

# 国立国語研究所学術情報リポジトリ

＜講演＞ウチから見た日本語の多様性：  
言語研究のインフラ整備：  
日本語コーパスからみえてきたもの

メタデータ	言語: Japanese 出版者: 公開日: 2017-03-15 キーワード (Ja): キーワード (En): 作成者: 前川, 喜久雄 メールアドレス: 所属:
URL	<a href="https://doi.org/10.15084/00000950">https://doi.org/10.15084/00000950</a>

## 〈ウチから見た日本語の多様性〉

## 言語研究のインフラ整備

## ～日本語コーパスからみえてきたもの～

言語資源研究系教授

前川 喜久雄

私は、国立国語研究所が開発している言語資源、コーパスについてお話しします。まずコーパスの必要性について触れ、次にこれまでのコーパス整備の経緯とこれからの計画を紹介します。その後、近年に開発したいくつかのコーパスを紹介し、最後にコーパスを使うとどのような検索ができるか検索例をお示しします。

## なぜコーパスが必要か

コーパスとは、言語を研究するために大量の用例を組織的に収集して、コンピュータで効率的に検索できるようにしたデータのことです。それでは、なぜコーパスが必要か。それは言語には大きな多様性があるからです。多様性が大きすぎるので、単に頭で考えているだけでは、言語の実態を把握することができないのです(図1)。

世界に何千種類も言語があるという意味でも言語は多様なのですが、一つの言語、例えば日本語の内部にもさまざまな多様性が認められます。そのような言語内多様性の源はさまざまですが、よく知られているのは歴史的な多様性と地理的な多様性でしょう。言語は時間とともに

に変化します。それが歴史的な多様性を生み出します。変化のなかには日本語独自の变化も、外国語からの影響による変化もあります。古くは中国語、近年では英語が代表的な外国語です。地理的な多様性、つまり方言の問題については、さきほど木部先生のお話のテーマでもあったので、省略します。

言語には創造的使用と呼ばれる変化も生じます。言語、表現には一定の意味があるわけですが、それをあえて変化させて使うのが創造的使用です。最近「やばい」という言葉がポジティブな評価に使われるようになってきているのはその一例です。

その他、近年目立つようになった多様性の要因として、日本語を母語とし



前川 喜久雄(まえかわ きくお)

言語資源研究系教授。博士(学術)(東京工業大学)。音声学が専門ですが、自発音声の研究のために『日本語話し言葉コーパス』(CSJ)の開発に携わったことがきっかけとなって、1999年来コーパスの設計と実装に深く関係するようになりました。そのため最近では第二の専門として言語資源学をなっています。主な著書は『講座日本語コーパス』(朝倉書店)、A Frequency Dictionary of Japanese(Routledge)、『音声は何を伝えているか』(コロナ社)など。

ない人たちが使う日本語があります。学習者の日本語です。さらに、コンピュータがつくりだす日本語（機械翻訳）も、今後、多様性の源になるかもしれません。

また、何が原因かはよくわからないけれども、ある語にふたつ以上の語形があつて、明瞭な規則性もなしに使われていることがあります。「日本」がニホンかニッポンか、「矢張り」が「ヤハリ」か「ヤツバリ」か「ヤハシ」か、等々。これらは一種の確率的な変動であつて、言語の変異と呼ばれることがあります。

さて、このような言語内多様性を正確に把握したいのですが、はどうやって把握するか。思いつくままに、こんながある、あんなのもあるといつても、正確ではありません。客観的な方法で調べたデータが必要になります。その際、データに求められる特性としては、以

- ・多様性の源
  - －歴史的多様性
    - ・内発的变化
    - ・外国語の影響
  - －地理的多様性
  - －創造的使用
  - －確率的変動（言語変異）
  - －その他
    - ・非母語話者
    - ・機械翻訳

図1 言語内多様性

- ・直観だけでは把握できない（例は後で）
- ・客観的なデータが必要
  - －本当に使われたことのある用例のデータ
  - －対象を偏りなく代表するデータ
  - －できるだけ大量のデータ
  - －検索用の情報がついたデータ
  - －コンピュータで利用できる形式（機械可読形式）のデータ
  - －誰でも利用可能な公開されたデータ



言語コーパス (corpus) の整備  
～言語資源 (language resources) の整備

図2 言語内多様性を把握する手段

下のものがあります（図2）。

まず、頭でつくりだしたものではなく、実際に使われたことが分かつていること（実用例であること）。第二に、対象となる言語の一部分だけではなく、全体を偏りなく代表するデータになっていること（均衡性）。第三に、できるだけ大量のデータであること（大規模性）。第四に、検索用のいろいろな情報が付加されていて、コンピュータで検索できること。最後に、データをつくった人だけが利用したり、ある特殊な機関に所属している人間だけが使えるのではなく、誰でもが利用できる公開されたデータであること。

そのような条件を備えたデータのことを、われわれはコーパス (corpus) と呼ぶのです。また、そのコーパスを構築・利用するためのノウハウや検索ツール、さらにはコーパスから二次的に派生された種々の二次的データ（例えば辞書）などもふくめて、言語資源 (language resources) と呼ぶことがあります。

### 国語研によるコーパス開発の経緯

これまでの国語研究所によるコーパス開発の経緯をまとめてみます。国語研究所は一九四八年に創立されました。直後の一九五〇年代から、新聞、雑誌などを対象とした各種の「語彙調査」が実施されています。これは簡単にいえば、共通語の語彙を確定するための基礎調査でした。方法論的には優れたことをやっていたのですが、残念ながらデータを公開

しませんでした。国語研の研究者が使って、結果を報告書にまとめて、それでおしまいでした。その意味でコーパスとはいえません。

国語研がコーパスを開発しはじめたのははるかに遅く、一九九〇年代末からでした。それから現在までに構築してきた代表的な日本語コーパスを図3に示します。

最初に公開したのは『日本語話し言葉コーパス(CSJ)』(構築一九九〇～二〇〇三年度、公開二〇〇四年)でした。これは現代語の話し言葉を対象としたコーパスです。次は、明治から昭和初期にかけての書き言葉を対象とした『太陽コーパス』(構築一九九五～二〇〇四年度、公開二〇〇五年)、『太陽』と

いうのは当時広く読まれた総合雑誌の名前です。三番目の『現代日本語書き言葉均衡コーパス(BCCWJ)』(構築二〇〇六～二〇一〇年度、公開二〇一一年)は、現代語の書き言葉を、書籍・雑誌・新聞・白書・広報紙・ネット掲示板・ブログ・詩歌・法律など幅広く収集したもので、現在もとても活発に利用されている書き言葉のコーパスです。規模はちょうど一億語です。

『日本語歴史コーパス(CHJ)』(構築二〇一〇～)は、奈良時代までさかのぼることのできる過去の日本語を対象としたコーパスで、現在も構築中ですが、一部は公開されており、日本語史の研究者にとっては必須のコーパスになっています。

- ・ 1950年代から各種「語彙調査」を実施してきたがデータは公開しなかった
  - ・ 1990年代末にコーパス開発始動
    - －『日本語話し言葉コーパス(CSJ)』(構築 1999～2003、公開 2004)
    - －『太陽コーパス』(構築 1995～2005、公開 2005)
    - －『現代日本語書き言葉均衡コーパス(BCCWJ)』(構築 2006～2011、公開 2011)
    - －『日本語歴史コーパス(CHJ)』(構築 2010～、段階的に公開)
    - －『国語研日本語ウェブコーパス(NWJC)』(構築 2011～2015、公開 2016 予定)
    - －『多言語母語の日本語学習者横断コーパス(I-JAS)』(構築 2012～、部分試験公開 2016)
- 迫田の発表

図3 国語研によるコーパス開発の経緯

『国語研日本語ウェブコーパス(NWJC)』(構築二〇一〇～二〇一五年)はインターネット上の日本語を大量に収集したもので、規模は二百五十億語あります。来年度(二〇一六年度)に公開の予定です。さらに、きょうのちほど迫田先生のお話しに出てくる『多言語母語の日本語学習者横断コーパス(I-JAS)』もあります。これは、日本語を勉強している人たちの言語行動を記録したコーパスで、近日公開予定です。

ここでもう一度、一九九〇年代末にもどりましょう。その時期に、日本語学の研究者が利用することのできた日本語のデータには図4に示すものがありました。毎日新聞などの新聞社が有償で公開するテキストデータがありました。もう少し古い時代のデータとしては、著作権の切れた文芸作品をもとにした青空文庫が使えました。新潮社が過去の文芸作品をデジタル化した『新潮文庫の百冊』もしばしば利用されましたが、著作権の問題が解消されていたかどうかは不明です。これがすべてです。日本語の全体像を知るには、明らかに偏ったデータです。

そこから二〇年ほど頑張ってきて現在の整備状況を示したのが図5です。さきに説明したように、書き言葉に関しては『現代日本語書き言葉均衡コーパス』があり、話し言葉については『日本語話し言葉コー

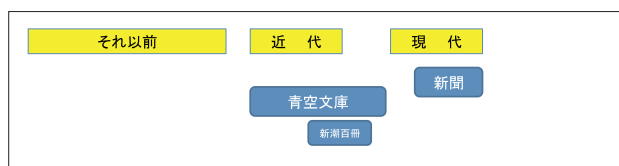


図4 日本語コーパス整備の経緯Ⅰ：1990年代

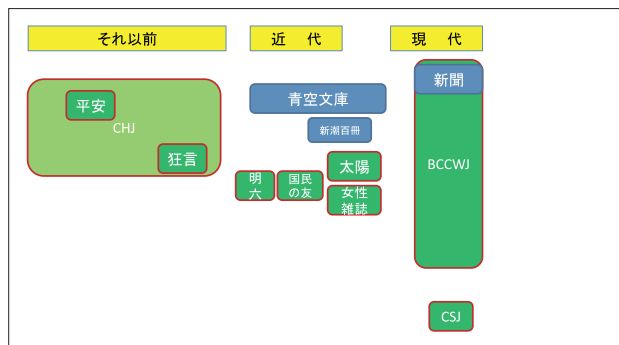


図5 日本語コーパス整備の経緯Ⅱ：現状

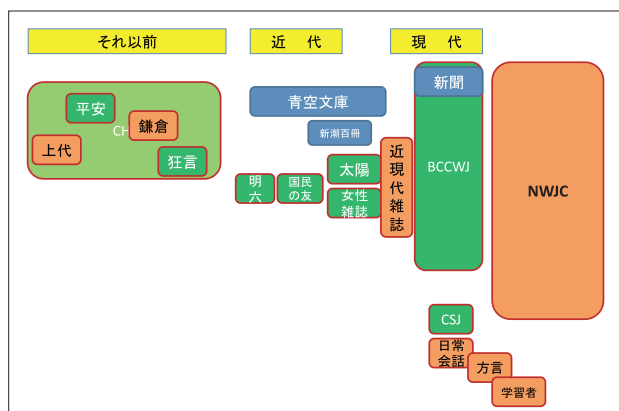


図6 日本語コーパス整備の経緯Ⅲ：2021年の目標

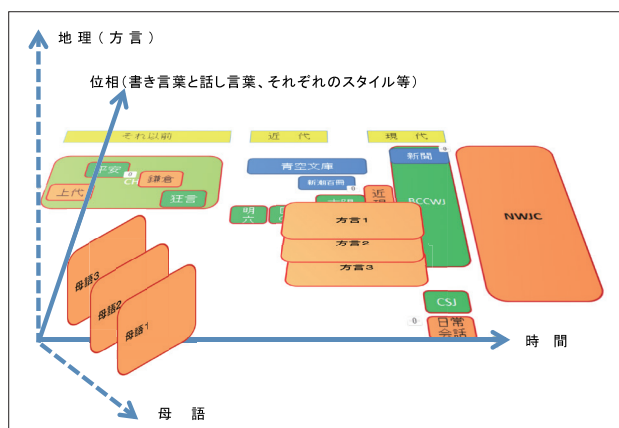


図7 各種コーパスの相互関係

パス』があります。近代語の各種雑誌のコーパス(『明六雑誌』『国民の友』など)にくわえて、『日本語歴史コーパス』のうち、平安時代と室町時代のデータが現時点で公開されています。

このように、二〇年間でかなり進んではきましたが、まだ、いろいろ穴があいています。そこで、これから六、七年の期間になにをするかという、図6のような目標を立てています。

先ほど触れた、『国語研日本語ウェブコーパス(NWJC)』、対話とか、多人数の会話を記録した、日常会話のコーパス、方言のコーパス、日本語学習者のコーパス(I・JAS)、そして、『日本語歴史コーパ

ス』も上代や鎌倉時代のデータを充実させていき、奈良時代から現代まで、細い線でよいからなんとかつながるように整備する計画です。

もう少し別の見方をすると図7のようになります。時間の軸があり、話し言葉や書き言葉という、いわゆる位相の軸があつて、その上に地理的な差異、そして話し手の母語の影響をいれる多次元空間が、日本語の内的多様性の全体です。今後そう遠くない時期に、この空間全体を対象として、包括的なコーパス検索を可能にする利用環境を整備していく予定です。

## オンライン検索ツール

さて、こういったコーパスは、公開しただけではあまり活用してもらえません。コーパスのデータは複雑な構造をしているので、検索にはかなり高いコンピュータリテラシーが必要とされるからです。そこ

**文字列検索ツール**  
『少納言』2007

- ・登録不要
- ・BCCWJ(1億語)が対象

**形態論情報検索ツール**  
『中納言』2010

- ・要登録(無償)
- ・BCCWJとCHJが対象

**超大規模コーパス**  
**検索ツール『梵天』**  
(開発中、2016公開予定)

- ・NWJCに対する、文字列検索、形態論情報検索、係り受け検索

**図8 オンライン検索ツール**

で『現代日本語書き言葉均衡コーパス』からは、検索用ツールもあわせて公開し、継続的に作りこむ努力をしてきました。現在、『少納言』と『中納言』という二種類のオンライン検索ツールが稼働しており、今年の秋からは新たに『梵天』というオンラインツールも公開する予定です(図8)。

一番広く使われている検索ツールは『少納言』です。これは登録不要でどなたにもお使いいただけます。現在は『現代日本語書き言葉均衡コーパス』のデータ一億語を対象としており、年間で八〇万回程度利用されています。

『中納言』では『現代日本語書き言葉均衡コーパス』と『日本語歴史コーパス』のデータを形態論の情報を利用して検索することができます、年間で三〇万件ほどの検索があります。近日中に『多言語母語の日本語学習者横断コーパス(I-JAS)』も『中納言』で検索可能になる予定です。『中納言』の利用も無償ですが、著作権保護の関係で利用申請をお願いしています。

図9は『中納言』の検索結果画面です。検索しているのは、動詞「そびえる」の終止形が、それ自身で文末を構成している例です。画面が細かすぎてよく見えないと思いますが、画面に表示されている用例をみると検索対象の動詞は「そびえる」と仮名で表記されていたり、「聳える」と漢字仮名交じりで表記されていたりします。そのような表面的な表記の相違に惑わされずにすべての用例を検索できるのが、解体論情報を使った検索の強みです。もちろん種々の活用形の違いなども吸収することができます。ちなみに『現代日本語書き言葉均衡コーパス』には、どこが文末かの情報も付与されているので、それを検索に



13 件の結果が見つかりました。

□ テーブルの幅を固定 短 ▾

(検索対象語数: 124,100,964、空白・記号・補助記号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	活用形	レジスター	執筆者	書名/出典	出版者	出版年
LBs2_00036	を[潜ると直ぐ]中央[広場]で、[堂々]たる[大階段]上[に]街の[シンボル]、ドゥオモが]	繰える	。#十[世紀]起源、十[三]世紀[拡張]、十[八]世紀[バロック]様式[に]改築、十[九]	終止形 一般	図書館・書籍	斐 澄 (著)	イタリア再発見	中央公論事業出版	2004
PM31_00272	施設「あかつき」の[村]は[小高い]丘の[上]にある。#背後[に]は[赤城]連峰が]	そびえる	。#村の[入り口]には、[門]も[柵]もない。[村長]の[石川]龍也[神父]。	終止形 一般	出版・雑誌	潮川 正仁 (著)	暮しの手帖	暮しの手帖社	2003
OY15_03179	大分[由布市]湯布院[町]([旧国]豊後[国])にある[温泉]で[すぐそば]に]	繰える	。#由布[岳]の[恵み]を[受けた]豊富な[湯量]を[誇る]かつては[ひなびた]温泉で[団体]観光	終止形 一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
LB02_00097	間の[進行]方向[左]に[男性]山([二千四百][八十四]メートル)などの[日光]連山が]	そびえる	。#鬼怒川[川]は[栃木]県[北部]の[山地]に[みなもと]を[発し]、[茨城]県[南部]で	終止形 一般	図書館・書籍	竹内 均 (著)	竹内均の日本の地誌	ニュートンプレス	2000
PB56_00113	の[晴れた]日など[は]電車[の]バック[に]雪を[]いた[三千]メートル級の[山々]が]	そびえる	。#東京[に]東京[急行]電鉄が[世田谷]の[下町]に[世田谷線]を[運行]するが、[環状]7	終止形 一般	出版・書籍	谷川 一巳 (著)	ローカル線こだわりの旅	角川学芸出版、角川書店(発売)	2005
LBK2_00051	入ったのだ[な]、[と]いう印象を[オレ]に[与えた]。#前方[は]険しい[山]が]	そびえる	。#右側を[谷]に[して]、[登って]きた[時]は[りり]急な[坂]を[下っ	終止形 一般	図書館・書籍	池田 拓 (著)	南北アメリカ徒歩縦横断日記	無明舎出版	1996
PB49_00244	。#現在[の]社殿[は]伊達家[が]運営[した]もので、[二百][二]段の[石段]が]	そびえる	。#境内[にある]シオガマクラ[は]国の[天然記念物]に[指定]されている。[塩竈]神社	終止形 一般	出版・書籍	実著者不明	奥の細道	学習研究社	2004

動詞「そびえる」の終止形がそれ自身で文末を構成している例

図9 『中納言』の検索結果画面

図10 『中納言』：検索条件指定画面

利用できます。

図10は、『中納言』の検索条件指定画面です。単語が「そびえる」、活用形が終止形で、文末から二語以内にある例を探せ、と指定しています。このような指定を行うと、内部的には検索式が形成され(図11)、これを保存することができ、後日、同じ検索を実行することが可能になります。

図12は『国語研日本語ウェブコーパス』のために開発中のオンライン検索ツール『梵天』の画面です。動詞の「そびえる」を検索の対象としていますが、それだけでなく、山がそびえる、ビルがそびえる、のように「名詞+が」が「そびえる」を修飾している例

(両者が係り受けの関係にある例)を検索しています。図13が検索結果です。係り受けの関係にある語は隣接しているとは限りません。たとえば、「櫓がひときわそびえる」のように、あいだに一語入っている場合がありますが、このような用例も検索できます。なかには、「レインボーブリッジが、その名の通り虹のような空に弧を描いて東京湾の出口に高くそびえている」のように、遠距離の係り受けが生じることもあります。これらも一網打尽にひっかけることができます。

## コーパスが捉えた現代日本語の変異

さて、ここからはコーパスを利用して現代日本語の多様性の実態を調べてみることにしましょう。とりあげるのは、いずれも内省するのが難しい例です(図14)。

最初に「NHK」はどのように発音されているでしょうか？ 少し考えてみてください。いろいろな発音の仕方がある

キー：(語彙素 = “聳える” AND 活用形 LIKE “終止形%”) WITHIN 2 WORDS FROM 文末 WITH OPTIONS unit= “1” AND tglBunKugiri= “#” AND tglWords= “20” AND limitToSelfSentence= “0” AND tglKugiri= “|” AND endOfLine= “CRLF” AND encoding= “UTF-16LE” AND tglFixVariable= “2”

動詞「そびえる」の終止形がそれだけで文末を構成している例を検索した際に自動生成される検索式。保存して再利用できる。

図11 『中納言』の検索式

超大規模corpus検索システム (仮称)

文字列検索 品詞列検索 係り受け検索

### 係り受け検索

0 1

<表層形>  
名詞  
<品詞2>  
<品詞3>  
<品詞4>  
<活用型>  
<活用形>  
語彙素読み  
語彙素

が

<品詞1>  
<品詞2>  
<品詞3>  
<品詞4>  
<活用型>  
<活用形>  
語彙素読み  
語彙素

1:

<表層形>  
<品詞1>  
<品詞2>  
<品詞3>  
<品詞4>  
<活用型>  
<活用形>  
語彙素読み  
語彙素  
聳える

URLドメイン

※ac.jpなど末尾2パート

検索

検索条件を初期化する

「名詞+が」が「そびえる」を修飾している(係っている)例の検索

図12 『国語研日本語ウェブコーパス』オンライン検索ツール『梵天』



ことはわかると思いますが、どれくらいあつて、どれが一番多いでしょうか。

また、いわゆる、ら抜きことばの「来られる」と「来れる」は、話し言葉で検索したとき、どっちが多いか？ これについては、皆さん意見が一致すると思いますが、どれくらい多いかも考えてみてください。

さらに、動詞に「です」がつく、「読むです」「行くです」の形。話し言葉で使う人はいるようですが、書く人はいるでしょうか？

もう一つ、可能の意味で「読める」「行ける」ではなく、「読めれる」「行けれる」と書く人はどのくらいいるか？ そんな人はいないと思うかもしれませんが、実はいるんですね。

そして、「～しそわない」と「～なさそう」はどちらが多いか？ 「～すべきでない」と「～しないべき」はどうか？ 少し考えてみてください。

では、これから実際の検索結果を紹介しましょう。まず、「NHK」については、

22	名詞 助詞 [路地の] [彼方に]	名詞 助詞 [高層ビルが] [動詞 助詞 そびえて]	
23		名詞 助詞 [櫓が] [ひとざわ] 動詞 助詞 動詞 助動詞 [聳えて 見えます]	
24		名詞 助詞 [連山が] [形容詞 遠く] [動詞 補助記号 聳え、]	名詞 名詞 名詞 助詞 助詞 [反対側正面に] [神社の] [鳥居越し 動詞 名詞 名詞 名詞 助詞 に] [標高] [504mの] [八剎山 補助記号 名詞 (観音)]
25	名詞 名詞 助詞 [煉瓦造りの] [城壁で] [囲まれ] 名詞 助詞 助詞 [城塞には] [形容詞 接尾辞 高] [名詞 名詞 記号 助詞 80mの] 名詞 名詞 助詞 [トロイツカヤ塔など] [19の]	名詞 助詞 [尖塔が] [動詞 助詞 そびえて 動詞 いる]	
26	名詞 助詞 [予想は] [外れまし た が、] 助詞 連体詞 [大きな]	名詞 名詞 助詞 [レインボーブリッジが] 補助記号 連体詞 名詞 、 [その] [名 助詞 名詞 名詞 助詞 形容詞 の] [通り 虹のよう 助動詞 名詞 助詞 な] [空に] [名詞 助詞 弧を] [動詞 助詞 描いて] [名詞 接尾辞 助詞 名詞 東京 港の] [出口 助詞 形容詞 動詞 に] [高く] [聳え 助詞 動詞 ている]	名詞 助詞 [ことは] [形容詞 良い] [ことです]

図13 係り受け検索結果の画面

- ・「NHK」はどのように発音されているか？
- ・「来られる」と「来れる」は話し言葉でどちらが多いか？
- ・「読むです」「行くです」等と書く人はいるか？
- ・「読めれる」「行けれる」は？
- ・「～しそわない」と「～しなさそう」はどちらが多いか？
- ・「～すべきでない」と「～しないべき」は？
- ・Etc.

図14 内省してみてください

発音	頻度
エヌエチケー	132
エネーチケー	24
エヌエッチケー	9
エヌエイチケー	7
エヌエチケ	3
エネーチケ	3
エネエチケー	2
エヌエスケー	1
エヌチケー	1
エネーシケー	1

←発音辞書の見出し

←発音辞書の見出し

図15 『日本語話し言葉コーパス』の検索結果

次は、話し言葉で「来られる」と「来れる」のどちらが多いか。これはいうまでもなく年齢差と関係しています。図16は、横軸が話者の生まれた年代を示しています。このグラフの左半分は文化庁の国語課が二〇〇一年に実施した世論調査のデータで、ご覧の通り、「来れる」の抜き言葉がどんどん増えてきて、一九七〇年代生まれの人のグループでは伝統的な「来られる」を逆転しています。これに対して、『日本語話し言葉コーパス』で

15の結果が得られます。これは『日本語話し言葉コーパス』に記録された日本語の独話データの分析ですが、一位は「エヌエチケー」で、圧倒的に高い数字を示しています。ご覧のように圧倒的な一位なのですが、これをあてられる人はほとんどいません。

日本語の発音辞典として有名なNHKのものと三省堂のものを調べてみると、一位の「エヌエチケー」だけでなく、二位の「エネーチケー」もみだしにのっていません。三位の「エヌエッチケー」と「エヌエイチケー」がでてくるだけです。辞典にはそれぞれの編集方針がありますから一概に批判はできませんが、実態を捉え損ねていることはたしかです。

分析すると、図16右の結果となります。全体のパターンは同じだと思われるかもしれませんが、逆転の時期が三〇年ズレています。一九四〇年代生まれの人は、二〇〇〇年前後に調べたとき、すでに逆転しています。アンケートで意識を調べた場合と、実際の言語行動を調べるのでは、このようなズレがでてきます。これは言語調査に携わる者にとって非常に重要な問題だと私は考えています。

ここからは書き言葉の例になります。まず、「読むです」「行くです」等の「動詞+です」の形(図17、1行目)。これを『現代日本語書き言葉均衡コーパス』で調べると、一億語に対して八二例が見つかります。それに対して、二五〇億語の『国語研日本語ウェブコーパス』の検索では一二、〇〇〇例近くみつかります(念のために注意しておく、『国語研ウェブコーパス』は現在インデックス作成中であり、二五〇億語全体が検索されているとはかぎりません。以下同様です)。

図18に『現代日本語書き言葉均衡コーパス』の検索結果の一部を示しました。村上春樹や柳田邦男といった有名な著述家の書いた文章が含まれています。村上春樹さんの小説『世界の終わり』とハードボイルド・ワンダーランド』にでてくる例は、マッドサイエンティストの「博士」が変なしゃべり方をしているもので、いわゆる役割語です。一方、柳田さんの例はそのような例ではありません。ともかく、「動詞+です」はサザエさんのタラちゃんだけでなく、書き言葉でも、けっこう普通に用いられているわけです。その条件を分析するといろいろおもしろいことがわかってきますが、きょうはここまでとします。

次は可能を表す「読める」「行ける」などの形。可能動詞に可能の助動詞がついたものとみて、私は二重可能形と呼んでいます。最



『日本語話し言葉コーパス』における行動

図16 『日本語話し言葉コーパス』の検索結果

図17 『現代日本語書き言葉均衡コーパス』(BCCWJ)と『国語研日本語ウェブコーパス』(NWJC)の検索結果

図18 「動詞+です。」

近では音の特徴から「レ足す言葉」と呼ぶ人が多いようです。これは、『現代日本語書き言葉均衡コーパス』には六例しかなく、また動詞も「行く」にかぎられています。これだと間違いかないという気もしますが、『国語研日本語ウェブコーパス』を調べると、多くの動詞に生じていることがわかります(図19)。二百五十億語という規模がものをいって、生起確率の低い現象が拾いあげられています。

次の「しそうにない」と

BCCWJ

動詞	レル	レナイ	レタ
行ケ	3	2	1
聞ケ	0	0	0
書ケ	0	0	0
遊ベ	0	0	0
歩ケ	0	0	0
出来	0	0	0
描ケ	0	0	0
飛ベ	0	0	0
聴ケ	0	0	0
読メ	0	0	0

NWJC

動詞	レル	レナイ	レタ
行ケ	32	28	2
聞ケ	5	1	2
書ケ	4	1	0
遊ベ	4	0	0
歩ケ	3	0	0
出来	2	0	0
描ケ	2	2	1
飛ベ	1	0	1
聴ケ	1	1	0
読メ	1	1	0

図19 「二重可能」(レ足す言葉)

15	名詞 動詞 助動詞 助動詞 [柚葉] [カナでも]	動詞 [くっつか]	形容詞 形容詞 助動詞 助動詞 名詞 [なさそう] [の] [こと]
16	名詞 形容詞 補助記号 感動詞 補助記号 名詞 助動詞 [遠慮なく] [「ああ、」] [大久保で] 動詞 助動詞 名詞 動詞 助動詞 助動詞 助動詞 [やったら] [カウンター] [食ったの] [で]	名詞 動詞 [反抗し]	形容詞 形容詞 助動詞 助動詞 名詞 助動詞 名詞 動詞 助動詞 [なさそう] [な] [どこで] [溜飲下げるの] 助動詞 補助記号 助動詞 [ね] [と]
17	名詞 名詞 助動詞 動詞 名詞 助動詞 [躊躇半端に] [踊る] [自分を]	動詞 [許さ]	形容詞 形容詞 助動詞 助動詞 助動詞 補助記号 副詞 代名詞 [なさそう] [な] [の] [で] [多分] [それ] 形容詞 名詞 助動詞 動詞 助動詞 [正しい] [気が] [します]
18	名詞 助動詞 [趙雲の] 名詞 助動詞 [子供って] 動詞 助動詞 助動詞 助動詞 助動詞 [出でこないか] 助動詞 助動詞 動詞 助動詞 補助記号 名詞 接尾辞 助動詞 [なあと] [思うけど] [三國志に] [形容詞 詳しい] [家人曰く] [あまり] [活躍して] 助動詞 助動詞 助動詞 助動詞 名詞 助動詞 助動詞 助動詞 [なかつた] [の] [こと] [な] [の] [で]	名詞 動詞 [期待でき]	形容詞 形容詞 助動詞 [なさそうです]
19	名詞 助動詞 [贅沢は]	動詞 形容詞 形容詞 助動詞 [好まなさそう] [だ] 助動詞 補助記号 [けど] [、]	代名詞 助動詞 動詞 名詞 助動詞 形容詞 助動詞 [そこに] [ある] [材料で] [簡単に] 形容詞 名詞 助動詞 動詞 助動詞 助動詞 形容詞 [美味しい] [ものを] [作れるのは] [凄い] 助動詞 動詞 助動詞 [と] [思ってる]
20	名詞 助動詞 [誓は]	動詞 形容詞 形容詞 助動詞 [つまらなさそうに]	名詞 助動詞 動詞 助動詞 [溜息を] [ついた]
21	名詞 助動詞 補助記号 補助記号 [数] [は] [・] 動詞 助動詞 動詞 [入れてもらえ]	名詞 助動詞 形容詞 助動詞 助動詞 補助記号 補助記号 補助記号 [なさそう] [だし] [・] [・]	
22	名詞 名詞 助動詞 副詞 形容詞 助動詞 [薔薇乙女は] [正直] [怖くて]	動詞 [読め]	形容詞 形容詞 助動詞 助動詞 補助記号 名詞 接尾辞 助動詞 [なさそう] [だ] [が] [く] [薔薇的に] 補助記号 [・]

図20 「しなそう」

六二九に對し、「ししてしなさそう」が七五で、「しなさそう」が少なくともないので、ウェブコーパスで調べると、一七、〇〇〇に對して一一、〇〇〇くらいと、頻度差がほとんどなくなっています。もはや誤りだと切つて捨てることができないう状態です。ウェブコーパスでは、「くつつかなさそう」「反抗しなさそう」「許さなそう」など、動詞もいろいろなのができます(図20)。

「しすべきでない」と「しなすべき」では、前者が正しいといわれています。誤りとされている「しなすべき」は、『国語研日本語ウェブコーパス』でもさほど多くは觀察されませんが、分布パターンがちよつとおもしろい(図21)。「萌えるべきなのか、萌えないべきなのか」「消すべきか、消さないべきか」「分けるべきだ分けないべきだと論争する」のような文脈、(ハムレット文脈と私は呼んでいます)が非常に多く、この文脈で変化が先行していることがわかります。

最後に、コーパスはただ大きければよいのではないという例を示しましょう。例としていわゆる自動詞の「泣く」と「死ぬ」が目的語を伴って他動詞のように用いられている例を検索します。「しを泣く」「しを死ぬ」の頻度は非常に低いものの、絶無ではありません。『現代日本語書き言葉均衡コーパス』を調べると、「しを泣く」が一例、「しを死ぬ」が四例見つかります。

さて、それでは『国語研日本語ウェブコーパス』を調べたらもっとたくさん見つかるかというと、実はまったく見つかりません。これはなぜでしょうか。

『現代日本語書き言葉均衡コーパス』に見つかった「しを死ぬ」

3	「 <small>名詞 助詞</small> 嗚呼、」 「 <small>動詞 助動詞 助詞 補助記号</small> 萌えるべきなのか、」	「 <small>動詞 助動詞 助動詞 助詞</small> 萌えないべきなのか」	
4	「 <small>名詞 助詞</small> 話は」 「 <small>動詞 助動詞 助詞 補助記号</small> 変わりますか、」 「 <small>記号 補助記号</small> Q :」 「 <small>名詞 助詞</small> 会社の」 「 <small>名詞 助詞</small> トイレの」 「 <small>名詞 補助記号</small> 電気、」 「 <small>名詞 助詞</small> 必要の」 「 <small>形容詞</small> ない」 「 <small>名詞 助詞</small> どこに」 「 <small>動詞 助動詞 助詞 補助記号</small> 消すべきか、」	「 <small>動詞 助動詞 助動詞 助詞</small> 消さないべきか」	
5	「 <small>動詞</small> 話し」 「 <small>動詞 助詞</small> 変わって」 「 <small>空白 記号</small> p」 「 <small>動詞 助詞</small> 入れるか」	「 <small>動詞 助動詞 助動詞 助詞</small> 入れないべきか」	「 <small>動詞 助動詞 助詞 助動詞 助詞 助詞 記号</small> 迷ってるんだよなaw」
6		「 <small>動詞 助動詞 助動詞 助動詞 助詞</small> 話さないべきなの」 「 <small>助詞 助動詞 助動詞</small> かなんです」	
7	「 <small>動詞 助詞 助詞</small> 話している」 「 <small>名詞 助詞</small> 側の」 「 <small>名詞 助動詞</small> 勝手な」 「 <small>名詞</small> 見解」 「 <small>助詞</small> に」 「 <small>動詞 補助記号</small> より、」 「 <small>動詞 助動詞 助動詞</small> 分けるべきだ」	「 <small>動詞 助動詞 助動詞 助動詞 助詞</small> 分けないべきだ」と 「 <small>名詞 動詞 助詞 助詞</small> 論争するのは」 「 <small>動詞 助詞 動詞 助詞</small> ばかけていると」 「 <small>動詞</small> 思う」	
8	「 <small>形容詞 助動詞</small> 余計な」 「 <small>名詞 助詞</small> 口出しを」	「 <small>動詞 助動詞 助動詞 助詞</small> しなすべきか」	
9	「 <small>名詞 名詞 名詞 助詞</small> 予防接種以外の」 「 <small>名詞 助詞</small> ことに」 「 <small>動詞 助詞 助詞</small> ついても」 「 <small>補助記号</small> 、」 「 <small>接尾辞 名詞 接尾辞 助詞</small> お医者様の」 「 <small>名詞 助詞 補助記号</small> 立場から、」 「 <small>副詞</small> かなり」 「 <small>動詞 助動詞</small> 突っ込んだ」 「 <small>名詞 助詞</small> 意見や」 「 <small>名詞 名詞 助詞 助詞</small> 具体例なども」 「 <small>動詞 助動詞 助詞 助詞 補助記号</small> 記されていて、」 「 <small>補助記号 動詞 助動詞 助詞</small> 「受けるべきか」	「 <small>動詞 助動詞 助動詞 助詞</small> 受けないべきか」 「 <small>補助記号 助詞</small> 」を」	「 <small>名詞 動詞 助詞 助詞</small> 検討するには」 「 <small>副詞</small> ども」 「 <small>名詞 助詞</small> 参考に」 「 <small>動詞</small> でき」 「 <small>助動詞 助動詞</small> ました」
10	「 <small>名詞 接尾辞 助詞</small> 予備校に」 「 <small>動詞 助動詞 助詞</small> 行くべきか」	「 <small>動詞 助動詞 助動詞 助詞</small> 行かないべきか」 「 <small>補助記号</small> ...」	
11	「 <small>名詞 接尾辞 助詞</small> 予備校に」 「 <small>動詞 助動詞 助詞</small> 行くべきか」	「 <small>動詞 助動詞 助動詞 助詞</small> 行かないべきか」 「 <small>補助記号 名詞 名詞 助動詞</small> ...」 「 <small>助動詞 助動詞 助詞</small> BIGLOBEな」 「 <small>助動詞 助動詞 助詞</small> んでも」	「 <small>名詞 接尾辞</small> 相談室」

図21 「～シナイベキ」

はすべて「彼は自分の死を死んだ経験者だった」のような「死を死ぬ」の例であり、書き手は文学者・評論家（有島武郎、田村隆一、五島勉、南伸坊）にかざられています。『現代日本語書き言葉均衡コーパス』では、韻文を含めて多くの文芸書がサンプリングの対象になっていますが、どうもウェブには日本語の文学作品はあまり載っていないようです。規模は大きくなくても、綿密に設計して構築した均衡コーパスには固有の価値があることを示す例といえるでしょう。

## まとめにかえて

最後に、言語資源を整備すると、今後の言語研究にどのような影響が及ぶかという問題を少し考えてみたいと思います。今日の話の後半で紹介した例でおわかりいただけたと思いますが、内省やアンケートに頼らず言語内多様性を把握しようとしてもうまくいかない例がたくさんあります。コーパスを利用することで、言語内多様性を実際の言語行動のデータに基づいて研究する可能性ができました。これが重要だと私は考えています。さきほどの、「しなさそう」のように、量的にみるとはや逆転が生じそうな現象の場合、それでも「しきそうにない」が正しくて「しなさそう」は誤りだと主張するためには、その根拠をきちんと示すことが要請されます。単なる直観では説明になりません。

従来の言語研究は、ややもすると規範的で正しいと思われるもののだけを対象として進められる傾向がありました。コーパスの存在を前提とした今後の研究では、正しくないとされているものでも、実際

に用いられているものは、すべて対象とした研究が行われるようになるだろうと思います。

要するに、複雑多様な言語現象を過度に単純化せず、複雑なものは、複雑なままに理解しようとする姿勢が求められています。そのためには、従来の言語研究法にくわえて、情報科学や統計科学との連携が不可欠になってくるでしょう。昨今、文理融合という言葉が頻繁に耳にするようになってきましたが、コーパスを用いた言語研究はその好例を提供できるのではないのでしょうか。

これで私の発表をおしまいします。

