＜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

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～

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変化も、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のっています。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って、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「ヤッパリ」か「ヤハシ」か、等々。これらは一種の確率的な変動であって、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、ではどうやって把握するか。思いつくままに、こんなのがある、あんなのもあるといっても、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っていること（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。

・多様性の源
　－歴史的多様性
　　・内発的変化
　　・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

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）』（構築一九九九～二〇〇三年度、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・広報誌・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、現在もっと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）』（構築二〇一一～二〇一五年）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。現代語については、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ていましたが、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

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と、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って、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。

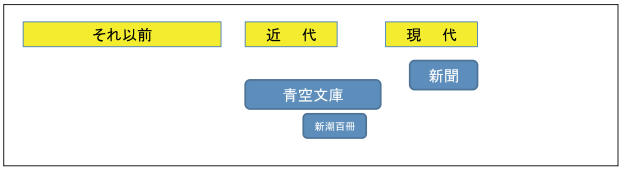


図4　日本語コーパス整備の経緯Ⅰ：1990年代

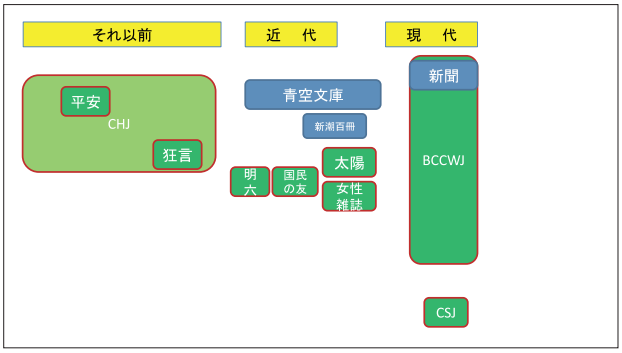


図5　日本語コーパス整備の経緯Ⅱ：現状

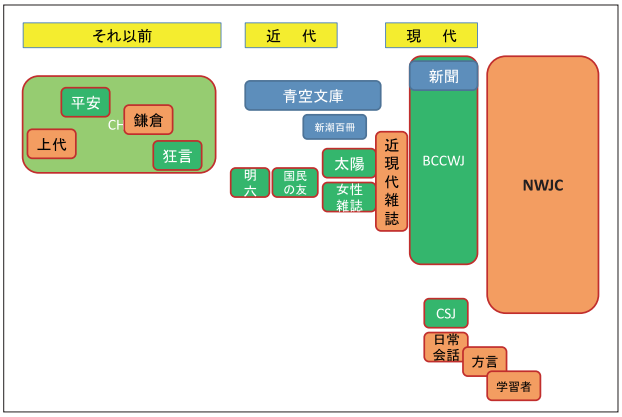


図6　日本語コーパス整備の経緯Ⅲ：2021年の目標

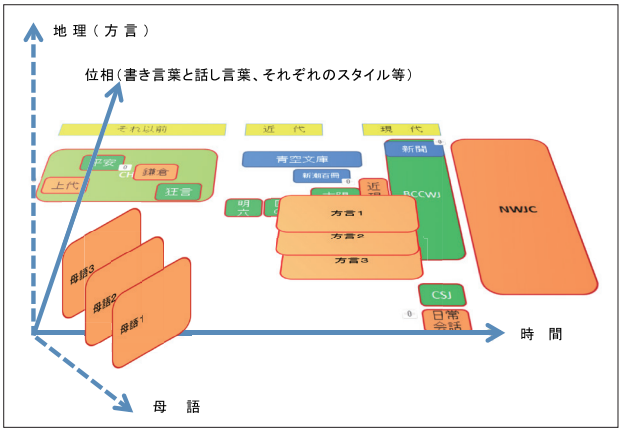


図7　各種コーパスの相互関係

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、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

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

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|	聳える	|。#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	塞　滋(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|能|也|神父|。	終止形-一般	出版・雑誌	瀬川　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|	聳える	#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
LBo2_00097	間|の|進行|方向|左手|に|男体|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|は|栃木|県|北西|部|の|山地|に|みなもと|を|発し|、|茨城|県|南西|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|パック|に|雪|を|頂い|た|三千|メートル|級|の|山々|が|	そびえる	|。#東京|:|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|より|も|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|造営|し|た|もの|で|、|二百|二|段|の|石段|が|	そびえる	|。#境内|に|ある|シオガマサクラ|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ます。したがって、自分がどんな検索をしたかをすべて記録に残しておくことができ、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かぎりません。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が、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

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

「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、どれくらいあって、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そうですが、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、「～しそうにない」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」ではどうか？　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す。まず、「NHK」については、図

22	名詞 助詞 ［路地 の ］ 代名詞 助詞 ［彼方 に ］	名詞 名詞 助詞 ［高層 ビル が ］ ［ 動詞 助詞 そびえ て ］	
23		名詞 助詞 ［櫓 が ］ 副詞 ［ひときわ ］ 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 は ］ 名詞 助詞 ［神社 の ］ 名詞 接尾辞 ［鳥居 越し 動詞 名詞 ［に ］ ［標高 ］ 名詞 名詞 名詞 名詞 助詞 ［５ ０ ４ ｍ の ］ 名詞 名詞 ［八 剣山 補助記号 名詞 （ 観音 ］
25	名詞 名詞 助詞 ［煉瓦 造り の ］ 名詞 助詞 ［城壁 で ］ 動詞 助動詞 補助記号 ［囲ま れ 、 ］ 名詞 助詞 助詞 ［城塞 に は ］ 形容詞 接尾辞 ［ 高 さ ］ 名詞 名詞 記号 助詞 ［ ８ ０ ｍ の ］ 名 詞 名詞 助詞 ［トロイツカヤ 塔 など ］ 名詞 名詞 助詞 ［ １ ９ 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 ］ 動詞 助動詞 助動詞 助詞 補助記号 ［外れ まし た が 、 ］ 形容詞 ［高く 助詞 て ］ 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 形状詞 の ］ ［通り 虹 の よう 助動詞 名詞 助詞 な ］ ［空 に ］ ［ 名詞 助詞 動詞 助詞 弧 を ］ ［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 ］ ［ 動詞 動詞 て いる ］	名詞 助詞 形容詞 名詞 助動詞 ［こと は ］ ［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？
・「～しそうにない」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

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。

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」のら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とでは、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。

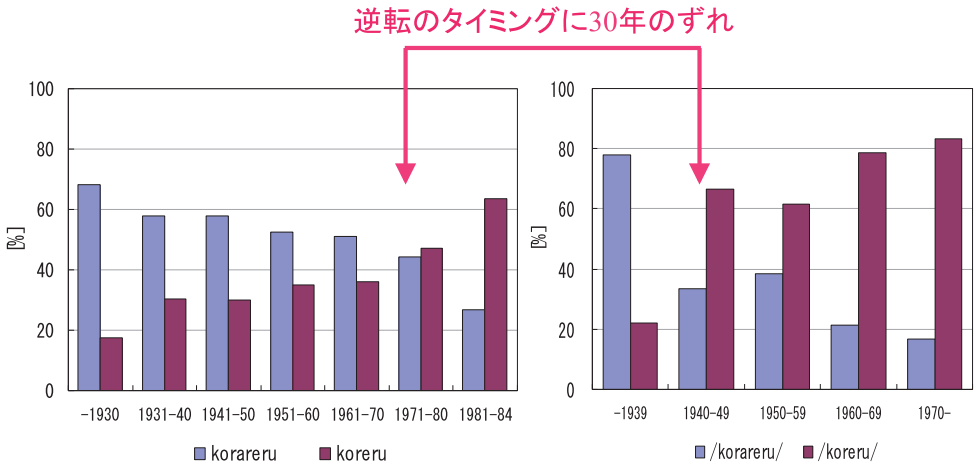
発音	頻度	
エヌエチケー	132	
エネーチケー	24	
エヌエッチケー	9	←発音辞書の見出し
エヌエイチケー	7	←発音辞書の見出し
エヌエチケ	3	
エネーチケ	3	
エネエチケー	2	
エヌエスケー	1	
エヌチケー	1	
エネーシケー	1	

図15 『日本語話し言葉コーパス』の検索結果

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と『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「読めれる」「行けれる」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が、最



文化庁国語課による世論調査 2001　　『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

表現	BCCWJ（1億語）	NWJC（200億語）
動詞＋デス	82	7,172
行ケレル、行ケレナイ、行ケレタ	6	62
～シソウニナイ／～シナサソウ	629 / 75	17,077 / 10,943
～スベキデナイ／～シナイベキ	245 / 11	3,018 / 205

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

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
OB2X_00159	あんたの|おっしゃるとおり|です。#そのことに|について|は|私|も|私|なり|に|反省|して|	おる|です	。#後悔|は|せんが|反省|は|して|おる|です。|しかし|弁解|する|わけ|じゃ|ない|です	終止形-一般	特定目的・ベストセラー	村上 春樹(著)	世界の終りとハードボイルド・ワンダーランド	新潮社	1985
OY14_44167	の|で|気|が|付かない|って|パターン|が|多い|です。#起きて|たら|ぎっと|どうぞ|って|言う|と|	思う|です	|^-^#あくまで|も、|起きて|たら|ね|(笑)|greatmz	終止形-一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
OC14_03979	・|塑節|＋|醤油#イヌ|ご飯|・|・|味噌|汁#たまごかけ|ご|・|・|生卵|＋|醤油|を|	かける|です	。#この|場合|おかず|は|無く|、|漬物|ぐらい|です。|おかず|が|ある|なら|、|普通|に|白米	終止形-一般	特定目的・知恵袋		Yahoo!知恵袋	Yahoo!	2005
OB2X_00159	知の|拡張|を|前||にする|と|それ|以外|の|状況|が|眼中||にな|なって|しまう|きらい|が|	ある|です	。#また|それ|なれ|ば|にそ|科学|も|間断|なき|進歩|を|遂げて|きた|わけだ|。	終止形-一般	特定目的・ベストセラー	村上 春樹(著)	世界の終りとハードボイルド・ワンダーランド	新潮社	1985
OY14_46352	二|mm|用|です|た|～|＞＜|ま|～|良い|です|わ|また|十|一|、|五|mm|用|入手|	する|です	。#新型|NOVAROSSI|と|O|S|T|Z|用|です|ね|早速|破損|した|SIRIO|で|練習|です	終止形-一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
OY14_43223	だ|と|いう|の|は|に|・|・|・|、#まだまだ|寒い|です|なあ|～#で|も|最近|感じる|こと|が|	ある|です	。#日|が|長く|なった|よ|なあ|～#daira|の|会社|は|十|七|に|三十|が|終業	終止形-一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
OY15_01489	、|違う|って！|」#糞|夫人#「|は|わ|わ|わ|・|・|・|み|、|身|の|危険|を|	感じる|です	。#二|は|後退|です」#母|「|高順|、|糞|夫人|を|追撃|して|撃破|して	終止形-一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
OC09_03684	。|　|ホコリ|や|寒暖|の|差||に|気|を|付け|て|下さい|ね|。#そして|規則|正しく|、|寝る|	食べる|です	。#私|は|幸い|、|薬|が|効い|た|の|と|生活|習慣|を|見直し|た|事||により	終止形-一般	特定目的・知恵袋		Yahoo!知恵袋	Yahoo!	2005
PB24_00273	に|綱|を|持っ|とる|家|が|二|軒|あって|、|人|夫|に|なって|ら|共|に|綱|引き|を|	手伝う|です	。#、いろんな|魚|が|取れ|て|。#(|中略|)|村|の|者||は|気心|知れ|た|者|でしょ	終止形-一般	出版・書籍	柳田 邦男(著)	ホスピス通りの四季	新潮社	2002
OY03_08881	も|何|も|解決|しない|の|で| |Que| |Sera| |Sera| |ケ・セラ・セラ| #なる|よう|に|	なる|です	。#|スペイン|語|ですが|・|ト|ゥ#英語|では| What| |will| |be|	終止形-一般	特定目的・ブログ		Yahoo!ブログ	Yahoo!	2008
LBb7_00014	調|の|ド|レ|ミ|が|読め|れ|ば|全部|の|調|の|ド|レ|ミ|が|読め|た|に|と|に|	なる|です	。#　だって|、|われ|われ|日本|人|は|1|つ|の|漢字|を|いろいろ|に|読み分け|て|いる	終止形-一般	図書館・書籍	相原 末治(著)	やさしい楽譜の読み方	音楽之友社	1987
OB2X_00153	は|極秘|事項|で|部外|者|は|誰|も|知ら|ない|はず|です|」#「|私|は|知っ|て|	おる|です	。#|組織|の|上層部|と|は|かなり|太い|パイプ|が|通じて|おり|まして|な	終止形-一般	特定目的・ベストセラー	村上 春樹(著)	世界の終りとハードボイルド・ワンダーランド	新潮社	1985

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あ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「～しなさそう」について、普通、「しなさそう」は間違いだといわれます。しかし、コーパスの検索結果では驚くべき結果がでてきます（図17、3行目）。もともと『現代日本語書き言葉均衡コーパス』でも、「～しそうにない」が

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 助詞 助詞 ［釉薬］［かけ て も］	動詞 ［くっつか］	形容詞 形状詞 助詞 助詞 名詞 ［なさ そう と の］［こと］
16	名詞 形容詞 補助記号 感動詞 補助記号 名詞 動詞 ［逡巡 なく］［「 ああ 、 ］［大久保 で］ 動詞 助動詞 名詞 動詞 助動詞 助詞 助動詞 ［やっ たら］［カウンター］［食らっ た の で ］	名詞 動詞 ［反抗 し］	形容詞 形状詞 助動詞 名詞 動詞 名詞 動詞 助詞 動詞 ［なさ そう な］［とこ で］［溜飲 下げる の 助詞 補助記号 助詞 ね 」 と］
17	名詞 名詞 助動詞 動詞 名詞 助詞 ［躊躇 半端 に］［踊る］［自分 を］	動詞 ［許さ］	形容詞 形状詞 助動詞 助詞 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 助詞 動詞 助動詞 ］［正しい］［気 が］［し ます］
18	名詞 助詞 名詞 助詞 動詞 助詞 動詞 助動詞 助詞 ［趙雲 の］［子供 って］［出 て こ ない か 助詞 助詞 動詞 助詞 補助記号 名詞 接尾辞 助詞 なぁ と］［思う けど 、 ］［三國 志 に］［ 形容詞 名詞 名詞 副詞 名詞 動詞 助動詞 詳しい］［家人 日く］［あまり］［活躍 し て 助動詞 助動詞 助詞 助詞 名詞 助動詞 助詞 助動詞 なかっ た と の］［こと な の で］	名詞 動詞 ［期待 でき］	形容詞 形状詞 助動詞 ［なさ そう です］
19	名詞 助詞 ［贅沢 は］	動詞 形容詞 形状詞 助動詞 ［好ま なさ そう だ 助詞 補助記号 けど 、 ］	代名詞 助詞 動詞 名詞 助詞 形状詞 助動詞 ［そこ に］［ある］［材料 で］［簡単 に］ 形容詞 名詞 助詞 動詞 助詞 助詞 形容詞 ［美味しい］［もの を］［作れる の は］［凄い 助詞 動詞 助動詞 と］［思っ てる］
20	名詞 助詞 ［譽 は］	動詞 形容詞 形状詞 助動詞 ［つまら なさ そう に］	名詞 助詞 動詞 助詞 ［溜息 を］［つい た］
21		名詞 助詞 補助記号 補助記号 ［藪 は ・ ・ 動詞 助詞 動詞 入れ て もらえ］	形容詞 形状詞 助動詞 助詞 補助記号 補助記号 補助記号 ［なさ そう だ し ・ ・ ・ ］
22	名詞 名詞 助詞 副詞 形容詞 助詞 ［薔薇乙女 は］［正直］［怖く て］	動詞 ［読め］	形容詞 形状詞 助動詞 助詞 補助記号 名詞 接尾辞 助動詞 ［なさ そう だ が］［（ 薔薇 的 に 補助記号 ）］

図20 「～シナサソウ」

六二九に対し、「～してしなさそう」が七五で、「～しなさそう」が少なくはないのですが、ウェブコーパスで調べると、一七、〇〇〇に対して一一、〇〇〇くらいと、頻度差がほとんどなくなっています。もはや誤りだと切って捨てることができない状態です。ウェブコーパスでは、「くっつかなさそう」「反抗しなさそう」「許さなそう」など、動詞もいろいろなものがでてきます（図20）。

「～すべきでない」と「～しないべき」では、前者が正しいといわれています。誤りとされている「～しないべき」は、『国語研日本語ウェブコーパス』でもさほど多くは観察されませんが、分布パターンがちょっとおもしろい（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。「～を泣く」「～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、「～を泣く」が一例、「～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「～を死ぬ」

3	［嗚呼 、 ］ ［萌える べき な の か 、 ］	［萌え ない べき な の か ］	
4	［話 は ］ ［変わり ます が 、 ］ ［Q ： 会社 の ］ ［トイレ の ］ ［電気 、 ］ ［必要 の ］ ［ない ］ ［とき に ］ ［消す べき か 、 ］	［消さ ない べき か ］	
5	［話し ］ ［変わっ て ］ ［ p ］ ［入れる か ］	［入れ ない べき か ］	［迷っ てる ん だ よ なぁ w ］
6		［話さ ない べき な の か な ん です ］	
7	［話し ている ］ ［側 の ］ ［勝手 な ］ ［見解 に ］ ［より 、 ］ ［分ける べき だ ］	［分け ない べき だ と ］	［論争 する の は ］ ［ばかげ ている と ］ ［思う ］
8	［余計 な ］ ［口出し を ］	［し ない べき か ］	
9	［予防 接種 以外 の ］ ［こと に ］ ［ついて も 、 ］ ［お 医者 様 の ］ ［立場 から 、 ］ ［かなり ］ ［突っ込ん だ ］ ［意見 や ］ ［具体 例 など も ］ ［記さ れ ている て 、 ］ ［「 受ける べき か ］	［受け ない べき か 」 を ］	［検討 する に は ］ ［とても ］ ［参考 に ］ ［でき ］ ［まし た ］
10	［予備 校 に ］ ［行く べき か ］	［行か ない べき か … ］	
11	［予備 校 に ］ ［行く べき か ］	［行か ない べき か … - BIGLOBE な ん で も ］	［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）にかぎられています。『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てきました。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、量的にみるともはや逆転が生じそうな現象の場合、それでも「～しそうにない」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ているも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必要不可欠になってくるでしょう。昨今、文理融合という言葉を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でしょうか。

これで私の発表をおしまいとします。