現代のテキストコーパス

Contemporary Written Corpus of Japanese

山崎　誠

1. 書き言葉コーパスの展開

日本語研究におけるコーパスの利用は 1990 年代から主に言語処理の分野で始まり，21 世紀に入る頃からは人文系の日本語研究・日本語教育研究においても，次第に利用されるようになり，現在では現代語研究のみならず，日本語史研究においても盛んに利用されるようになってきた．研究の分野は，語彙，文法，文体，表記，音声，音韻，コミュニケーション，日本語教育，国語教育と多岐にわたる．日本語研究以外でも社会学，心理学などでもコーパスは利用されている．実用的な場面では，国語辞書編纂の基礎資料としても利用されている．

日本語研究でコーパスという名前が付くデータが登場したのは，「京都大学テキストコーパス」，略称「京大コーパス」が最初であろう．このコーパスは，「毎日新聞」の 1995 年の記事約 4 万文に対して，形態論情報(用語)，構文情報(用語)を付与したものである(1)．なお，公開されているのは，タグのみであり，コーパスを利用するためには原文の新聞記事データを購入しなければならない．京大コーパスは，専ら自然言語処理分野で利用されている．京大コーパスのデータが新聞記事であったことからも分かるように，1980 年代後半から各新聞社が自社の記事をテキストファイルとして有料で売り出すようになり，それがコーパスとして利用されるという研究手法が生まれた．ただし，新聞記事データはかなり高額であったため，研究費が潤沢でなく，個人研究の割合が高い人文系日本語研究においては，利用が盛んではなかった．その代わり，「新潮文庫の 100 冊」が手軽なコーパスの代用としてよく使われた．

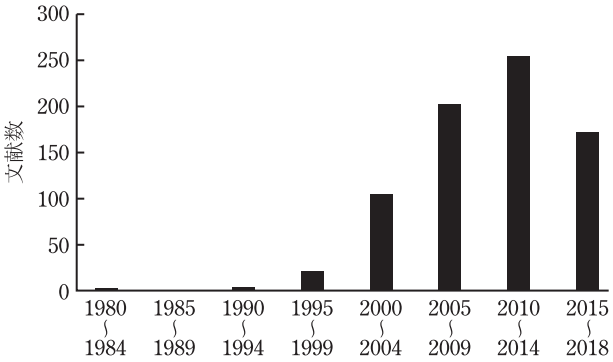


図 1　タイトルに「コーパス」を含む文献数の推移

図 1 は，国立国語研究所の「日本語学・日本語教育文献データベース」でタイトルに「コーパス」を含む文献数の推移を示したものであるが，これからも 2000 年代から文献数が急増していることが分かる(注1)．また，2015〜2018 年の「コーパス」の頻度は 172 件であるが，同様の検索における「文法」は 193 件．「アクセント」は 155 件であった．このことは，コーパスが日本語研究の中にしっかり根付いたことを表していると言えよう．

日本語研究におけるコーパスの利用は，2000 年代の後半，「現代日本語書き言葉均衡コーパス(用語)」(BCCWJ : Balanced Corpus of Contemporary Written Japanese(注2)) の登場によって一変する．これまでの書き言葉コーパスが，新聞あるいは小説という一つのジャンルから構成されていたのに対し，BCCWJ は，多種のジャンル(注3)を含む均衡コーパス（balanced corpus）であったため，研究における信頼性が高いと判断されたのである．2010 年代からは，日本語史研究のための「日

山崎　誠　国立国語研究所言語変化研究領域
E-mail　yamazaki@ninjal.ac.jp
Makoto YAMAZAKI, Nonmember (Language Change Division, National Institute for Japanese Language and Linguistics, Tachikawa-shi, 190-8561 Japan).
電子情報通信学会誌　Vol.102 No.6 pp.549-553 2019 年 6 月

(注 1)　2018 年 11 月 27 日更新版を利用．検索日は 2018 年 12 月 20 日．
(注 2)　現代日本語書き言葉均衡コーパス（BCCWJ），https://pj.ninjal.ac.jp/corpus_center/bccwj/
(注 3)　BCCWJ では，「レジスター」という用語を使っている．

本語歴史コーパス」(CHJ:Corpus of Historical Japanese)，Web上のテキストから成る「国語研日本語ウェブコーパス」(NWJC:NINJAL Web Japanese Corpus[注4])も加わり，コーパスの利用は今後ますます充実してくることが期待される．

2. 主な書き言葉コーパス

本章では，国立国語研究所で開発した現代語の書き言葉のコーパスを二つ紹介する．

2.1 現代日本語書き言葉均衡コーパス（BCCWJ）

人文系日本語研究においては最もよく利用されているコーパスである．国立国語研究所・コーパス開発センターのHPには，BCCWJをはじめ各コーパスを利用した文献の一覧が掲載されているが，BCCWJは，2018年7月31日現在848件の利用実績がある．

BCCWJは，1億語規模のコーパスで，書籍，雑誌，新聞，白書，広報誌，Web上のテキストなど13個のジャンルから構成されている（図2）．

BCCWJの特徴として，短単位と長単位という2種類の言語単位で解析されていることが挙げられる．言語の分析のためには，まずテキストを単語に分解する必要があるが，通常の日本語は分かち書きされていないため，いわゆる単語に当たるものを客観的な基準で設計しなければならない．そこで，意味を持つ最小の言語単位である形態素（morpheme）の結合の度合いによって，短単位と長単位という2種類の言語単位を設けた．例えば，「日本語書き言葉コーパス」という語は，短単位では，「日本／語／書き言葉／コーパス」と4語に分かれるが，長単位ではこれ全体で1語である．短単位は主に基本語彙の選定に，長単位はジャンルの比較などに使われる．

2.2 国語研日本語ウェブコーパス（NWJC）

NWJCは，Webを母集団とした，約258億語のコーパスである．NWCJの構築目的は，「稀言語現象の言語学的，心理学的及び情報処理的視点からの究明の可能性を開くこと」（HPから）である．稀言語現象の例として，単語と単語の組合せがある．BCCWJでは，助詞「まで」は163,960件，助詞「こそ」は，16,642件と高頻度であるが，「までこそ」という連接はBCCWJで1件しか検索されない[注5]．しかし，NWJCでは，「までこそ」は139件ヒットし，言語学的な分析が可能になる．NWJCは，「梵天（ぼんてん）」[注6]という検索ツールを使用して検索する．

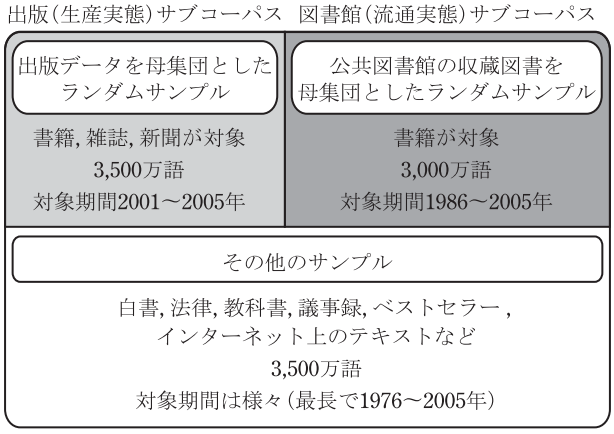


図2 BCCWJの構成

3. 検索ツールと関連技術

3.1 検索ツール「中納言」

コーパスの利用を促進したのが検索ツールの存在である．BCCWJなど多くのコーパスでは，「中納言」[注7]という検索ツールを利用している．中納言は，コーパスに付与された形態論情報を基に検索するツールで，BCCWJの検索のために作られたものである．現在では，日本語歴史コーパス，日本語話し言葉コーパス（CSJ），多言語母語話者の日本語学習者コーパス（I-JAS），名大会話コーパス，現日研・職場談話コーパス，日本語日常会話コーパス（CEJC）［モニター公開版］など，七つのコーパスが検索できる．将来的にはこれらのコーパスを1回の検索で串刺し検索できる機能を搭載する予定である．図3は，BCCWJで「国語」という語を検索した結果である．画面上は500件しか示されない

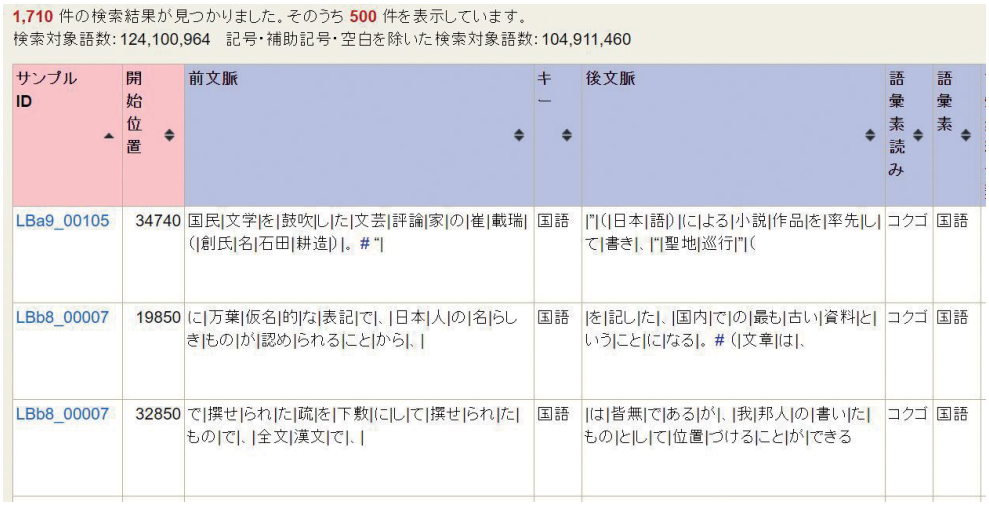
■ 用 語 解 説

形態論情報 テキストや談話に形態素解析を施し，言語単位に分割した際に，それぞれの言語単位に対して付与する，見出し語形，表記形，品詞など語に関する情報を指す．活用語であれば，活用に関する情報も含む．

構文情報 テキストの構文的理解に必要な情報．文節境界の情報やそれを基にした係り受けの情報，「ヲ格」，「ガ格」などの格関係の情報や指示・照応などの情報がある．

均衡コーパス 一般に言語現象はジャンルが変わると異なる様相を呈することがある．そこで，特定のジャンルに偏らず，様々なジャンルのテキストや談話を集めたコーパスが必要となる．そのようなコーパスは均衡コーパス（balanced corpus）と呼ばれる．

（注4） 国語研日本語ウェブコーパス（NWJC），https://pj.ninjal.ac.jp/corpus_center/nwjc/
（注5） 「M & A はいまでこそ」という文脈で，誤解析である．正解は「今でこそ」．（サンプルID:LBs2_00038，開始位置：15840）
（注6） 梵天，http://bonten.ninjal.ac.jp/string_search
（注7） 中納言，https://chunagon.ninjal.ac.jp/



1,710 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	開始位置	前文脈	キー	後文脈	語彙素読み	語彙素
LBa9_00105	34740	国民|文学|を|鼓吹|した|文芸|評論|家|の|崔|載瑞|(|創氏|名|石田|耕造|)|。#"|	国語	|"|(|日本|語|)|による|小説|作品|を|率先|し|て|書き|、|"|聖地|巡行|"|(	コクゴ	国語
LBb8_00007	19850	に|万葉|仮名|的|な|表記|で|、|日本|人|の|名|らし|き|もの|が|認め|られる|こと|から|、|	国語	|を|記し|た|、|国内|で|の|最も|古い|資料|と|いう|こと|に|なる|。#|(|文章|は|、	コクゴ	国語
LBb8_00007	32850	で|撰せ|られ|た|疏|を|下敷|に|し|て|撰せ|られ|た|もの|で|、|全文|漢文|で|、|	国語	|は|皆無|で|ある|が|、|我|邦人|の|書い|た|もの|と|し|て|位置|づける|こと|が|できる	コクゴ	国語

図 3　中納言で「国語」を検索した結果（一部）

が，検索結果がダウンロードできる(注8)ので，Excel 等で読み込ませたりすれば，細かな分析が可能になる．

3.2　形態素解析用電子化辞書「UniDic」

2.1 で分かち書きをしない日本語の分析にはまずテキストを単語に分割する必要があると述べたが，そのために必要なのが形態素解析(注9)という技術である．形態素解析のためには，解析を行うツールである解析器とそれに利用する辞書が必要である．解析器は JUMAN(注10)，ChaSen(注11)，MeCab(注12)などがよく利用されている．解析に使用する辞書は，IPADic が一般的であったが，言語学的には言語単位としての長さが不統一であるという問題があった．そのため，BCCWJ を構築するにあたって，新たな形態素解析用辞書 UniDic(注13)を構築した．UniDic は，短単位辞書であり，テキストを短単位に分割するために利用するものである．現代書き言葉用，現代話し言葉用，古文用などテキストの種類に応じた複数のバージョンが公開されている．

4．コーパスを利用した研究

4.1　コーパスにより変わる研究方法

文献(2)では，日本語研究においてコーパスの利用により，研究方法も変わると述べ，量的把握，網羅的記述，非母語話者への貢献，試行的研究のしやすさの 4 点を挙げている．以下この 4 点について簡単に説明する．

量的把握とは，印象でしか語れないことを具体的な数字で言えるようになることである．例えば，表記において「バイオリン」と「ヴァイオリン」，「賛美歌」と「讃美歌」のどちらが多いかは実際に調べてみないと分からない(注14)．コーパスの利用は，数字で分布や傾向を把握する客観的な研究姿勢がより普及するきっかけとなる．

表 1　「～の発生」における名詞（頻度順）

順位	名詞	頻度	順位	名詞	頻度
1	事故	212	11	がん	34
2	災害	99	12	損害	31
3	事件	86	13	事態	24
4	地震	66	14	赤潮	22
5	被害	62	14	雑草	22
6	火災	59	16	暴力	21
7	公害	55	16	汚染	21
8	障害	49	18	ガス	18
9	犯罪	45	19	死者	17
10	問題	43	20	悪臭	16

網羅的記述とは，データの取扱いの公平性と関係する．コーパスを使わず，頭で思い付く用例だけで研究を進めていくと，往々にして自分に都合の良い用例に偏りがちである．しかし，コーパスを使うと，自分の主張に合った都合の良い例だけでなく，都合の悪い例も出てくる可能性がある．それらを排除せず，むしろ研究の枠組みを広げる方向で再検討することにより，現象の本質に迫ることができるのではないか．

また，近年日本の大学院にはアジア圏を中心とする留

(注 8)　ただし，1 回の検索でダウンロードできるのは 10 万件までという上限がある．
(注 9)　形態素解析という用語は言語学的には正確でない．テキストを形態素に分割するわけではないからである．
(注 10)　JUMAN，http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
(注 11)　ChaSen（茶筌），http://chasen-legacy.osdn.jp/
(注 12)　MeCab（和布蕪），http://taku910.github.io/mecab/
(注 13)　UniDic，https://unidic.ninjal.ac.jp/
(注 14)　BCCWJ では，「ヴァイオリン」415 例，「バイオリン」293 例，「讃美歌」47 例，「賛美歌」89 例である．文部科学省の外来語の表記や常用漢字表における例外の方が多く使われていることが分かる．

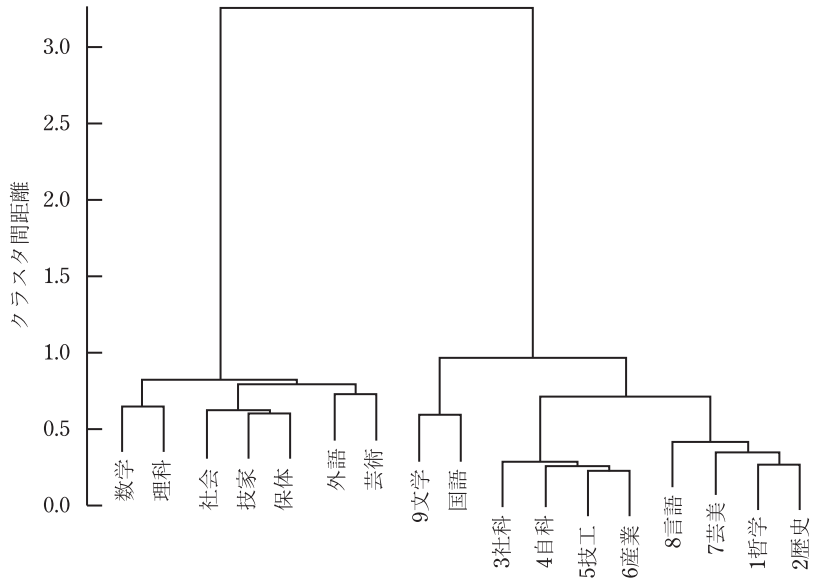


図 4　BCCWJ のジャンルのクラスタリング

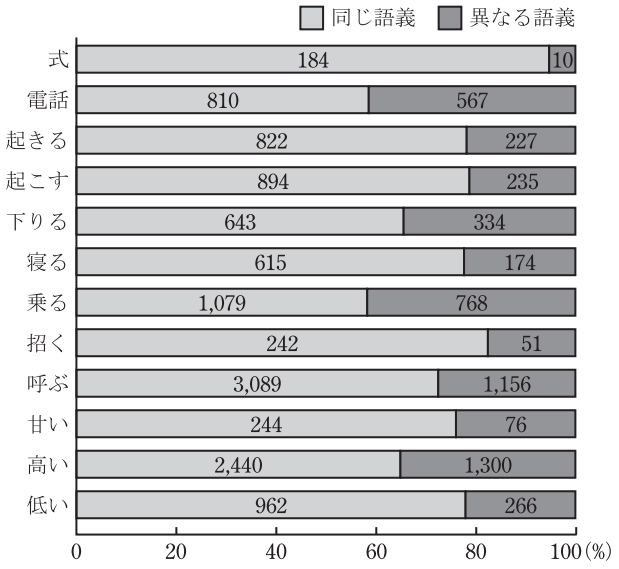


図 5　BCCWJ のサンプルにおける多義語の意味の分布

学生が多く在籍するようになった．彼ら・彼女らが日本語研究を行う際には，これまでは非母語話者であるという障壁があったが，コーパスを使うことにより，少なくともデータ収集に関しては母語話者と同じスタートラインに立つことができる．ただし，現代の日本語研究の中心的分野である文法研究においては，コーパスは用例収集を目的として使われている場合も目立ち，単純な傾向把握という分析も多く，本格的な定量的研究はそれほど広まってはいないという印象がある．

中納言のような検索ツールを使うと，検索結果がすぐに得られるため，いろいろな実験を試してみることができる．表 1 は，名詞＋「の発生」というパターンに入る名詞のリストである．これを見ると，このパターンに入る名詞にはマイナスの意味の語ばかりであることに気付く(注15)．このような例は国語辞書の記述に役立つばかりでなく，非母語話者の作文教育にも貢献する．

4.2　ジャンル・文体の分析

BCCWJ が複数のジャンルで構成される均衡コーパスであることから，ジャンルや文体による違いということが意識されることになった．図 4 は文献(3)による BCCWJ に収録されている書籍の図書分類と教科書の各教科をそれらに出現する上位 100 語の動詞でクラスタリングしたものである．内容が近いジャンルが近くに集まっていることが分かる．

4.3　従来の研究の見直し

コーパスを利用した研究で大きく期待されるのは，従来の研究の見直しや検証を通した研究の精緻化であろう．事例を二つ挙げる．

一つは文献(4)による二重目的語構文の基本語順の分析である．二重目的語構文とは 1 文中に「を」と「に」が同時に現れる構文で，「太郎が花子に本を貸す」「太郎が本を花子に貸す」のように二通りの語順がある．その語順について Web から取得した 100 億語規模のコーパスを用いて，従来指摘されている幾つかの仮説を検証したものである．その結果，「約 6 割の動詞は『にを』語順，4 割の動詞は『をに』語順が優勢である」，「省略されにくい格は動詞の近くに出現する傾向がある」などの

(注 15)　頻度は少ないが，「エネルギー」，「酸素」，「キノコ」のようなマイナスの意味を持たない語もある．

具体的な結果を得ている．この研究は直観あるいは少数のデータを元に主張されてきた現象をコーパスを用いて検証した好例である．

もう一つは，文献(5)で，文献(6)で提唱されている“one sense per discourse”（一つの談話で用いられる語義は一つだけ）という仮説を日本語で検証したものである．図5は，「式」，「起きる」，「甘い」などの多義語の意味が文章中でどのような分布を示すかを調べたものである．これによると，必ずしも多義語が文章中で一つの意味に限定されて用いられているわけではないが，一つの意味で用いられているサンプルがおおよそ60〜80%あることから，一つのテキストに一つの語義という傾向が緩やかに認められるとしている．

5. ま　と　め

本稿では日本語研究における書き言葉コーパスの展開とその利用について概観した．ここで触れることのできなかったコーパスや研究事例も多い．近年では，言語資源という概念の下に，コーパス，シソーラス，形態素解析技術などがまとめられ，総合的な研究領域として確立しつつある[7]．日本語研究におけるコーパスの利用はまだ始まったばかりである．今後，様々なコーパスが構築され，より一層の研究の進展が見込まれるだろう．

また，忘れてならないのは普及活動である．講習会の開催やコーパス研究のための授業用テキストの編さんなども合わせて取り組むべき課題である．

文　献

（1） 黒橋貞夫，長尾　眞，“京都大学テキストコーパス，”言語処理学会第3回年次大会発表論文集，pp. 115-118, March 1997.
（2） 山崎　誠，“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，vol. 35, no. 13, pp. 12-17, 2016.
（3） 内田　諭，藤井聖子，“クラスター分析とフレーム分析による語彙のジャンル別特徴―「現代日本語書き言葉均衡コーパス」を用いて―，”言語文化論究，no. 34, pp. 21-34, March 2015.
（4） 笹野遼平，奥村　学，“大規模コーパスに基づく日本語二重目的語構文の基本語順の分析，”自然言語処理，vol. 24, no. 5, pp. 687-703, Dec. 2017.
（5） 山崎　誠，“テキストにおける多義語の語義の分布―『現代日本語書き言葉均衡コーパス』を利用して―，”計量国語学，vol. 27, no. 7, pp. 251-262, Dec. 2014.
（6） W. Gale, K. Church, and D. Yarowsky, “One sense per discourse,” Proc. the 4th DARPA workshop on Speech and Natural Language, pp. 233-237, Harriman, NY, Feb. 1992.
（7） 前川喜久雄，“仮想講義「言語資源学入門」，”日本語学，vol. 35, no. 13, pp. 2-11, Dec. 2016.

（2019年1月1日受付　2019年1月18日最終受付）

山崎（やまざき）　誠（まこと）

昭59筑波大大学院文藝言語研究科退学，同年から国立国語研究所勤務．語彙調査，シソーラス，コーパスの構築に従事．現在，同所言語変化研究領域教授．博士（学術）．著書「テキストにおける語彙的結束性の計量的研究」など．