コーパスのオンライン利用と言語研究の可能性

佐野　真一郎

1. はじめに

近年の技術発展に伴い，様々な研究資源の整備・拡充が進んでいる。コーパスも例に漏れず，これを用いた研究も増加の一途を辿っている。2010 年代以降は，コーパス利用のオンライン化も進められ，利便性も更に高まっている。例えば，現在様々なコーパスが利用可能であるが，共通のオンライン検索システムにより，複数のコーパス間比較を比較的容易に行うことができる。また，異なる複数のコーパスを同一のプラットフォームで検索することで，研究対象について同一基準で多角的に分析・検証ができる。本稿では，日本語・英語それぞれについてオンライン利用が可能な代表的コーパスを用いた研究事例を報告し，言語研究の新たな可能性を提案したい。まず日本語について，国立国語研究所による『中納言』から利用できるいくつかのコーパスを用いた「ら抜き言葉」の数量的研究を紹介する。英語については，English-Corpora.org からオンライン利用のできるコーパスを用いて，単語使用の通時的変遷に関する分析を紹介する。

2. 「中納言」を用いたら抜き言葉の分析

ら抜き言葉は，一段動詞・カ変動詞の可能形に現れる進行中の言語変化であり，規範形「見られる」に対するら抜き形「見れる」などの変異として観察される（金田一ら 1995）。ら抜き言葉の分布は，話者の属性や発話場面などの影響を受ける（Sano 2011 他）。本研究では，レジスター，スタイルなどに関して，性質の異なる複数のコーパス間比較を通して，ら抜き言葉の性質を推測する。データ検索には「中納言」を用いた。「中納言」を介して検索できるコーパスは現在 10 種類あるが，本研究では表 1 に示す 5 つのコーパスを利用した。各コーパスが持つ特徴のうち，本研究と直接関係のあるものを表 1 にまとめる。これらのコーパスから抽出した可能形のデータ 10,365 件（規範形 9,328 件，ら抜き形 1,037 件）を対象として分析を行った。

表 1. 各コーパスの特徴

コーパス	話・書	データ収録時期	発話形式	スタイル
現代日本語書き言葉均衡コーパス	書き言葉	中間 (1976-2005)	−	−
昭和話し言葉コーパス	話し言葉	古い (1950s-1970s)	会話+独話	中間
日本語話し言葉コーパス	話し言葉	新しい (1999-2002)	独話 (+対話, 朗読)	あらたまった
名大会話コーパス	話し言葉	新しい (2001-2003)	会話	くだけた
日本語日常会話コーパス	話し言葉	最新 (2016-)	会話	くだけた

図 1 が示すように，ら抜き率の分布はコーパスごとに大きく偏っている。表 1 の各コーパスの特徴を可能形の使用文脈と捉えると，ら抜き言葉の性質に関する以下の一般化が導かれる：1）書き言葉よりも話し言葉に，2）古いデータよりも新しいデータに，3）独話よりも会話に，4）あらたまった発話よりもくだけた発話に現れやすい。

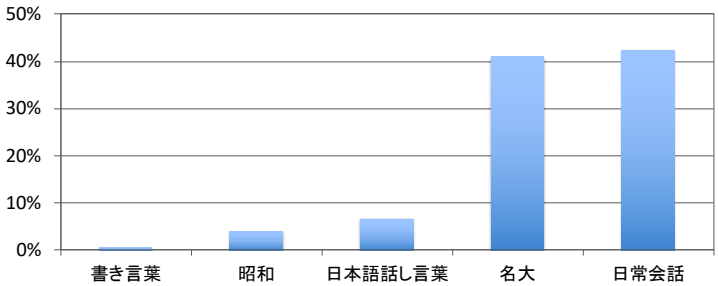

図 1. 各コーパスにおけるら抜き率の分布[注1]

以上から，進行中の言語変化であるら抜き言葉の性質が確認された。また，発話場面における相手とのやり取りの有無の影響や，書き言葉と同様にあらたまった発話が言語変化への耐性を持つことも示された。

3. English-Corpora.org を用いた単語使用の通時的分析

20 世紀以降，国際化が加速し，情報技術も大幅に発展した。これに伴い，新たな語彙が生み出され，その使用実態も変化してきている。本研究では，このような社会的変化の影響を受け，近年使用が拡大していると思われる英語の名詞 8 種を取り上げ，単語使用の通詞的変遷，意味拡張，スタイル差，ジャンル差について複数のコーパスを用いた数量的分析を行った。データ検索には，英語を対象とした，時代やジャンルの異なるコー

パスが（Virtual Corpora を介して）オンライン検索できる English-Corpora.org を用いた。

まず，1810 年代から 2000 年代までのアメリカ英語を収録している Corpus of Historical American English（COHA）を用いて単語使用の変遷を見る。図 2 は，各単語の頻度を年代ごとにまとめたものである。

	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
international	5.08	2.31	4.94	5.73	7.22	20.46	15.41	18.7	25.97	37.24	64.54	103.81	112.75	127.07	126.34	118.95	110.52	139.4	125.66	133.42
globalization	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.28	3.15	9.06
computer	0	0	0	0	0	0	0.16	0	0	0.09	0.04	0.04	0.16	0.29	3.95	18.81	50.56	112.93	140.51	108.77
internet	0	0	0	0.12	0.36	0	0.11	0.1	0.24	0.09	0	0.19	0.04	0	0	0	0	0	45.99	84.59
online	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0.04	0	0.04	0.28	24.12	63.55
blog	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.29	5.85
mouse	8.47	1.73	2.98	2.93	8.44	8.85	8.84	7.63	8.79	9.01	15.07	16.37	8.33	13.1	8.76	7.8	9.62	11.06	19.08	20.33
web	1.69	5.2	6.24	6.73	4.74	9.15	7.87	6.01	5	5.7	5.15	4.79	4.92	4.81	5.99	5.67	6.3	9.64	42.95	82.56

図 2. COHA における語彙頻度の通時的変遷（数値は 100 万語あたりの調整頻度）

図 2 の分布特徴（網掛け部分）は以下のようにまとめられる：1）“international”は 1900 年代から増加している，2）“globalization”は 1980 年代から観察され，2000 年代に急増している，3）“computer”は 1970 年代から増加している，4）“internet,” [注2] “online,” “blog”は，1990，2000 年代から増加している，5）“mouse,” “web”は 1990 年代以降増加しているが，1800 年代から一定数が比較的安定して観察される。

ここで“mouse,” “web”が比較的古くから観察される点に注目したい。これらは，既存語彙の意味拡張によってコンピュータ用語としても使われるようになった。従って，前者は「ねずみ」「マウス」，後者は「蜘蛛の巣」「インターネット」といった多義性を示す（Blank 1999）。COHA で共起する単語を検索すると，“mouse”の場合“mickey,” “cat”などと共起する「ネズミ」の用例は 1800 年代から広く分布しているが，“click”などと共起する「マウス」の意味では 1990 年代以降に限られる。“web”の用例のうち“spider”, “spun”などと共起する「蜘蛛の巣」の意味では 1800 年代から多く見られるが，“site(s),” “page(s)”などと共起する「インターネット」の意味では 1990 年代以降に限られる。このように，近年の社会変化を反映した単語使用の変化と，単語内での用法の変化が確認された。

スタイル差については，“internet”と“web”を取り上げ，COHA と法律文書に特化した Corpus of US Supreme Court Opinions での分布を比較した。前者では差が無かったが，後者ではよりあらたまった場面で使われる internet がより多く観察され，スタイルの単語選択への影響が確認された。

ジャンル差については，“web”を対象に，インターネット上の記事を収録する Corpus of Online Registers of English を用いて検索したところ，情報提供・技術指南に関する記事でよく使われることが分かった。

4. 結論

本稿では，日英両語のオンライン検索システムを利用したコーパス間比較分析の事例を報告した。性質の異なる複数のコーパスを比較することで，分析対象の特徴を多角的・効率的に分析するなど，コーパスのオンライン利用は言語研究の可能性を拡大すると言える。一方で，このような方法は対象の全体像を把握するには便利だが，各コーパス内にも細かな違いがあるため，詳細な理解にはコーパスごと・用例ごとに吟味が必要である。また，コーパスごとに設計方針などが異なるため，それぞれの特徴を踏まえた研究計画が必須である。

[注1] ソーシャル・ネットワーキング・サービス，ブログなどは話し言葉に似た性質を備えているため，「現代日本語書き言葉均衡コーパス」におけるら抜き率は，「Yahoo!ブログ」，「Yahoo!知恵袋」の用例を除いたものである。また，同コーパスにおけるその他のら抜き形は，「書籍」の 1 例を除き全て「雑誌」で観察された。

[注2] COHA における“internet”の 1800 年代の用例は，“Internet Archive”のようなものばかりであった。これらは 1800 年代の資料を収録する（近年）編纂されたアーカイブの注釈内の用例である。従って，“internet”が 1800 年代に使用されていたということではない。

主要参考文献

Blank, A. (1999) Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Blank, A. and Koch, P. (eds.) Historical Semantics and Cognition, pp. 61-89. Berlin: Mouton de Gruyter.

金田一春彦・柴田武・林大（編）(1995)『日本語百科大事典 縮刷版』東京：大修館書店

Sano, Shin-ichiro. (2011) Real-time demonstration of the interaction among internal and external factors in language change: A corpus study. *Gengo kenkyuu* 139, 1-27.