BATJ Journal No.16: pp.63-76

第 17 回 BATJ 大会：基調講演

コーパス研究が切り開く新しい日本語教育

筑波大学人文社会系　　李在鎬

1. はじめに：コーパスとは何か

言語研究のために設計・構築された文章の集合を「コーパス」（corpus）と言います。コーパスとは何であり、日本語教育の中でどのような利用可能性が見込まれるかについては、本稿全体を通して説明したいと思いますが、この節では、コーパスのイメージをつかんでもらいたいと思います。

さて、コーパスの中身はどんなふうに構成されているのでしょうか。一言でいえば、大量のテキストファイルが入っているものです。そして、テキストファイルの出典が記録されたレジスタファイルがセットになっているデータベースをイメージすれば良いと思います。この大量の電子ファイルをコンピュータで探索していくのがコーパスを使った言語研究の基本的なやり方です。イメージとしては、Google サイトを使って、キーワード検索をするのと似ています[1]。

さて、コーパスの中身を構成するテキストファイルにはどんな種類のものがどうやって収録されているのでしょうか。まず、「どんな種類」ということに関して言えば、話し言葉の会話だけを集めたものもあれば、書き言葉の新聞記事だけを集めたものや、その両方を集めたものもあり、多種多様です。次に「どうやって」ということに関して言えば、特定のジャンルを大規模に集めるというやり方といろいろな文章をバランスよく取り入れるというやり方があり、利用目的によって違う設計をします。2 節以降では、このコーパスとは何であり、どのように使うのかということについて説明したいと思います。まず、2 節ではなぜコーパスを使うのか、3 節ではコーパスとして認定するための条件について考えます。そして、4 節と 5 節では、コーパス研究の手法としての性質と理論としての性質をそれぞれ見ます。最後に、6 節では日本語教育に対してどのような示唆があるか、考えたいと思います。

2. コーパスによって可能な言語研究

コーパスを使う最大にして唯一の理由は、個人の言語的直感では得られない科学的・一般的な言語事実が発見できる点です。多くの理論言語学では、分析者自らが用例を作る形で分析を進めますが、しかし、どんなに優れた言語的直観を持った分析者でも、個人の創造性には限界があり、必然的にデータが不足していたり、観察が偏っていたりするものです。しかし、大規模な言葉のデータベースであるコーパスを使うことで、個人の言語感覚に頼ることなく、科学的な手続きによって仮説を検証したり、時には新たな事実を発掘したりすることもできます。

そして、こうした手続きによって得られた知見を言語教育に活かすことは、非常に意味のあることと言えます。

さて、大規模なデータベースといっても、それがどんなものかあまり実感がわかないかと思うので、具体的な例を挙げてみます。近年、コーパスを使う研究分野では、1 億語のデータベースであれば大規模なコーパスであると言います。日本語のコーパスで言えば、国立国語研究所が作成した「現代日本語書き言葉均衡コーパス」（以下、BCCWJ）がそれに該当します。では、1 億語とはどの程度のサイズでしょうか。物理的な広さであれば、東京ドームの何倍といった記述ができますが、コーパスは物量ではないので、少し違う指標が必要になります。ここでは、1 日の発話量という観点から捉えてみることにします。

さて、皆さんは 1 日の中で、どのくらいの語を産出していると思いますか。過去に国立国語研究所で調べた結果によると、都会で生活する一般成人の 1 日の平均的な発話量は、おおよそ 1000～1500 語だと言われています。最大数をとり、1500 語で 1 億語を計算してみてください。まず、日数で言えば、66,667 日間において話したデータ量ということになります。さらに年数で言えば、約 183 年間において話したデータ量ということになります。これで 1 億語の巨大さがある程度、伝わったのではないでしょうか。

こうした大規模なデータを使うことで、個人の言語的直観では得られない言語的事実を発見することができます。具体例を挙げます。もし、あなたが自分のクラスの学生に「食べにくい」と「食べづらい」の意味を説明してくださいと求められたらどうしますか。おそらく「食べる」の意味については、簡単に説明できるだろうと思いますが、うしろの「にくい」や「づらい」の意味の違いを聞かれて、即答できる人は少ないだろうと思います。これらはいわゆる難易表現と呼ばれているものです。この難易表現を上述した BCCWJ を使って調べてみます。調べる時のポイントとして、どのような動詞と一緒に使われるかを調査し、全体的な意味の違いを調べてみることにします。次の表 1 はその結果です。

表 1: 「～づらい」と「～にくい」の使用例の上位 20 位

順位	「～づらい」の例	使用頻度	「～にくい」の例	使用頻度
1	分かりづらい	215	分かりにくい	853
2	言いづらい	81	見えにくい	255
3	使いづらい	79	言いにくい	251
4	読みづらい	70	考えにくい	248
5	取りづらい	38	出にくい	164
6	聞きづらい	37	入りにくい	151
7	入りづらい	28	使いにくい	150
8	見えづらい	22	取りにくい	130
9	生きづらい	19	読みにくい	125
10	話しづらい	18	扱いにくい	93
11	考えづらい	16	聞き取りにくい	85
12	歩きづらい	14	受けにくい	79
13	出づらい	13	落ちにくい	62
14	書きづらい	13	歩きにくい	60
15	食べづらい	12	起こりにくい	55

16	住みづらい	11		滑りにくい	52
17	絡みづらい	11		溶けにくい	52
18	扱いづらい	10		住みにくい	47
19	答えづらい	9		答えにくい	47
20	動きづらい	8		直りにくい	45

表 1 では、BCCWJ での各用例に対する使用頻度の高い順から「づらい」と「にくい」の上位 20 件を挙げています。コーパスを使った調査では、どの程度、使われているのか、どの程度出現しているのかという事実を大切にするため、使用頻度をもとに分析します。

表 1 から観察される事実として、まずは次の 2 つが挙げられます。1 つ目に、使用頻度そのものを見た場合、「づらい」よりは「にくい」のほうが多く使われています。すなわち日本語の中では、「にくい」のほうがより一般的であり、学習者にとって目にする確率が高いということが言えます。2 つ目に、「～づらい」と「～にくい」は、大部分の動詞に関して一緒に使えることです。この 2 つ目の側面から、「～づらい」と「～にくい」は類義語の関係にあることが分かります。

では、意味の違いはどう捉えれば良いでしょうか。意味の違いを見るためには、一方の要素と高頻度で使われている動詞に注目します。そうすると「づらい」の特徴的な傾向として「話す」「書く」「聞く」「食べる」のように身体を使った身近な動作を表す語と一緒に使われることが多いことが分かります。次に「にくい」に関しては「落ちる」「溶ける」「起こる」「滑る」のように自動詞と一緒に使われることが多いということが見て取れます。このことから考えると、「づらい」は人間が意思をもって行う日常的行為の難易を表すことが多い表現であり、「にくい」は自然発生的に起こる出来事の難易を表すことが多い表現であると言えるのではないでしょうか。こうした事実は、個人の言語的直観ではなかなか見えてこないものであり、また、明示的に分析することが難しいのではないでしょうか。

3. コーパスの条件

コーパスに基づく言語研究の総称として「コーパス言語学」という分野があります [2]。このコーパス言語学では、2 節の表 1 で見たように出現頻度などの数量的特徴に基づいて言語研究をします。コーパス言語学の分野では、コーパスは、研究を行うための重要な資源であると認識されているため、コーパスとしての定義が存在します。つまり、何をもってコーパスと呼べるかということに関する共通認識が存在します。

まず、何をコーパスと呼ぶかについては、2 つの見方が存在します。1 つ目に、言語研究のためのテキストデータであれば良いとする立場があります。これは、コーパスを広い意味で解釈しているので、広義のコーパスと言えます。広義のコーパスでは基本的には研究目的さえあれば、何をコーパスと称するかについては、あまりこだわりがないということになるので、これ以上の説明は不要ということになります。2 つ目に、何からの条件を課していて、それを満たすものだけをコーパスと呼ぶという立場があります。これは、コーパスを狭い意味で解釈しているので、狭義のコーパスと言えます。では、どのような条件を課しているのでしょうか。まず、第 1 の条件としては「代表性」、第 2 の条件としては「大規模性」です。まず、「代表性」の条件とは、コーパスを作る過程で収録するデータをどのように決定するかということが問題

になります。次のような事例を考えてみましょう。コーパスAは、田中さんというサラリーマンが東京の通勤電車で読んでいる本をすべて集めてコーパスにしました。コーパスBは、東京にある公立図書館の蔵書リストと書店の在庫リストにある本をすべて集めてコーパスにしました。当然のことながら、コーパスAとコーパスBは、量的にかなり違います。しかし、ここでは量の違いはさておいて、質の違いを考えてみましょう。コーパスAは田中さんの本の好みによって収録されるデータが大きく左右されます。また、通勤電車という環境の制約も受けることになります。さらに、サラリーマンという職業上の制約もあり、あまり専門的な本は対象にならないだろうと思います。一方、コーパスBはどうでしょうか。公立図書館となると、様々な利用者に対応せねばなりませんので、個人の好みによる偏りの問題は解消されます。また、図書館と書店という複数の環境を取り入れることにより、環境的な制約も比較的受けにくいと考えられます。また、公的施設であるため、利用者の職業による偏りも出ないだろうと思います。

コーパスAとコーパスBはかなり対極的な例ではありますが、コーパス言語学の分野では、コーパスAは代表性がないと言い、コーパスBは(完全ではないが)代表性があると言います。一般的なコーパス言語学の研究者は、代表性があるコーパスBのほうを好んで使います。それはなぜでしょうか。コーパス言語学の研究者がコーパスを使う大前提として、「コーパス＝当該言語の縮小図」という暗黙知があります。つまり、コーパスは、その言語をミニマルに再現していて、それを調べることによって、その言語の実態が明らかにできる、という前提が必要です。この前提があってこそ、コーパスを使う意味があるのです。以上の理由から狭義のコーパスとしての第1要件として、代表性が重視されます。この代表性は、コーパスデータの信頼性・妥当性を考える上で重要な指標であり、この代表性を持つコーパスを均衡コーパス（Balanced Corpus）と呼びます[3]。

次に、「大規模性」について考えてみましょう。大規模性とは、コーパスの物理的なサイズに関することです。この大規模性は、コーパスの研究資源としての客観性、科学性を保証する性質であると考えられています。コーパスの物理的なサイズの大小によって何が違うのでしょうか。第1に考えられるのは、検索語を入れた場合の用例のヒット件数です。小規模であれば、ヒット件数は少ないだろうし、大規模であれば、件数は増えます。表2は、石川（2012）が報告しているものです。

表2: 異なる大きさのコーパスにおける語の出現件数

		検索語			
		animal	large	run	Gradually
英語コーパス	Brown（100万語）	72	378	246	51
	BNC（1億語）	6,634	33,034	21,547	3,592
	ukWaC（15億語）	109,131	467,724	356,555	35,244
	enTenTen（32億語）	211,862	790,802	626,372	64,591
		動物	大きい	走る	徐々に
日本語コーパス	BCCWJ（1億語）	9,362	9,757	3,756	2,607
	JpWaC（4億語）	26,505	30,373	9,980	7,825

（　）内の数値はコーパス全体の延べ語数

表 2 では同じ語を異なる大きさのコーパスで検索し、出現の頻度を示しています。それを見ると、animal については、Brown コーパスで 72 件、BNC コーパスで 6,634 件など、数が違うことが分かります。それでは、このヒット件数は何に関係してくるのでしょうか。それは、分析結果の安定性につながります。ヒット件数が少ないと、限られた標本をもとに分析を行うことになり、分析の結果も不安定になります。なぜなら、限られた標本によって得られた結果は、偶然性の影響を受けやすいからです。しかし、ある程度の大きさの標本数で分析を行った場合、相対的に偶然性の影響を受けにくく、調査結果も安定します。ただ、用例を目視で細かく検討する必要がある研究では、あまりヒット件数が多すぎても、研究の障害になります。大まかな傾向を確認する程度であれば、すべての用例を見る必要はないと思いますが、一つ一つの用例を細かく吟味する必要がある場合は、数千レベルが限度ではないかと思います。標本数やコーパスサイズの問題は、研究目的や実施可能性とのバランスを見ながら決めていく必要があると言えます。

4. 研究手法としてのコーパス

コーパス言語学は、どのような特徴を持っているのでしょうか。この疑問に答えるためには、いわゆるコーパスを使わない言語研究の手法と比べてみる必要があります。以下では、中本・李（2010）をもとに説明します。

中本・李（2010）では言語研究のために、データをあつめ、分析する作業を 3 つの次元で分類しています。第 1 の次元は「作例によるデータ収集か実例による収集か」、第 2 の次元は「研究者の内省による判断かそれ以外の方法による判断か」、第 3 の次元は「質的な方法による分析か、量的な方法による分析か」です。

まず、第 1 の次元は、データとなる言語表現をどこから、どのように収集したかという問題です。作例基盤の研究では研究者自身が自らの仮説に従って、用例を作成するのに対し、実例基盤の研究ではコーパスなどの既に存在する言語表現の中から収集します。作例基盤の研究はある種の思考実験的な性質を持っていますが、実例基盤の研究は、実験的な性質を持っています。

次に、第 2 の次元は、収集した言語表現に対して何らかの分析（測定）をする時、研究者自身の言語的直観、すなわち内省を使うのかそれ以外の方法を使うのかという問題です。研究者自身がある文が適当な表現として認められるかどうかを判断することで議論を進めていくような場合は、研究者の内省による研究であると言えます。それと対照されるタイプとしては様々なものが考えられます。まず、コーパスに準拠するタイプとしては次のような例が考えられます。特定の語句がいくつ含まれているかといった頻度の計測、どのような語句と一緒に使われるかの共起数を数える共起頻度の計測、文中での特定の語句の出現位置の計測などが挙げられます。次に、被験者を使うタイプとしては、次のような例が考えられます。何らかの言語表現（作例か実例かは問わない）に対して、第 3 者に容認性を判断してもらったり、その表現の意味を述べてもらったり、その表現を読んで理解するのにかかる時間を計測したりするタイプがあります。このように言語使用者である被験者の反応から収集されるデータを行動データ（behavioral data）と呼びます。コーパスに準拠した頻度データも、被験者の反応に準拠した行動データも研究者自身による内省ではない点で、客観データと呼ぶことができます。

最後に、第 3 の次元は、「この語句の意味は～である」といった形で表されるようなものか、観察あるいは計測の結果が数値として得られる（その結果統計的な処理が可能になる）ものか

という問題です。ここでは、前者が質的、後者が量的な研究であるとします。
以上の 3 つの次元を組み合わせると表 3 のようになります。

表 3: 言語研究の方法論

タイプ	第 1 次元	第 2 次元	第 2 次元	内容
1	作例	内省	質的	従来の理論言語学で主流であった方法
2	作例	内省	量的	基本的には存在しない(容認可能性に程度を認めて判定を行う研究はこれに相当するかもしれない)
3	作例	客観	質的	方言やまだ記述されていない言語について複数のインフォーマントに対するインタビューを通して記述を行う場合など。
4	作例	客観	量的	心理学的な実験や調査を行う研究の多くが含まれる
5	実例	内省	質的	コーパスの事例を研究者自身が内省により分類し、用法を特定するような場合
6	実例	内省	量的	コーパスの事例に対し研究者自身が内省に基づいてコーディングを行い、多変量解析で分析を行う場合
7	実例	客観	質的	コーパスの事例に対して、複数のインフォーマントから情報を得て、用法や用例の特徴を定性的に記述するような場合
8	実例	客観	量的	自然言語処理の技術などを使用して、コーパスの語の頻度や共起関係を利用した分析を行う場合

本章で取り上げているコーパスを使った研究は、タイプ 5〜8 のものに該当しますが、5 や 6 のように分析者の内省を取り入れたタイプもあれば、7 のように行動データ的な手法を取り入れることもあれば、8 のように完全なプログラムによる処理で行う研究タイプもあります。一方の 1〜4 の研究タイプについては、作例に基づく研究ということになりますが、こうしたタイプの研究の場合、統制すべき要因が明示的な場合にのみ、有効と言えます。つまり、統制すべきパラメーターがはっきりしない場合は、思考実験で終わる危険性もあります。コーパスの場合、データに対する探索的な研究ができるので、理想的には、1〜4 の手法と 5〜8 の手法を組み合わせることが求められます。

5. 研究の手法から理論へ

コーパス研究は、4 節で紹介した通り、言語研究のための方法論の 1 つとして理解されています。しかし、近年、Neo-Firthian と呼ばれている研究グループは、コーパス研究が言語の理論モデルの 1 つになりうることを示しています（具体的には McEnery & Hardie 2012 を参照)。Neo-Firthian の研究で、重要なキーワードになるのが「コロケーション」と「談話」です。そして、研究姿勢において特徴的と言えるのが、「言語の使用文脈でもって現象を捉える」ところです。

5.1 コロケーション分析

コロケーション分析とは何でしょうか。コロケーション分析では、どのような考え方のもとで言語事実にアプローチしているのでしょうか。その考え方を理解するためには、次のことを抑えておく必要があります。コロケーション分析では、ある語が別の語と「一緒に使われる」という事実に注目します。そして、「一緒に使われる」事実には「強弱」もしくは「濃淡」があると考えます。つまり、ある単語のよく使われる組み合わせや自然な語の連鎖をデータに基づいて捉えるのです。

コロケーション分析そのものは、コーパスを使わなくても可能です。前節の表 3 のタイプ 1 もくしはタイプ 3 の手法を用いてコロケーション分析をすることも可能です。しかし、タイプ 1 やタイプ 3 の場合、分析者個人の内省に基づく研究ですので、網羅性と記述の一般性の部分で課題が残ります。この課題を解決するため、コロケーション分析では、コーパスを利用します。以下では、McEnery & Hardie（2012）の議論にそって、コロケーション分析の背後にあるコーパス研究の考え方を紹介します。そして、これを通して、コーパス研究は言語理論の 1 つとして位置付けられることを示したいと思います。

5.2 コロケーション分析と語の意味

コロケーション分析では、語の意味はどこに存在するのかという問題に対して、少しユニークな考え方をします。私たちは、語の意味を知りたい時に、辞書を調べます。辞書を調べれば、見出し語に対して何らかの説明が与えられているので、その語の意味を知ることになります。これは、語という単位に何らかの固有の意味を内在化させ、記述することができることを前提にしています。

しかし、コロケーション分析では、語の意味を語そのものに単体として内在化させることはしません。コロケーション分析では、語の使用環境、すなわち当該語と高頻度に共起する他の語や構造との特徴的な関係の中で語の意味を記述することを目指します。これは、Firth（1968）が指摘する「統辞レベルにおける抽象関係（an abstraction at the syntagmatic level）」に基づく考え方です。

ただ、何をコロケーションと呼ぶかについては、必ずしも普遍的な定義があるわけではありません。というのは、いわゆる慣用句的なものだけをコロケーションと呼ぶ立場もあれば、一定の頻度で共起する語と語の組み合わせをコロケーションと呼ぶ立場もあるからです。コーパス研究では、後者の立場で捉えることが多く、コロケーションはクリアカットに規定できるものではないと認識されています。つまり、連続的に規定されるべきものと考えられています。

5.3 3 つの論点

コロケーション分析を理解するためには、次に述べる 3 つのポイントを押さえておく必要があります。

- 第 1 ポイント：どんなデータをもとにコロケーションを定義するか
- 第 2 ポイント：コロケーションをどのような手続きで決定するか
- 第 3 ポイント：コロケーションはどこに存在するか

まず、第1のポイントについて説明します。これは、コロケーションは用いるデータによって変わることが知られています。表4のデータで説明します。

表4: 「積む」のサブコーパス別のコロケーション

順位	全体	書籍	雑誌	国会会議録	知恵袋
1	経験（268）	経験（191）	経験（20）	経験（10）	経験（39）
2	石（46）	石（44）	キャリア（8）	金（4）	エンジン（7）
3	修行（41）	修行（38）	エンジン（8）	実績（4）	金（5）
4	荷物（35）	訓練（29）	トレーニング（5）	割（3）	荷物（5）
5	訓練（35）	荷物（26）	燃料（5）	研修（2）	修行（3）
6	金（28）	修業（22）	バイク（4）	悪行（2）	キャリア（3）
7	修業（28）	トレーニング(21）	葉（4）	研さん（2）	メモリ（3）
8	キャリア（27）	修練（20）	荷物（3）	訓練（1）	訓練（2）
9	トレーニング(26）	体験（18）	修業（3）	荷物（1）	燃料（2）
10	エンジン（24）	金（17）	石（2）	修練（1）	箱（2）
11	修練（23）	練習（15）	訓練（2）	燃料（1）	鍛錬（2）
12	体験（20）	キャリア（15）	金（2）	もの（1）	水（2）
13	練習（19）	研鑽（14）	練習（2）	兵器（1）	CPU（2）
14	研鑽（14）	石垣（11）	レンガ（2）	土（1）	研修（1）
15	実績（12）	荷（11）	死体（2）	研究（1）	修練（1）
16	徳（12）	稽古（10）	バッテリー（2）	つくし（1）	もの（1）
17	燃料（12）	石炭（10）	実績（1）	機材（1）	つくし（1）
18	花（11）	善行（9）	花（1）	ファンド(1）	機材（1）
19	荷（11）	エンジン（9）	物（1）	引き当て(1）	体験（1）
20	石垣（11）	実績（7）	ブロック（1）	学歴（1）	徳（1）

（ ）内の数値は延べ頻度

表4は、「中納言」（https://chunagon.ninjal.ac.jp/）を使ってBCCWJで「積む」の用例2636例を取り出し、「Nを積む」というコロケーションパターンに対して、調査したものです。（ ）に示した延べ頻度は「積む」のコロケーションをサブコーパス別に数え上げたもので、上位20位までを挙げています。左から２列目のセルにBCCWJ全体における順位と延べ頻度を、そして、左から3列目に書籍[4]、4列目に雑誌、5列目に国会会議録、6列目にウェブデータとしてYahoo!知恵袋での延べ頻度を示しました。

表4で注目すべきは、次の3つです。1）いずれのサブコーパスでも「経験を積む」というコロケーションがもっとも多いこと、2）出来事を表す抽象名詞（経験、修行、訓練、体験、実績）との共起例が多いこと、3）サブコーパスによる相違として、雑誌ではカタカナ語が上位に来ていること、Yahoo!知恵袋では「メモリ」や「CPU」のようにコンピュータ用語が出ていることです。1）と 2）はサブコーパスによらない普遍的な現象と言えますが、3）はサブコーパス、すなわち使用のコンテキストに依存するもので、個別的な現象と言えます。こうした普遍性と個別性が交差する側面こそが、コロケーション分析の醍醐味ですし、BCCWJのような均衡コーパスがあるからこそ、実現できる研究と言えるのではないでしょうか。

次に第 2 のポイントについて説明します。これは、コロケーションをどのような手続きで決定するかという問題です。この問題については、2 つのアプローチが用いられています。1）キーワード検索をしたあと、ヒットした用例を目視することで、手作業で決めていく方法、2）ヒットした語同士の頻度を総合的に分析し、数学的な方法でコロケーション性を決めていく方法の 2 つです。1）の方法では分析者の内省に依拠して結果が表現されるのに対して、2）の方法では、コーパスが同じであれば、同じ分析結果が再現できるというメリットがあります。なお、1）の手法に関しては、カイ二乗統計量や対数尤度比や相互情報量などを使ったスコア法が提案されており、指標によって取り出されるコロケーションがかわることもよく知られています（詳細は石川 2012 を参照）[5]。

表 5: 複数の指標に基づく「積む」のコロケーション

順位	コロケーション	頻度	コロケーション	MI スコア	コロケーション	ログダイス
1	経験を積む	246	福運を積む	14.02	経験を積む	9.28
2	石を積む	44	研鑽を積む	13.85	研鑽を積む	9.14
3	荷物を積む	39	土嚢を積む	13.46	修練を積む	8.9
4	修行を積む	35	習練を積む	13.34	修業を積む	8.87
5	訓練を積む	32	銀塊を積む	13.02	修行を積む	8.8
6	トレーニングを積む	29	鍾乳石を積む	13.02	荷物を積む	8.31
7	金を積む	29	魚探を積む	13.02	トレーニングを積む	8.22
8	研鑽を積む	25	修練を積む	12.86	キャリアを積む	8.1
9	修業を積む	25	ボールベアリングを積む	12.64	善行を積む	7.8
10	練習を積む	24	善根を積む	12.58	徳を積む	7.75

表 5 は、NINJAL-LWP for BCCWJ（http://nlb.ninjal.ac.jp/）で「N を積む」パターンで用例検索し、頻度と MI スコアとログダイスを降順で並び替え、上位 10 位ずつを示したものです[6]。3 つの指標を比べた場合、 MI スコアと頻度の間には差が大きく、「経験を積む」は高頻度のコロケーションということになりますが、MI スコアとしては 10.77、一方、「研鑽を積む」の場合、実際の頻度としては 25 回しかないのに、MI スコアとしては 13.85 と高い値を示しています。これは、MI スコアの場合、両方とも高頻度の語である場合、スコアが低くなる性質を持っているからです。

最後に、第 3 のポイントについて説明します。これは、コロケーションの存在論的位置づけに関する問題です。この問題については、生成文法の考え方と対比させながら説明します。生成文法では、語はそれ自身で意味を参照することはなく、形式的・数学的ルールによって生成された統語構造内のスロットに投入されるだけであると考えられてきました。これに対して、Sinclair（2004）などではコロケーションパターンこそが意味の中核をなすと考えています。そして、言語現象における意味の重要性、コンテキストの重要性をコロケーション研究によって明らかにしているのです。

Sinclair（2004:18）では、自然言語の文章はコロケーションの連続によって成り立っていると指摘しています。そして、言語の生成メカニズムとして、話者は語を選択しているのではなく、

意味の単位を選択している、意味の単位は複数の語から成り立っていると主張しています。こうした考え方に基づき、「文法とは意味の文法であって、語の文法ではない」と主張しています。さらに、複数の語の連結、すなわちコロケーションによって生じる意味は、それらの語が個別的に持つ意味とは異なるとも述べています。

6. 学習者コーパスを利用した日本語教育研究

5 節で取り上げた BCCWJ のようなコーパスは、日本語母語話者の産出データを集めたものです。これとは別に、特定の外国語を学習する人たちの産出データを集めたコーパスがあり、それは「学習者コーパス」（Learner Corpus）と総称します。これは、いわゆる母語話者によって産出された母語話者コーパスに対立するものとして理解されています。

学習者コーパスに注目した最初の研究が Granger（1998）でありますが、そこでは、学習者コーパスの研究意義を次のように述べています。「コーパス言語学の手法やツールを使ってオーセンティックな学習者の言語をもっと深く洞察するもの」。学習者コーパスを使うことで、いわゆる教師や研究者目線ではない、学習者自身による真の言語使用実態を見ることができるのです。

6.1 学習者コーパス

Leech（1998）では、学習者コーパスに基づく習得研究の具体的な研究タスクとして、以下の 5 つを提案していますので、紹介したいと思います。

1. 母語話者と比較した場合、学習者が有意に「過剰使用」または「過小使用」しがちな目標言語の言語的特徴にどんなものがあるか。
2. 学習者の目標言語における振舞いには、母語からの影響（母語転移）がどの程度あるか。
3. 学習者が目標言語で十分に表現できない場合は、「回避ストラテジー」を使うが、その言語領域にはどんなものがあるか。
4. 学習者が母語話者的に運用したり、あるいは非母語話者的に運用したりする言語領域にはどんなものがあるか。
5. 非母語話者的に運用する言語領域で A 国の学習者が苦手とし、特別な助けを必要とする言語領域にはどんなものがあるか。

この 5 つのトピックは、いわゆる学習者コーパスを使った第二言語習得研究の具体的な研究課題であると同時に、学習者コーパスを使った研究の方向性とも言えます。言語教育に関わる研究者は、学習者コーパスを使うことで、様々な語や表現の使用傾向を考察することができます。そして、各データには母語や習熟度などの情報が紐づけられていることが多いため、使用傾向と母語の問題の関連を検討することもできます。こうした研究の蓄積によって、学習者の中で何が起きているかを捉えることができますし、その成果は、言語教育においてもダイレクトに活用できるわけです。

6.2 研究の現状と展望

学習者コーパスを使った日本語教育研究としては、いろいろなものがあります。

まず、学習者コーパスの紹介を主な目的とする研究としては次のものがあります。話し言葉の学習者コーパスとしては、「KY コーパス」が広く知られています。この KY コーパスを紹介した研究として、鎌田（2006）があります。そして、KY コーパスにタグを付与する「タグ付き KY コーパス」というコーパスがあります（詳細は、李 2009 を参考）。作文コーパスとしては、「作文対訳 DB」があり、宇佐美・籠宮・椙本（2001）で詳細を確認することができます。そして、最近、公開されたデータとしては、「日本語教育のためのタスク別書き言葉コーパス」があり、金澤（2014）においてコーパスそのものの開発プロセスやコーパス構築で利用したタスクの解説、さらにはコーパスを利用したケース・スタディまで収録しています。

次に、学習者コーパスを利用した研究についていくつか紹介したいと思います。

まず、音声の習得を扱った研究は非常に少ないですが、河野（1999）では動詞のアクセントの習得に関して、戸田・カッケンブッシュ（1999）では外来語アクセントの形成について、それぞれ考察しています。

次に、語彙・文法の習得研究は、学習者コーパスを利用し、もっとも活発に研究されている分野です。例えば、野田（2001）では英語・中国語・韓国語という異なった母語の学習者全てに共通する仮説構築の過程を検証しています。文法の研究として、コソアの習得過程を考察した迫田（2001）、アスペクト表現「ている」についての研究である許（2000）、 崔（2011）などがあります。山内（1999）は、文の習得という観点から学習者のレベルの違いを考察しています。伊藤（2012）は対のある自他動詞（「あく」対「あける」）の誤用を分析し、レベルによって誤用に質的な違いがあることを明らかにしています。KY コーパスを自然言語処理の方法で調査したものとして李・井佐原（2006）があります。李・井佐原（2006）では、クラスター分析を使い、KY コーパスに表れる助詞「に」の用法を分類しており、助詞「に」の共起語に見られる傾向と習得レベルの関連性を考察しています。

次に、コミュニケーションの方策としての日本語習得を扱う論文について紹介します。小林（1999）は質問-応答の発話連鎖に注目し、レベルが上がるに従って学習者の質問が質的に変化することを指摘し、コミュニケーションを円滑にすすめる質問ができるのはある程度レベルが上がってからであることを発見しました。そして、言語テストとコーパスの関連性を論じた研究としては李・宮岡・林（2013）があります。そこでは言語テストの得点により、作文の到達度に差があることを明らかにしています。

最後に、学習者コーパスを利用したこれからの研究について述べたいと思います。ここで特に注目したのは、テキストマイニングを利用した研究です。テキストマイニングとは、テキストデータに対する統計的な分析のことを指します。これによって、情報を発掘したり、一般化したりする研究領域です。このテキストマイニングの分野では、既存のデータから何らかの学習モデルを構築し、新しいデータを予測するための分析が活発になされています。これらの分析モデルは、第二言語習得研究においても様々な活用ができます。例えば、学習者コーパスに含まれている文字データから産出者の属性を予測するタイプの研究が可能です。具体的に、李・鎌田（2013）では、産出データから初級、中級、上級といった習熟度を予測できることを示しています。こうしたアプローチによる研究が進んだ場合、学習者の産出データを自動評価するシステムの開発につなげることもできます。

7. 終わりに

本稿では、母語話者コーパスを利用した研究事例を紹介しました。そして、学習者コーパスを利用した日本語教育研究の可能性についても考えてきました。これらは、日本語教育の実践と研究におけるコーパス利用の可能性を示唆するものです。例えば、母語話者コーパスを利用することで、真正性のある教育コンテンツの作成ができます。そして、現在、教材開発やテスト開発に関して様々な利用が試みられています。さらに、学習者コーパスを利用した日本語教育研究を実践することで、学習者目線で教育コンテンツを選定することが可能になります。

最後に、近年の動向に関連して、2 点述べたいと思います。一点目に、近年、コミュニケーション能力の向上を言語教育の目標として定める傾向が非常に強くなってきました。言語学習に関する学習者のニーズも多様化していることを受け、日本語教育の中でも様々な工夫が求められています。本稿で取り上げたコーパスというツールは、絶対ありませんが、近年の日本語教育が抱えている課題を解決するための強力なツールになり得ると思います。二点目に、近年、コーパスに対する関心の高まりから、様々な「コーパス」という名前が付いた書籍が刊行され、科研費などの公的資金によるデータベースの開発プロジェクトが次々に始動しています。こうした動きは、第二言語習得や外国語教育研究の分野で、確認されています。特にコーパスを利用した教育コンテンツの開発や学習者データをあつめ、コーパス化するプロジェクトは、現在、たくさん出現しており、今後も活発化すると予想されます。こうしたプロジェクトでは、言語学、言語教育学、言語工学の集合知でもって課題解決することが求められており、競争の時代から共創の時代に移行しつつあるように感じます。そして、国内外の日本語教育関係者もこうした動きに積極的に参加する必要があるように感じます。

*謝辞：本稿は、第 17 回英国日本語教育学会年次大会における基調講演の内容をもとに執筆したものであります。基調講演においては会長の松本スタート洋子先生（エディンバラ大学）をはじめ、北川利彦先生（リージェンツ大学ロンドン）、ボールディング敏美先生（ケンブリッジ大学）、小木曽左枝子先生（カーディフ大学）に大変お世話になりました。また論文執筆においては村田裕美子先生（ミュンヘン大学）に大変お世話になりました。以上の先生方にお礼申し上げたいと思います。

注

[1] 研究者によってはWebをコーパスと見なす人もいて、Web検索で得られた結果を言語研究に使おうという考え方もあります。荻野（2014）をご参照ください。ただし、コーパス研究者の多くはWeb検索はコーパス研究とは違うと考えています。

[2] コーパスを使った言語研究については、分野によって「コーパス研究」（corpus research）、「コーパス言語学」（corpus linguistics）などの呼び方がありますが、コーパスデータの観察に基づく言語記述を目指すという意味では相違はありません。

[3] 均衡コーパスは複数のジャンルからランダムサンプリングなどの統計的な方法でコーパスに収録するデータを選択していきます。均衡コーパスを作るのには莫大なコストがかかるため、個人単位で作ることは非常に難しく、一般的には国家規模の研究資金を使って作ります。均衡コーパスの作り方の詳細は、李・石川・砂川（2012）をご参照ください。

[4] 書籍については、出版サブコーパスと図書館サブコーパスを合算した値を示しました。
[5] 使う指標によって、取り出される結果がかわることは、コロケーション指標の欠陥ではありません。用いるツールによって結果がかわることは、自然なことでありますし、結果の反証可能性を示すものであるため、むしろコロケーション指標がもつ科学性を示すものと理解することができます。
[6] MI スコアとは相互情報量とも言いますが、2 つの単語のうち、一方が与えられた時、もう一方の単語をどの程度予測できるかを指標化したものです。具体例として、BCCWJ で「戦闘帽をかぶる」は 1 例のみであるが、MI スコアは 15.9、「水をかぶる」は 46 例あるが、MI スコアは 6.5 である。この 15.9 と 6.5 の違いは、予測可能性の違いを示すものと言える。「戦闘帽」が与えられてから、後部要素として「かぶる」が来る可能性と「水」が与えられてから、後部要素として「かぶる」が来る可能性を比較した場合、次の事実に気づく。つまり、「水」に関しては「水を飲む」「水を浴びる」「水を入れる」「水をかける」など様々な可能性があるが、「戦闘帽」に関しては「かぶる」以外の語が生起する可能性はほとんどない。たくさんの可能性の中で 1 つのコロケーションが選択された場合は、MI スコアの値は低いが、その逆の場合は、MI スコアが高くなる。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。

参考文献

石川慎一郎（2012）『ベーシックコーパス言語学』，ひつじ書房.
伊藤秀明（2012）「学習者は『対のある自他動詞』をどのように使っているかー中国人日本語学習者の中級から超級に注目してー」，Journal of International and Advanced Japanese Studies (4)，43-52.
宇佐美洋・椙本総子・籠宮隆之（2004）「「日本語学習者による日本語／母語発話の対照言語データベース」の設計」,『電子情報通信学会技術研究報告』SP2004-24，29-34.
荻野綱男（2014）『ウェブ検索による日本語研究』，朝倉書店.
金澤裕之（編）（2014）『日本語教育のためのタスク別書き言葉コーパス』，ひつじ書房.
鎌田修（2006）「KY コーパスと日本語教育」,『日本語教育』130 号，42-51.
許夏珮（2000）「自然発話における日本語学習者による『テイル』の習得研究ーOPI データの分析結果からー」,『日本語教育』104 号，20-29.
小林ミナ（1999）「KY コーパスにあらわれた疑問詞疑問文ーインタビュー・パートにおける学習者からの質問に注目して」，カッケンブッシュ寛子『第 2 言語としての日本語の習得に関する総合研究』(平成 8-10 年度文部省科学研究費補助金基盤研究 (A) (1) 研究報告書[課題番号 08308019])
河野俊之（1999）「動詞のアクセントの習得」，カッケンブッシュ寛子『第 2 言語としての日本語の習得に関する総合研究』（平成 8-10 年度文部省科学研究費補助金基盤研究（A）（1）研究報告書［課題番号 08308019］）
迫田久美子（2001）「第 2 章 学習者の文法処理方法」，野田尚史・迫田久美子・渋谷勝己・小林典子著『日本語学習者の文法習得』大修館書店, 25-44.
崔亜珍（2011）「自然発話における日本語テンス・アスペクトの習得研究ーR 時の認識を中心にー」，『小出記念日本語教育研究会』19，5-19.
戸田貴子・カッケンブッシュ寛子（1999）「中間言語における外来語アクセントの形成と日本人話者による評価」，カッケンブッシュ寛子『第 2 言語としての日本語の習得に関する総合研究』（平成 8-10 年度文部省科学研究費補助金基盤研究（A）（1）研究報告書［課題番号 08308019］）
中本敬子・李在鎬（編著）（2010）『認知言語学の方法論入門』，ひつじ書房.
野田尚史（2001）「第３章　学習者独自の文法の背景」,『日本語学習者の文法習得』野田尚史・迫田

久美子・渋谷勝己・小林典子編著，大修館書店, 45-62.
山内博之（1999）「初級及び中級レベルにおける『文』の習得について」，カッケンブッシュ寛子『第2言語としての日本語の習得に関する総合研究』（平成8-10年度文部省科学研究費補助金基盤研究（A）（1）研究報告書［課題番号08308019］）
李在鎬（2009）「タグ付き日本語学習者コーパスの開発」，『計量国語学』27-2，60-72.
李在鎬・石川慎一郎・砂川有里子（2012）『日本語教育のためのコーパス調査入門』，くろしお出版.
李在鎬・井佐原均（2006）「第二言語獲得における助詞「に」の習得過程の定量的分析」，『計量国語学』25-4，163-180.
李在鎬・宮岡弥生・林炫情（2013）「学習者コーパスと言語テストー言語テストの得点と作文のテキスト情報量の関連性」，『言語教育評価研究』3，22-31.
李在鎬・鎌田修（2013）「OPIコーパスの妥当で効果的な使用について：コンピュータによる自動判定に向けての試案」（4th International Conference of the Italian Association for Japanese Language Teaching）
Firth, J.R. (1968). “A synopsis of linguistic theory 1930-1955.” in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman.(http://annabellelukin.edublogs.org/files/2013/08/Firth-JR-1962-A-Synopsis-of-Linguistic-Theory-wfihi5.pdf)
Granger, S. (*1998) Learner English on computer*. London: Longman.
Leech, G. (1998) “Learner corpora: What they are and what can be done with them”, in S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, xiv-xx.
McEnery, Tony and Andrew Hardie. (2012)*Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
Sinclair, John. (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.