国立国語研究所学術情報リポジトリ

日本語コーパスの紹介とその利用

メタデータ	言語: Japanese 出版者: 公開日: 2020-09-04 キーワード (Ja): キーワード (En): 作成者: 山崎, 誠 メールアドレス: 所属:
URL	https://doi.org/10.15084/00003009

日本語コーパスの紹介とその利用

山崎　誠
国立国語研究所

要旨

本稿では、日本語研究で用いられ、日本語教育にも役立つコーパスを紹介する。コーパスの構築の目的やどのようなデータが格納され、どのような研究情報（アノテーション）が付与されているかなどを概観するとともに、代表的な検索ツールとして『中納言』『NLB』を紹介する。また、日本語コーパスを利用した研究事例を紹介しつつ、コーパスの活用法や研究における位置付けについて述べる。

【キーワード】　コーパス、書き言葉、話し言葉、定性的研究、定量的研究
Keywords:　corpus, written language, spoken language, qualitative research, quantitative research

1　はじめに

近年、コーパスを初めとする言語データを言語資源という概念で捉えることが増えてきた。言語資源[1]とは、言語研究に利用する資源という意味で、コーパスだけでなく、解析用のツールや電子化辞書なども含むとされる（前川 2016: 2）。また、ELRA（European Language Resources Association）のホームページには、"What is a Language Resource?" というタイトルのもとに、以下のような説明がなされている[2]。

> The term **Language Resource** refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.
> Examples of Language Resources are **written** and **spoken corpora**, **computational lexica**, **terminology databases**, **speech collection**, etc. Basic software tools are also important for the acquisition, preparation, collection, management, customisation and use of these Language Resources and other resources.
> （http://www.elra.info/en/about/what-language-resource/）より

また、前川（2016）では、「言語の研究に使うデータやツールの望ましい姿に関する研究、そのデータやツールを作るための方法論の研究、そして開発されたデータやツールを使った言語の研究の全体」を「言語資源学」と位置付け、ひとつの研究領域として位置付けている。

2　コーパスとは

2.1　コーパスの定義

一般的なコーパスの定義は、以下のようなものになるだろう。

⑴ 実際に使われた言葉を、
⑵ 代表性を持たせるべく、
⑶ 大量に集め、
⑷ 研究用の情報を付け、
⑸ コンピュータで検索できるようにしたデータベース。

これらの条件の 1 つでも欠けるとコーパスとは言えない。⑵ の「代表性」とは、ランダムサンプリングにより、母集団から偏りのない抽出が行われているという意味の統計学の概念である。例えば、1 年分の新聞を母集団とする場合、日曜日の記事だけを抽出するとか、偶数ページだけ抽出すると抽出された記事に偏りが生じる可能性があるため、無作為に抽出することが望ましい。なお、代表性とは、価値判断を含んでいない。著名な作家の作品や文学賞受賞作品等をサンプリングで優遇することはしない。一方、コーパスにはゆるやかな定義も存在し、代表性を持たないテキストアーカイブ的なデータをコーパスと呼ぶ場合もある。新聞社が販売している新聞記事データベースは本格的な書き言葉コーパスが登場するまでは工学系の研究者を中心にコーパスとしてよく利用された。また、人文系の研究者では『新潮文庫の 100 冊』を利用することが多かった。これらは、1 つのジャンルからなるコーパスであり、後述の均衡コーパスとは異なる。

2.2　コーパスを利用する意義

コーパスを使う意義は、研究における客観性と再現性の確保が大きな理由である。コーパスとよく対比的に取り上げられる内省による研究は研究者自身の主観に左右されやすく、客観性が担保されにくい。また、コーパスを使った分析は、手法や手順を適切に示せば、第三者が結果を確認することが可能で、研究の透明性が高まることになる。客観性と再現性については、水谷静夫（1957: 61）において「良い調査に求められる条件」としても挙げられている。

3　コーパスの歴史

コーパスの開発は英語を対象としたものから始まった。大量のデータに基づく言語調査は、コーパス以前からあり、教育基本語彙の選定や速記法の研究のような実用的な目的で行われていた。実用という用途は現在のコーパスにも生きており、言語教育や辞書編纂などが主な用途である。

3.1　The Survey of English Usage (SEU)

最初のコーパスと言われるのが標題のコーパスである。この時点では、名称にコーパスという語は入っていない。教養層の英語の書き言葉・話し言葉のデータを集めたもので、5000 語×200 テキストで 100 万語（完成時）を収録している。また、当初は電子化されていなかった。

3.2　ブラウンコーパス（Brown Corpus）

電子化コーパスの嚆矢と言えるのがブラウンコーパスである。W. N. Francis と Henry Kučera が 1961 年に構築を始め、1964 年に完成した。正式名称は、The Standard Corpus of Present-day Edited American English である。このコーパスには、1961 年刊行の出版物におけるアメリカ英語を収録している。語数は、2000 語×500 テキストで約 100 万語である。このコーパスは、その後の汎用コーパスのモデルとなった。

しかし時代的に、1960～1980 年代のアメリカでは生成文法が言語研究の主流となり、コーパスを利用した研究は下火になった。生成文法では、理想の話者による言語能力の解明が主眼であり、データを必要としなかったためである。

3.3　LOB Corpus

その後、コーパス研究の主流はヨーロッパが主体となり、ブラウンコーパスにならって LOB Corpus が作られた。こちらは、1961 年刊行の英語の出版物から 2000 語×500 テキスト抽出、合計約 100 万語である。この設計はブラウンコーパスと同じである。

3.4　British National Corpus (BNC)

BNC は、出版社（OUP，Longman 等）、大学（Lancaster）の共同プロジェクトとして、開発された。1991 年に構築を開始し、1994 年に完成している。1975 年以降のイギリス英語を書き言葉で 9000 万語、話し言葉で 1000 万語収録している。コーパスの規模は約 1 億語で、ブラウンコーパスに変わって、これ以降の汎用コーパスのモデルになった。

3.5　日本におけるコーパスの歴史

日本でコーパスと名の付くデータが登場したのは、「京都大学テキストコーパス」（1997）が最初であろう。このコーパスは、1995 年発行の毎日新聞の記事４万文に対して、形態論情報や構文情報などを付加したものであるが、公開しているのは付加した情報（アノテーション）のみであり、利用者は別途毎日新聞記事データベースを購入する必要がる。自然言語処理では有用性が高く良く利用されている。

一方、コーパスという名称は付されていないが、国立国語研究所が行ってきた書き言葉、話し言葉の実態調査は、手法がコーパス言語学に近く、研究面では世界に先駆けていた面もあった。しかし、日本における一連のコーパス言語学的研究が世界的に認知されていないのは、英語での発信がなかったこと、また、国立国語研究所の定量的研究は、データを公開しなかったため、学界への普及、研究の発展を伴わなかったことがコーパス研究で世界に遅れをとった原因であろう。

4　コーパスの規模と種類

4.1　コーパスの規模

1960～1980 年代は 100 万語が中心であったコーパスの規模は 1990 年代以降、1 億語以上が標準的となった。21 世紀に入り、ウェブからコーパスを作るようになると 100 億語のオーダにまで拡大した。今後もコーパスの規模は増加しつづけると予想される。

4.2　コーパスの種類

「情報処理学事典」（p.68）には、コーパスの種類として、以下のようなものが挙げられて

いる。

言葉の種類：書き言葉／話し言葉
設計方針：サンプル／モニター[3]
利用目的：汎用／特殊目的
時代区分：共時／通時
言語の数：単言語／多言語
情報付与：生／注釈付き

5　日本語のコーパス

本稿では日本語研究および日本語教育研究有用なコーパスを採り上げる。

5.1　現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese: BCCWJ）

2011 年に公開された 1 億語の書き言葉コーパス。書籍、雑誌、新聞、ブログなど 13 個のレジスター（＝ジャンル）から構成されている均衡コーパスである。均衡コーパスとは、単一のジャンルではなく、さまざまなジャンルの言語を収録しているコーパスという意味である。データの年代はおよそ 1970 年代から 2000 年代である。『少納言』『中納言』という WEB 上のツールで検索できるほか、データ自体も有償で配布されている。

5.2　多言語母語の日本語学習者横断コーパス（International Corpus of Japanese as a Second Language: I-JAS）

「日本を含む 20 の国と地域で、異なった 12 言語を母語とする日本語学習者 1000 人の話し言葉および書き言葉を収集することを目標に」（HP[4]より）構築されたコーパスである。学習者の母語、年齢、性別、日本語能力などの背景情報が収録されている。『中納言』で検索可能。I-JAS は現在構築中であり、2018 年 7 月末現在の総語数は約 230 万語である。

5.3　BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2018 年版

相互行為としての会話を対人コミュニケーションや語用論の観点から分析するために構築されたコーパスである。2018 年版には 333 会話（約 79 時間）が収録されており、そのうち音声付きデータは 203 会話、約 40 時間である。会話参加者の年齢、性別、話題などが統制された形でデータが収集されており、様々な観点から比較・対照研究ができるようになっている。データは HP からダウンロードできる（要申し込み）。収録語数は約 93 万語である。

5.4　日本語話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ）

このコーパスは、日本語の自発音声（その多くはモノローグ）を大量に集めて、多くのアノテーションを施した話し言葉研究用のコーパスである。2004 年公開され、現在でも質・量ともに世界最高水準を保っている。収録語数は約 750 万語である。『中納言』で検索できるほか、データを有償で配布している。

5.5　名大会話コーパス（Nagoya University Conversation Corpus）

名古屋大学の大曽美枝子氏が科研費（2001 年～2003 年）で構築した日本人同士の会話の

コーパスで、129 会話、合計約 100 時間の雑談を収録している。語数は約 110 万語である。『中納言』で検索できる。書き起こしのみで音声は公開していない。

5.6　日本語歴史コーパス（Corpus of Historical Japanese: CHJ）

万葉集から昭和初期に至る、日本語史の主要な言語資料を収録したコーパスである。収録語数は 2019 年 2 月現在約 1560 万語である。『中納言』で検索できる。

6　主要な検索ツール

コーパスを利用するには、ツールが必須である。アノテーションが施されたコーパスの多くは、人間が目で読むためには出来ていないからである。ツールには、コンコーダンサーとワードプロファイラーの 2 つのタイプがある。前者は、検索した語を中心として前後の文脈が示される、いわゆる KWIC（Key Word in Context）形式のものであり、後者は検索語と一緒に使われる語（コロケーション）などを表示させるものである。

代表的なコンコーダンサーとしては『少納言』と『中納言』がある。少納言は、BCCWJ のみを対象としたもので、登録不要で誰でも使えるが、文字列検索しかできない。中納言は、登録制であるが、無料で、形態論情報に基づいた、より高度な検索ができる。例えば、玉葱の表記を調べたい場合、少納言では、可能性のある表記をすべてその都度検索しなければならないが、中納言では、語彙素（辞書の見出しに当たる）を指定して検索すれば、すべての表記が検索できる。

また、ワードプロファイラーとしては、NLB（NINJAL LWP for BCCWJ）あるいは、NLT（NINJAL LWP for TWC）がよく使われている。前者は BCCWJ を対象としたもの、後者は筑波大学が構築した筑波 WEB コーパスを対象としたものである。

図 1 に中納言の検索結果を、図 2 に NLB での検索結果をそれぞれ示した。図 1 では、「キー」の列に原文で出現した文字列が示され、それを挟んで前後に文脈が示される。前後の文脈の長さは 10 語から 500 語の範囲で指定できる。検索結果は自分のパソコンにダウンロードすることができ、エクセル等に取り込んで分析することができる。図 2 はエクセルのピボットテーブルの機能を使って、「玉葱」の表記の分布とレジスターの関係を示したものである。国会会議録だけが一つの表記で統一されているが、そのほかのレジスターでは、複数の表記が混在していることが分かる。

2,087 件の検索結果が見つかりました。そのうち 500 件を表示しています。
検索対象語数：124,100,964　記号・補助記号・空白を除いた検索対象語数：104,911,460

サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	レジスター	執筆者
LBcn_00027	ちょうど　いいでしょ。#　まだ　すこし、	たまねぎ	が　はいってるみたいだから、下の	タマネギ	玉葱	図書館・書籍	間所　ひさこ(著)
LBd3_00056	たのはまずジャガイモさんとニンジンさんと	タマネギ	さん。# それからチューインガムにソフトクリーム。# みんな	タマネギ	玉葱	図書館・書籍	松本　美津枝(著)
LBe9_00060	ながら、遅い昼食にニシンとサワークリーム、	玉ねぎ	、それに黒パンの簡単な食事を	タマネギ	玉葱	図書館・書籍	リチャード・ヘンリック(著)/小関　哲哉(訳)
LBf2_00018	一人が大鉢にワインをなみなみと入れ、	玉葱	のみじん切りにペッパー、油、香りの	タマネギ	玉葱	図書館・書籍	中島　暢太郎(著)

図1　中納言での検索結果（「玉葱」の例）

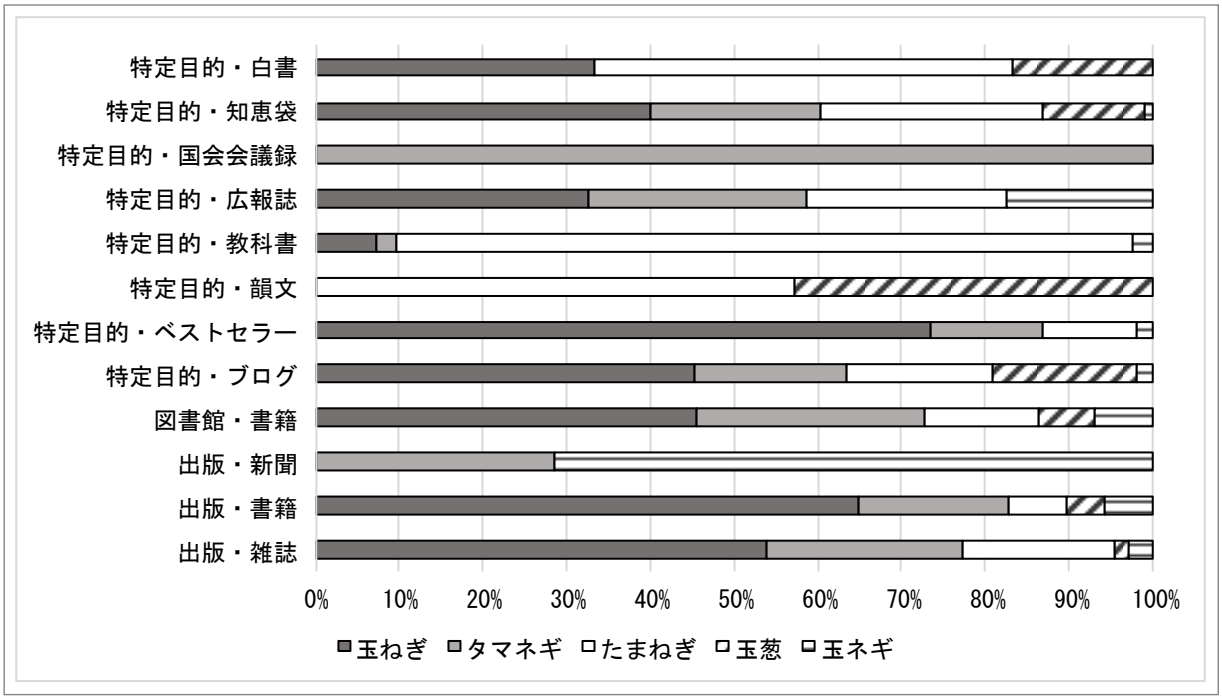

図2　「玉葱」の表記の分布とレジスターとの関係

図3はNBLで動詞「終わる」を検索した例である。ここでは、「名詞＋終わる」のパターンを表示させている。その中で「…を終わる」に着目して、どういうコロケーションがあるか、その頻度と共起の強さを測る尺度とともに示されている。また、一番右側のボックスには実際の例が示される。「…を終わる」は、「質問を終わる」が一番多く、その使用例は国会会議録に多いことが分かる。ワードプロファイラーは簡単にさまざまな用例を確認することができ、教材の作成にも役立つものと思われる。

図3　NLB の画面「終わる」の例

なお、上で紹介したコーパスやツールの多くは国立国語研究所のコーパス開発センターのサイトから情報を得ることができる。」

図4　コーパス開発センターのサイト
https://pj.ninjal.ac.jp/corpus_center/

7　コーパスの利用がもたらした変化

コーパスが言語研究にもたらした（あるいはもたらすであろう）変化は、定性的研究と定量的研究の関係の再構築という観点から捉えることができる。両者は、対立的な関係と思われがちであるが、むしろ、相補的な関係と捉えたほうがよいだろう。Leech（1992）でも、コ

ーパス言語学の特徴として、「質的な（qualitative）言語モデルのみならず、数量的な（quantitative）言語モデルにも中心を置く。」という点を挙げている。ここで、「質的な」と「数量的な」の両立を強調しているところが重要である。

定量的な研究の特徴として、網羅的記述が挙げられる（山崎 2016: 14）。コーパスから抽出されたデータを対象として分析を進めると、現象の全体像が把握できるのであるが、その中には例外があったり、自身の解釈にとって都合の悪い例があったりすることがある。そのような例を安易に切り捨てず、全体の中にうまく位置付けるような解釈を導くことができれば、研究をさらに精緻化することができる。内省だけを使った研究でも同じようなことができないわけではないだろうが、コーパスと同程度の網羅性を内省で実現するのはかなり難しいであろう。データが瞬時に集めることができるという点がコーパスのメリットである。

もう 1 つ重要な変化は、非母語話者への貢献である。データの検索に関しては、母語話者も非母語話者も同じであるから、非母語話者の参入がしやすくなったと言えよう。

8　コーパス言語学の特徴

8.1　コーパス言語学は普通の言語学である

Leech（1992）では、コーパス言語学の特徴として以下の 4 つを挙げている[5]。これらの特徴は生成文法に対するアンチテーゼと考えられよう。

⑴ 言語能力（competence）よりも言語運用能力（performance）を中心に置く。
⑵ 言語の普遍的特性（linguistic universals）の解明よりも個別言語の言語記述（linguistic description）に中心を置く。
⑶ 質的な（qualitative）言語モデルのみならず，数量的な（quantitative）言語モデルにも中心を置く。
⑷ 言語研究における合理主義的（rationalistic）な立場よりもむしろ，経験主義的な（empirical）立場に中心を置く。

宮島（2007: 41）は、「大量の用例をしらべることは，まさに言語学の王道」と述べ、コーパス言語学とは言語学そのものであると位置付けている。いわゆるコーパス言語学が言語学に中でどのような位置付けにあるかについては、2 つの立場がある。1 つは、上の宮島の主張ともつながるものである。例えば、園田（2002: 275）にコーパス言語学について以下のような指摘がある。

> ここで注意しなければならないのは、コーパス言語学だけを研究する純粋なコーパス言語学者は存在しないということです。言語研究者は何か自然言語についての問題を追及するためにコーパスを利用しますが、問題そのものはコーパス言語学のなかには存在しえないのです。

すなわち、コーパス言語学というのは、単に研究手段について言っているにすぎず、言語学的課題そのものは、言わばコーパスの外にあり、それは、文法であったり、語彙であったり、従来的な研究分野に属しているのだということを述べている。一方、前川（2007: 20）では、コーパスによる研究には corpus-based と corpus-driven の 2 種類があることを先行研究から引用している。

コーパス言語学の可能性を論じて Tognini-Bonelli (2001) は corpus-based investigation と corpus driven investigation の区別を主張している。前者は従来から言語研究において検討されてきた諸問題をコーパスを利用して解決しようとする研究である。一方後者は，コーパスそのもののなかから従来の言語研究では認識されてこなかった現象を発見し，それを解決しようとする研究である。前者にとってコーパスは研究ツールであるが，後者にとってのコーパスは研究対象そのものである。

その上で前川（2007: 22）は、corpus-driven な研究の可能性について文法性判断を例として以下のような考えを述べている。

従来の文法研究，とくに生成文法では文と非文との境界は明確に（二値的に）定まるものと考えてきた。しかし文法性判断に異動が存在する状態が稀な例外ではないとすれば，文と非文との関係を連続的な変化としてとらえることが考えられる。その場合，文法には正解が存在しないこととなり，文の候補として与えられた文字列の程度を評価することが新しい文法の主要な目的となるだろう。Corpus-driven な言語学がめざすべき目標には，このような文法性の程度を評価する連続量の計算法と，その評価値の高低が何に起因するかを説明するための理論が含まれていてしかるべきである。

9 研究事例

ここでは、森篤嗣（2014）を採り上げて、コーパス言語学の研究方法の特徴について考えてみたい。この論文の要旨には、以下のように書かれている。やや長くなるが、引用する。

本稿では，ナガラ節の二つの意味解釈「付帯状況」と「逆接」の意味判別を例として BCCWJ から抽出したナガラ節 13,830 件を人手で意味解釈したデータを用いて，文法研究によって蓄積された規則がどの程度，言語使用の実態を把握に資するかを計量化する。その結果，ナガラの前接品詞が名詞・ナ形容詞・イ形容詞の場合はほぼ逆接になるため，前接品詞が動詞で 10.02%，助動詞で 49.53%を占める逆接の 1,631 件をどのように意味判別するかが課題であることがわかった。ある規則（文法記述）が，この 1,631 件のうちどれくらいの意味判別に資するかを「寄与率」，当てはまる用例すべてのうち，意味判別に成功した割合を「判別精度」として分析した。現時点で検証できる判別精度 99%以上の規則での累積寄与率は 83.86%で，これをさらに上げて行くには，形態素解析辞書に動詞の意志性やアスペクト的特徴などを付与して行く必要があることもわかった。
（以上、森 2014: 84　より）

要旨には「寄与率」や「判別精度」など計量的な分析になじむ用語が使われ、従来の質的な研究とは異なる方法論であることが示されている。判別精度の例として表 1 を引用する。これは、接続助詞「ながら」に前節する動詞によって、その用法が付帯状況なのか、逆接なのかが判定できる度合いを示したものである。表のトップにある「知る」は、逆接でしか使用されないため、判別精度が 100%となっている。このような計量的事実は、コーパスを調べて初めて分かることである。この研究の行き先は、どのような条件があれば、これら 2 つ

の用法が分けられるのかを客観的に明らかにすることであり、その目的自体は定性的研究と変わるものではない。

表 1　判別精度が高い前接動詞（動詞：合計頻度 10 以上）（森 2014: 94）

	付帯状況	逆接	計	寄与率	判別精度	ナガラモ
知る	0	49	49	3.00%	100.00%	4
去る	1	11	12	0.67%	91.67%	0
認める	1	10	11	0.61%	90.91%	7
恐れる	2	9	11	0.55%	81.82%	0
会う	3	7	10	0.43%	70.00%	3
戸惑う	6	11	17	0.67%	64.71%	10
持つ	37	52	89	3.19%	58.43%	20
置く	23	27	50	1.66%	54.00%	2
伴う	6	4	10	0.25%	40.00%	4
装う	6	4	10	0.25%	40.00%	3

この論文の意義として、文法規則を計量的に評価することに成功していることが最も大きな点であろう。従来の文法規則はそれがどのくらいの事例に適用できるか、客観的評価が示されていなかったが、コーパスを使うことで、精度が示された。このように、指標を用いて客観的に文法規則の優劣を判断する手法は、今後の文法研究の中に適切に位置付けるべきものであろう。そのためには、そのためにはコーパスによる網羅的な把握が必要となってくる。

10　まとめ

7 節でコーパスの登場で定性的研究と定量的研究の関係が変化したと述べたが、山崎（2013: 2）では、以下のような表を挙げて、日本において、コーパス以前の文法研究からコーパス以後の文法研究で変化したのは、定性的研究から定量的研究へという研究手法ではなく、多くの場合は自説をサポートする用例の検索のためにコーパスが利用されていると述べている。すなわち、少なくとも人文系の日本語研究においては、表 2 の（1）から（3）への変化が起きていると考えられる。これは一面では無理からぬことで、現在の日本語研究を担っている世代は、統計やプログラミングの教育を受けている人が少ないことの反映である。しかし今後世代交代が進み、統計的な知識や手法の浸透により、人文系の日本語研究においても定量的かつ定性的なバランスのとれた言語研究が現れるようになるであろう。

表 2　研究手法とデータとの関係（山崎 2013: 2）

	定性的研究	定量的研究
作例（内省）	(1)	(2)
実例（コーパス）	(3)	(4)

付記

本稿は、2018 年 8 月 4 日、2018 年日本語教育国際研究大会（ICJLE2018）おいて行った、国立国語研究所の連続講義 3、4 にもとづくものである。

注.

[1] 言語資源という用語は、このほかにやや異なる意味でも用いられる。例えば、金水（2011:95）には、「言語資源論は、言語や言語変種（標準語・方言、文体変種など）、およびその構成要素（音韻、語彙、文法、文字・表記、語用論的規則など）を資源と捉え、人がそれを入手するための支出や労力（コスト）と、入手により得られる利益（ベネフィット）という二つの面から、人々の言語行動、言語の歴史的変化、言語の教育・学習等について考えるものである。」という考え方が示されている。

[2] http://www.elra.info/en/about/what-language-resource/

[3] モニターコーパスとは、時間が経つにつれ、データを追加（あるいは削除）し、随時変化しつづけるコーパスのことである。

[4] http://lsaj.ninjal.ac.jp/

[5] 邦訳は、齋藤俊雄他「英語コーパス言語学」（1998: 4）より。

＜参考文献＞

Leech, G. (1992) Corpora and theories of linguistic performance. *In J. Svartvik (Ed.) Trends in linguistics: Directions in corpus linguistics. Proceedings of Nobel Symposium 82 Stockholm* (pp.105-122). Berlin/New York: Mouton de Guyter.（邦訳 : 齋藤俊雄・赤野一郎・中村純作『英語コーパス言語学』（1998）研究社）

金水敏（2011）「日本語の将来を考える視点―「言語資源論」の観点から―」,『学術の動向』16(5), pp.95-99.

前川喜久雄（2016）「仮想講義『言語資源学入門』」,『日本語学』35(13), pp.2-11.

水谷静夫（1953）「語彙調査大体」,『国語学』15, pp.58-69.

宮島達夫（2007）「語彙調査からコーパスへ」,『日本語科学』22, pp.29-46, 東京 : 国書刊行会.

森篤嗣(2014)「意味判別における文法記述効果の計量化―ナガラ節の意味判別を例として―」,『日本語文法』14(2), pp.84-100.

山崎誠（2013）「形式語研究の方法論―定性的研究と定量的研究―」, 藤田保幸編『形式語研究論集』, pp.1-18, 大阪 : 和泉書院.

山崎誠（2016）「コーパスが変える日本語の科学―日本語研究はどのように変わるか―」『日本語学』35(13), 12-17.

＜参考 URL＞

国立国語研究所コーパス開発センター

https://pj.ninjal.ac.jp/corpus_center/

『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2018 年版』

https://ninjal-usamilab.info/lab/btsj_corpus/