athshala पाठशाला MHRD Govt. of India

Development Team

Principal Investigator:	**Prof. Anita Khanna** Jawaharlal Nehru University, New Delhi
Paper Coordinator:	**Prof. Prashant Pardeshi** The National Institute for Japanese Language and Linguistics (NINJAL)
Content Writer:	**Prof. Prashant Pardeshi** The National Institute for Japanese Language and Linguistics (NINJAL)
Content Reviewer:	**Prof. Emerita Yuriko Sunakawa** University of Tsukuba

Japanese	**Japanese Linguistics**
	コーパスに基づく日本語研究 (Corpus Based Research on Japanese)

Description of Module	
Subject Name	**Japanese**
Paper Name	**日本語学 (Japanese Linguistics)**
Module Title	**コーパスに基づく日本語研究 (Corpus Based Research on Japanese)**
Module ID	**JPN-P02-M29**
Quadrant 1	**E-Text**

コーパスに基づく日本語の研究

目的：このモジュールの目的は，「**コーパス言語学**」という言語研究の分野について解説し，コーパス言語学的な研究の事例として日本語の語彙的自他動詞対に関する最新の研究成果を紹介することである。

1. コーパスおよびコーパス言語学とは

言語の研究を行う場合は，研究者の内省に基づいて研究を行う方法と実際に使用されている言語データに基づいて研究を行う方法がある。パーソナルコンピューターの時代が到来するまでは，言語の研究は専ら研究者の直感に基づく作例や古い文献，小説，新聞などの小規模な言語データに基づいて行われていた。パーソナルコンピューターの普及や自然言語処理技術の発達に伴い，大規模な電子化された言語データを集積することや使用することが可能となった。大規模な電子化された言語データの集積をコーパス（corpus）という。また，コーパスを利用した言語の使用実態に基づいて分析を行う研究分野をコーパス言語学という。

コーパスを利用して様々な言語現象の使用実態を記述したり，分析したりするためには，大規模な言語データの森から記述・分析に必要なデータをピンポイントで抽出する必要がある。そのため，コーパスの言語データには様々なレベルの文法情報（品詞情報，係受け情報，句構造，節・文構造など）が付与されている。

コーパスの言語データに文法情報を付与する作業のことを**アノテーション**（annotation）という。アノテショーンの質および量はコーパスから抽出できる言語データの精密さの決め手となる。例えば，受動文がどのようなジャンルでどのくらい使われているのか，「飛び出す」「話し込む」などの複合動詞にどのようなタイプがあるのか，「帽子をかぶった子ども」のように主語（子ども）を修飾する関係節と「子どもがかぶった帽子」のように目的語（帽子）を修飾する関係節のどちらが多く使われているのかなどをコーパスで調べたいと思ったときは，アノテーションがどのように行われているのかを十分に理解しておくことが重要である。コーパス言語学は，まさに言語学とコンピューター科学を融合した学際的な分野であり，近年脚光を浴びている。

先にも述べたように，コーパスは大規模な電子化された言語データの集積である。コーパスの言語データの規模は，一般的に，コーパスに収録されている総語数を指す。

IT 技術の進歩と共にコーパスの規模も拡大し，その規模は近年 100 億語にも及んでいる。

以下に，日本語の代表的なコーパスを簡単に紹介する。

2. 日本語のコーパス

日本語のコーパスで代表的なものは，国立国語研究所が構築した『**現代日本語書き言葉均衡コーパス**』(BCCWJ) である（http://pj.ninjal.ac.jp/corpus_center/bccwj/）。BCCWJ はその名の通り，現代日本語の書き言葉のデータを収録したものであり，その規模は 1 億 430 万語である。さらに，このコーパスは書き言葉の様々なジャンル（書籍，雑誌，新聞，白書，ブログなど）から無作為にサンプルを抽出し，現代語の全体像を把握できるように構築されているため，均衡コーパスと呼ばれる。BCCWJ を無料・無登録で検索できる検索システムは次の二つである。

(1) 少納言（http://www.kotonoha.gr.jp/shonagon/）

(2) NINJAL-LWP for BCCWJ (NLB)（http://nlb.ninjal.ac.jp/）

ユーザー登録をすれば以下のシステムを使うことができる。

(3) 中納言（https://chunagon.ninjal.ac.jp/）

また，日本語のウェブサイトから収集した 11 億 3800 万語のデータを格納した**筑波ウェブコーパス**も公開されている。このコーパスはウェブサイトというジャンルのみからデータを収集しているため BCCWJ と違って均衡コーパスではないが，その規模は BCCWJ の約 10 倍である。筑波ウェブコーパスは NLB と同じ検索システム（NINJAL-LWP）で次のサイトから無料で利用できる（http://nlt.tsukuba.lagoinst.info/）。

さらに，100 億語規模の**『国語研日本語ウェブコーパス』**（NWJC）が国立国語研究所によって構築され，検索システム「梵天」によって検索できる。「梵天」を利用するには利用申請が必要である。（http://pj.ninjal.ac.jp/corpus_center/nwjc/subscription.html）

上記のコーパスは主に形態論的な情報（品詞情報，係受け情報）を付与したコーパスである。一方，統語・意味解析情報（句構造，節・文構造など）を付与した『NINJAL Parsed Corpus of Modern Japanese (NPCMJ)』の構築が国立国語研究所によって開始され（http://npcmj.ninjal.ac.jp/），2017 年 4 月現在で約 1 万文（17 万語）規模のコーパスが公開されている。このコーパスを検索するために複数の検索ツール（インターフェース）が用意され，ユーザー登録せずに利用することが可能である（http://npcmj.ninjal.ac.jp/interfaces/）。その中で一番簡単に利用できるインターフェースは「パターンブラウザー」である。これを利用し，上述の「帽子をかぶった子ども」

のように主語（子ども）を修飾する関係節と「子どもがかぶった帽子」のように目的語（帽子）を修飾する関係節のどちらが多く使われているのかを簡単に調べることができる。これから 5 年間で毎年 10,000 文ずつ追加され，最終的には 60,000 文（100万語）規模のコーパスになる予定である。

そのほか，日本語の話し言葉コーパス，歴史コーパス，近代語のコーパスについての情報が，国立国語研究所の「コーパス・データベース」のサイトに掲載されている。（http://www.ninjal.ac.jp/database/）。

以上は日本語母語話者のコーパスであるが，日本語学習者の話し言葉や書き言葉を集積した日本語学習者コーパスもある。上記の国立国語研究所の「コーパス・データベース」のサイトから，以下の学習者コーパスについての情報を得ることができる。

(1) 『中国語・韓国語母語の日本語学習者縦断発話コーパス』 (C-JAS)

(2) 『多言語母語の日本語学習者横断コーパス』 (I-JAS)

(3) 『日本語学習者による日本語・母語対照データベース』

(4) 『寺村誤用例集データベース』

上記の (2) については利用申請が必要だが，(1)，(3)，(4) については無登録で利用できる。学習者コーパスは日本語学習者の習得研究に大きく貢献するものであるとと

もに，学習者の日本語使用を日本語母語話者の日本語使用と比較することにより，日本語の分析にも活用することができる。

3. コーパスに基づく研究の事例：日本語の自動詞と他動詞

日本語を含む世界の多くの言語には，「立つ (tat-u)：立てる (tate-ru)」，「裂ける (sake-ru)：裂く(sak-u)」，「移る(utur-u)：移す(utus-u)」，「死ぬ (sin-u)：殺す(koros-u)」のような語彙的自他動詞対が存在する。特に日本語には語彙的自他動詞対が豊富にある。日本語の自動詞と他動詞の間の形式的な関係は表 1 に示すように分類できる。この表に挙げたもののほかに，「死ぬ (sin-u)：殺す(koros-u)」のように形式的な関係がない対もある。

表1 自他動詞の形態的な関係による分類

形態的な関係	例	派生の方向の有無		サイズの大小
他動化型	開く → 開ける 立つ → 立てる	方向あり	自動詞(無標) → 他動詞(有標)	自動詞 < 他動詞
自動化型	焼ける ← 焼く 裂ける ← 裂く	方向あり	自動詞(有標) ← 他動詞(無標)	自動詞 > 他動詞
均衡型	直る : 直す 移る : 移す	方向なし	自動詞 : 他動詞	自動詞 = 他動詞
自他同形型	開く = 開く 終わる = 終わる	方向なし	自動詞 = 他動詞	自動詞 = 他動詞

自然言語の様々なレベルの言語単位（音，語，句，文など）には形態上の非対称性がある。一方は形式的に単純で，そこから派生されたもう一方は，形式的に複雑になる。例えば，能動形の「建てる」と受動形の「建てられる」を比べると，能動形には何の標識も付いていないが，受動形には受動の標識である「られる」が付いている。標識論（markedness theory）では，特定の標識を用いて示される形態を有標（marked），標識を用いないで示される形態を無標（unmarked）と呼んでいる。上の表 1 では，他動化型においては，自動詞のほうが他動詞より短い（自動詞＜他動詞）ため自動詞は無標，他動詞は有標となる。一方，自動化型においては，他動詞のほうが自動詞より短い（自動詞＞他動詞）ため他動詞は無標，自動詞は有標となる。つまり，方向性のある自他対において，自動詞あるいは他動詞は無標（単純・短い）になる場合もあれば，有標（複雑・長い）になる場合もある。このような形式的な無標と有標の違いはどのような理由（動機付け）で決まるだろうか。言語類型論の分野では，このような疑問を解明するために多くの言語データに基づいた研究が行われ，以下の 2 つの動機づけが提案されている。

(1) 意味的な動機づけ：形式的な単純さ・複雑さ（すなわち形式的な短さ・長さ）は意味的な単純さ・複雑さを反映しているものである。認知，意味の面において

無標・単純な出来事のほうが，形態の面においても無標・単純である（すなわち短い）。逆に，認知，意味の面において有標・複雑な出来事のほうが，形態の面においても有標・複雑である（すなわち長い）。この立場は意味と形態の間の類似性（iconicity）を主張する。

(2) 経済的な動機づけ：使用頻度が高いものは形態上コンパクトに短く表現され，使用頻度の低いものは形態上長くなる傾向がある。この立場は形態の長さが使用頻度によって決まるという考え方である。

このモジュールでは，コーパスに基づく研究の事例として (2) の経済的な動機づけを主張するナロック・パルデシ・赤瀬川 (2015) の統語的派生に関する研究を紹介する。

4. コーパスに基づく研究の事例：ナロック・パルデシ・赤瀬川 (2015)

表 1 に示した形式的な関係による分類の中で，対をなす動詞の間にサイズの違いがあるのは他動化型と自動化型の 2 つのみである。他動化型，例えば「開く→開ける」の場合は，自動詞「（ドアが）開く」では動作主がなく，それに，動作主を加えて，他動詞「（だれかがドアを）開ける」を派生している。一方，自動化型，例えば「裂く→裂ける」の場合は，他動詞のほうにある動作主「（だれかがなにかを）裂く」を省い

て自動詞「（なにかが）裂ける」が派生されている。ナロック・パルデシ・赤瀬川はこのような派生を統語論的派生と名付け，次の仮説を立てている。

仮説：どのような統語論的派生パターンにおいても，形式的に派生された動詞のほうが頻度が低く，形式的な派生元となる動詞のほうが頻度が高い。

この仮説を検証するために，現代日本語の語彙的自他動詞対の包括的なリスト「現代語自他対一覧表」を作成し（この表は http://watp.ninjal.ac.jp/resources/からダウンロードできる），そのリストの他動化型，自動化型に属する動詞について，BCCWJを用いた頻度調査を行った。その結果を示したのが表 2 である。

表2　他動詞型と自動詞型の頻度と%

統語論的派生パターン	派生元の動詞（形態的に短い）		派生された動詞（形態的に長い）	
	頻度	%	頻度	%
他動化型	182	82%	38	17%
自動化型	103	74%	36	26%
合計	285	79%	74	21%

表 2 に見られるように，派生元の動詞は，他動化型が 82%，自動化型が 74%と，どちらも派生された動詞よりも頻度が高い。すなわち，他動化型も自動化型も形式的に短い動詞のほうが形式的に長い動詞よりも高い頻度で使用されていることが分かる。

さらに，その比率は自動化型より他動化型のほうが高いことから，他動化型の方が仮説

に合致する傾向が強いと言える。これは，他動化型の派生元の動詞，すなわち自動詞の頻度が高いことを示すものである。このことは，日本語の言語使用において，自動的表現法のほうが好まれることを示すものであると考えられる。また，形態の複雑さの度合いに差がない「直す：直る」のような対の調査においては，自動詞の頻度のほうが高いという結果が得られている（74% 対 26%）。この結果も自動詞がベースとなっている日本語の特質によるものであると考えられる。

以上のような議論を展開し，ナロック・パルデシ・赤瀬川 (2015) は，現代日本語の大量のコーパスデータを用いて，形式的に短い動詞のほうが頻度が高く，形式的に長い動詞のほうが頻度が低いという仮説を概ね裏付けている。

このように，コーパスを利用すれば，使用頻度という客観的な指標を用いて様々な言語現象を説明することが可能となる。

キーワード：

コーパス言語学　アノテーション　BCCWJ コーパス　筑波ウェブコーパス

NPCMJ コーパス　NLB　NLT

****