufer and Waldman, 2011, 堀, 2011) は多く存在する一方で、学習上の難しさも指摘されている。Altenberg and 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ている。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学習者データから示し、特に「作る」という意味のmake(creative make) は過剰使用の傾向があるのに対し、軽動詞のmakeや「お金を稼ぐ」(money make) という意味のmakeなどは過少使用の傾向があると指摘する。また、単語のレベル別のリストはJACET8000(相澤・村田・石川, 2005), SVL12000(アルク), CEFR-J(投野, 2013) などが存在するが、管見の限り、コロケーションのレベル別のリストは存在しない(cf. 内田, 2015)。このようなことを考えると、基本語のコロケーションおよびその特徴をCEFRレベル別に(= 難易度別に) 提示することができれば、教授面でも学習面でも非常に有益な情報になると考えられる。

本節では、CEFRレベルに基づいた教材のコーパスについて3.3.2(カテゴリー化) の処理までを経たものを用いる。これは品詞タグのエラーを回避する目的と、品詞を限定しないほうが構文的な特徴を効果的に観察できると考えたからである。コンコーダンサーはAntConc(ver. 3.2.4) を用い、それぞれの原形および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除外した。

クロス集計表から特徴を抽出する手法として、対応分析が挙げられる。対応分析は言語分析でも広く用いられ、その有効性が証明されている(e.g., 田畑, 2009, 後藤, 2007)。本研究でも、それぞれの基本語の共起語とCEFRレベルをマッピングするため対応分析を実施した。対応分析には、統計ソフトR(ver. 3.0.2) を用い、MASSライブラリのcorresp関数を使用し、biplot関数によって2軸上に描画した。以下、get, have, makeの順に結果を示し、考察を行う。図中、ラベルの重複により判読が難しいところがあるが、関連する部分については同時に出力される固有値を参照し、本文中で言及した。なお、以下のmakeの論述については内田(2015) で提示した議論に基づいていることを付記しておく。

4.1 getのコロケーション

図1 は get の共起語と CEFR レベルのクロス集計表に対して対応分析を行った

「シンポジウム」CEFR レベルに基づいた教材コーパス レベル別基準特性の抽出に向けて

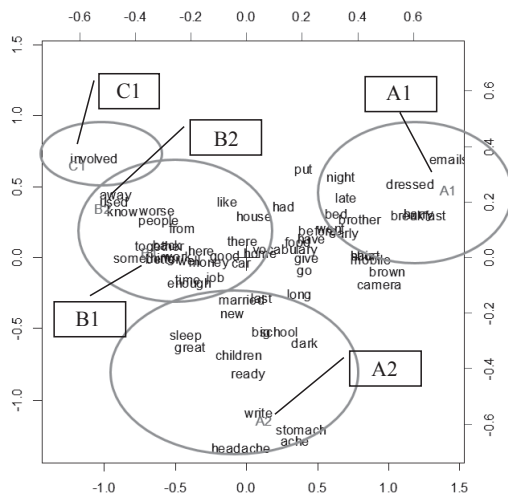


図1 対応分析からみた get のコロケーションと CEFR レベル

結果である。対応分析では、ラベルの近接性が関係の密接さを表す。特に2つのカテゴリー間(CEFRレベルと共起語)を同一次元上に配置することができ、それらの関係性を見出す上で対応分析は有効な手段である。ここではCEFRレベルごとに対応すると考えられる共起語を見ていく。なお、図中の楕円は、それぞれのレベルの大まかな範囲を示すために筆者が付加したものである。

まず、A1レベルと近接する共起語を見てみると、emails, breakfastなど、具体的なものが目立つ。また、get dressed(着替える), get ~ late(遅く ~ に到着する)など、日常的な定形表現も特徴的に見られることがわかる。

次にA2レベルについて考察する。A2レベルはY軸のマイナスの座標に配置されているが(第2主成分がマイナス)、同様の値を持つ共起語に注目すると、ready, dark, lost, marriedなどが見られ、A2レベルではgetは「~になる」(SVC構文)という意味で用いられていることが推測できる。また、headache, stomachacheなどが近接していることから、痛みが出てくるという意味でも使われていることが観察できる。

B1, B2, C1については同一の象限に配置されていることから、似た特徴があることがわかる。このレベルでは慣用表現と思われる用法が多い。例えば、B1ではget together, get here, B2ではget away, get used(to), get worse, C1ではget involved(in)などが特徴的である。

以上のことから，get のコロケーションは A1レベルでは具体物を指す名詞が多く，A2レベルでは SVC 構文で使用される形容詞が特徴的に見られ，Bレベル以上では慣用表現が目立つ，ということが明らかとなった。

4.2 have のコロケーション

次に，have のコロケーションについて考察する。図2 は対応分析の結果である。

A1ラベルについて，breakfast, lunch, dinner, coffee などとのコロケーションから「食べる」の意味の have が使われていることがわかる。また，children, car などの共起語から具体的なものの所有や，shower, party などから動作を表す用法で have が用いられるということが読み取れる。

A2レベルでは，fun, holiday, talk, question など，抽象物を表す(目では観察できない)単語と共起する傾向があることがわかる。

B1のラベルの付近には，asked, lived, finished, decided, seen, done, tired などの共起語が配置されている。今回の調査では品詞タグを指定していないので，これらの語は助動詞の have とのコロケーションであることが読み取れる。つまり，現在完了形や過去完了形などの用法が，このレベルと密接に関わっていると言え

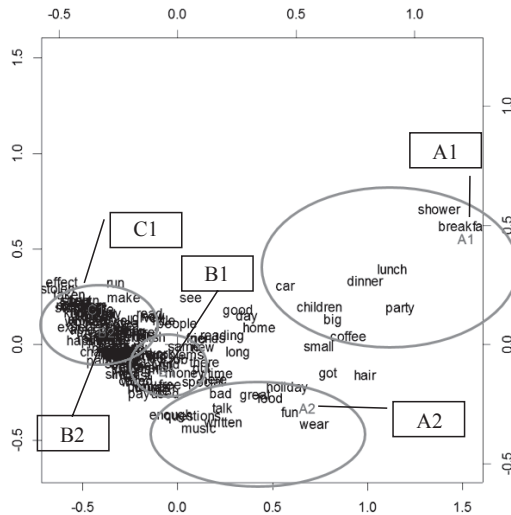


図2 対応分析からみた have のコロケーションと CEFR レベル

習者が過剰使用する傾向を指摘している。このことは日本人の学習者の *make* の用法は、A1レベルのものを繰り返し使っており、より上位のレベルのコロケーションは十分に使用できない可能性を示唆する。

次に A2レベルで共起する単語について考察すると、*present, clothes, friends* などの具体的なものを表す単語のほか、*appointment, questions, noise* などの抽象的な名詞が多いことが読み取れる。*make* の意味としては何かを生み出すという点で *creative make* として解釈できるものが多いが、A1レベルとは異なり、共起語に具体物と抽象物が混ざっていることが特徴である。

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が、*decisions, connections* など、より上位レベルの抽象名詞が観察されること、*easier, happy, simple, understand* などの形容詞や動詞が多く共起していることから SVOC の文型で *make* が用いられる傾向にあることが読み取れる。さらに、*sense, sure* などの単語が 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。

C1のラベルは B1・B2の少し外側に配置されており、はっきりと特徴を読み取ることが難しいが、*recommendations* のような比較的レベルが高いと考えられる語や *clear* (SVOC の構文)、*difference* (慣用的表現) などが見られ、多様な用法が用いられていることがわかる。

以上より、*make* のコロケーションは、レベルが上がるにしたがって、具体物から抽象物への段階的な変化が見られ、また形容詞等の出現から構文が複雑化し、慣用表現が用いられるようになる、とまとめることができる。

4.4 対応分析の結果のまとめ

ここまでの議論の結果、基本語のコロケーションをみることで、レベル別の特徴がはっきりと読み取れることがわかる。まず、コロケーションの特徴として、A1レベルでは特に具体的なものを表す名詞が多いという点が共通している。このことから、「基本語(基本動詞)と具体名詞のコロケーション」は、A1レベルの基準特性の1つとして挙げることができるだろう。統計的な結果で言うと、すべての基本語で A1, A2ラベルは第1主成分(X軸)がプラスになっており、それ以外はマイナスの値をとる。このことから、Aレベルとそれ以上では、基本語のコロケーションの傾向が大きく異なることが示唆される。また、レベルの上昇とともに構文的な複雑さが増すことも観察された。現段階では探索的な試みの結果に過ぎないが、*get* は A2レベルで SVC の構文、*have* は B1レベルで完了形の

「シンポジウム」CEFR レベルに基づいた教材コーパス レベル別基準特性の抽出に向けて

構文, make は B レベルで SVOC の構文と密接に関わっていることが示唆される。このような特徴もそれぞれのレベルを表す基準特性の候補として検証する価値のあるものである。

5 . 結 語

本稿では教材コーパスの作成過程とそれぞれの段階での課題を概観した。コーパスの作成は地道な作業であり労力がかかるが、本稿で示したような方法や注意点等を広く共有することで、効率的な構築につながると考えられる。教材の場合、著作権の関係で公開の範囲が限定されてしまうことが多いが、作業上のプロセスを共有し蓄積していくことは可能だろう。

教材コーパスは、教育上非常に有効なデータを抽出できる可能性を秘めている。その1つの試みとして、本稿では基本語のコロケーションから CEFR の各レベルの特徴に迫った。この結果は、CAN-DO statements に基づいた機能的なシラバスと伝統的な語彙・文法シラバスを橋渡しする可能性がある。例えば、B1レベルで完了形が多いという結果は、そのレベルを目指す学習者にとって現在完了形や過去完了形はマスターしなければならない事項であることが明らかになり、どちらのシラバスに重点を置いた場合であっても、教員および学習者の明示的な目標となる。

今後、教材コーパスを有効に利用していくためには出版社、(教材等の)著者、研究者間で連携を取ることが大きな課題となる。そのためのひとつの方法として、教材コーパスのシンポジウムを、継続的かつ定期的に続けていくことが重要だろう。

付 記

本稿は英語コーパス学会第40回大会シンポジウム「英語教育・研究のための教材コーパスの構築と利用：実践例と課題」での口頭発表の内容を基に作成したものである。発表時に有益なコメントをくださったシンポジウムの参加者、また建設的なご意見をくださった匿名の査読者に深く感謝する。

なお、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, 代表: 投野由紀夫)の助成を受けたものである。

参考文献

- Altenberg, B. and S. Granger(2001) " The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. " *Applied Linguistics* 22-2, pp. 173-195.
- Hill, J.(2000) " Revising Priorities: From Grammatical Failure to Collocational Success. " In Lewis, M.(ed.) *Teaching Collocation: Further Development in the Lexical Approach*. Hampshire: Heinle, Cengage Learning, pp. 47-69.
- Laufer, B. and T. Waldman(2011) " Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. " *Language Learning* 61-2, pp. 647-672.
- Lewis, M.(2000) " There is Nothing as Practical as a Good Theory. " In Lewis, M.(ed.) *Teaching collocation: Further development in the lexical approach*. Hampshire: Heinle, Cengage Learning, pp. 10-27.
- Negishi, M., T. Takeda and Y. Tono(2012) " A Progress Report on the Development of the CEFR-J. " *Studies in Language Testing* 36, pp. 137-157.
- Sinclair, J. M.(1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- 相澤一美・石川慎一郎・村田年(編)(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。
- 内田諭(2015) 『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』『言語処理学会第21回年次大会発表論文集』言語処理学会, pp. 880-883.
- 後藤一章(2007) 『統語機能別頻度分布に基づく名詞的特徴的コロケーションの発見』『多変量解析を用いたテキスト分析研究』統計数理研究所共同研究リポート 201, pp. 1-23.
- 田畑智司(2009) 『Gentleman in Dickens 多変量アプローチで見る文体意匠としてのコロケーション』『多変量アプローチによるテキストの計量研究』統計数理研究所共同研究リポート 231, pp. 1-22.
- 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。
- 堀正広(2011) 『例題で学ぶ英語コロケーション』研究社。
- 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, pp. 31-45.