ことばの波止場 vol.13

ISSN 2432-9207

ことばの波止場

2024 vol.13

特集
- 日本で話されていることば その危機とルーツ
- コンピュータと人間の言語
- BCCWJ 開発秘話

インタビュー
- 新しいことを！：前川喜久雄
- 今も昔も「集めて、比べる」：松本 曜

エッセイ
- 現在進行形のアイヌ語：中川 裕
- 自然言語処理「言葉の意味を表す技術」のブレークスルー：古宮嘉那子

研究室訪問
- 本の匂いの中で：高田智和

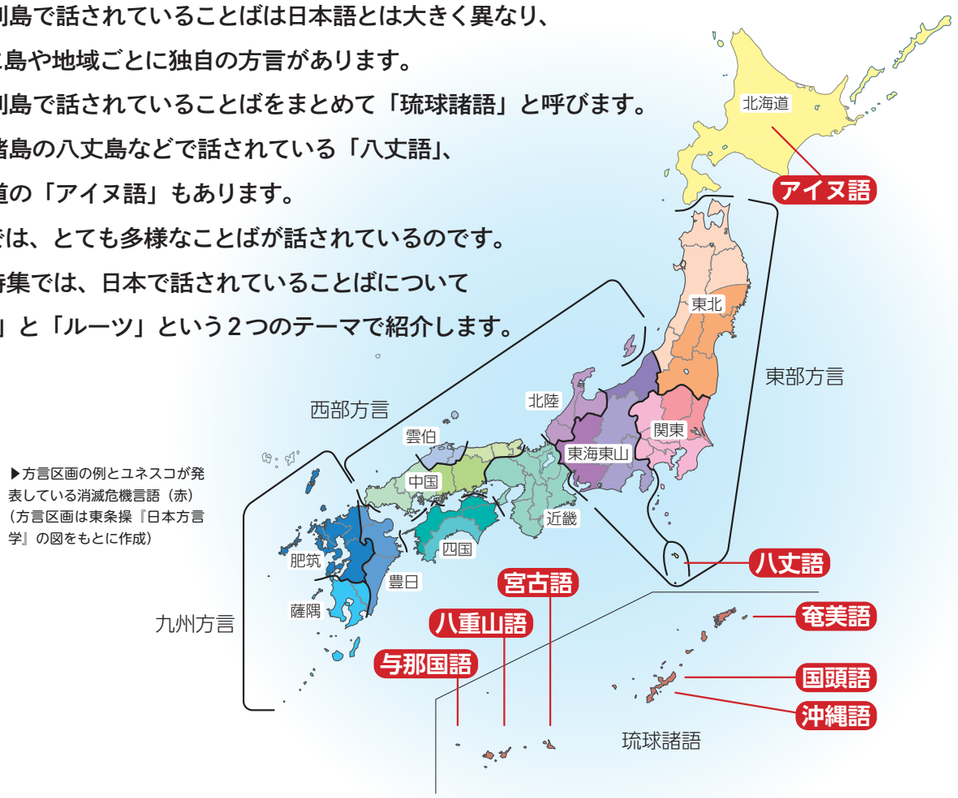
書籍紹介

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

特集

日本で話されていることば その危機とルーツ

日本で話されていることばは？ と聞かれたら、「日本語」と答える人が多いかもしれません。
日本語といっても地域によってずいぶん違う、さまざまなことばがあります。
琉球列島で話されていることばは日本語とは大きく異なり、
さらに島や地域ごとに独自の方言があります。
琉球列島で話されていることばをまとめて「琉球諸語」と呼びます。
伊豆諸島の八丈島などで話されている「八丈語」、
北海道の「アイヌ語」もあります。
日本では、とても多様なことばが話されているのです。
この特集では、日本で話されていることばについて
「危機」と「ルーツ」という2つのテーマで紹介します。



▶方言区画の例とユネスコが発表している消滅危機言語（赤）（方言区画は東条操『日本方言学』の図をもとに作成）

消滅の危機にあることばを保存・継承していくために

今、何もしなければ、なくなってしまう

世界に6,000から7,000ある言語のうち2,500が消滅の危機に瀕している――2009年、ユネスコ（国連教育科学文化機関）による発表です。その中には、日本で話されている8つのことばが含まれています。アイヌ語、八丈語、奄美語、国頭（くにがみ）語、沖縄語、宮古語、八重山語、与那国語です。実は、消滅の危機に瀕しているのはその8つだけでなく、日本各地の方言の多くも、今、何もしなければ、なくなってしまう恐れがあります。

ことばが消滅の危機に瀕しているとは、どういう状態をいうのでしょうか。皆さんがそれぞれの地域で使っていることばは、次の項目に当てはまりますか？

- 若い世代がそのことばを話していない。
- そのことばを話す場面が家庭の中だけなど限られている。
- そのことばで、またはそのことばの教育が行われていないか少ない。
- そのことばが新聞、テレビ、ラジオなどで使われていない。
- そのことばを話す人（話者）が地域の人口に対して少ない。
- 話者自身がそのことばを使うのを恥ずかしく思っている。

当てはまる項目が多いほど、そのことばは消滅の危機の度合いが高いと考えられます。

ことばの消滅がもたらすこと

ことばには、コミュニケーションの道具としての役割があります。ことばは、地域の自然環境や歴史や文化を反映して形づくられてきました。思考や感情の基盤にもなっていて、自分が自分であること、つまりアイデンティティの不可欠な要素です。自分が属するコミュニティのことばによって自分を表現できて、ほかの人とコミュニケーションできることは、私たちにとってとても大切なのです。

ことばの消滅は、アイデンティティや文化、歴史の喪失につながります。祖先に当たることばの姿を知る手がかりも失われてしまいます。私たちは、普段話している耳慣れたことばとは違うことばに接すると、知的好奇心が刺激され、学ぶことで知識が蓄積されていきます。ことばの多様性が減ると、そうした機会も失われます。

ことばを保存し継承する取り組み

国語研では、日本各地のことばについて、語彙の収集や、話している場面の録音・録画、文法の記述、辞書の作成などによる「記録保存」を進めています。しかし、ことばの記録が大量に保存されていても、そのことばを話す人がいなくなったら、ことばは消滅してしまいます。ことばが親から子、孫へ、世代を超えて継承される「継承保存」も不可欠です。国語研で行っている記録保存と継承保存の取り組みを右に紹介します。

沖永良部（おきのえらぶ）島の例で明らかなように、ことばの保存・継承は、研究者だけではできず、そのことばを話す地域の人々との協働が不可欠です。また、絵本のクラウドファンディングでは、その地域以外の人たちからも多くの支援をいただきました。ことばの消滅は、遠い南の島の問題ではなく、自分にも関係していることとして受け取ってくださったのだと思います。

今、何もしなければ、なくなってしまうことばがあります。皆さんも、自分の地域のことば、さまざまな地域のことばについて、考えてみませんか？

ことばを集めて保存・公開しています

・危機言語データベース

消滅の危機にある日本各地の言語・方言のさまざまな語彙を音声付きで公開しています。また、地域の行事や昔話を方言で語っている音声を、書き起こしと共通語訳を付けて公開しています。
https://kikigengo.ninjal.ac.jp/

・ことばのミュージアム

日本各地で話されているいろいろなことばを、見たり、聞いたり、学んだりできます。
https://museum.ninjal.ac.jp/

島のことばで伝承を語る絵本を出版しています

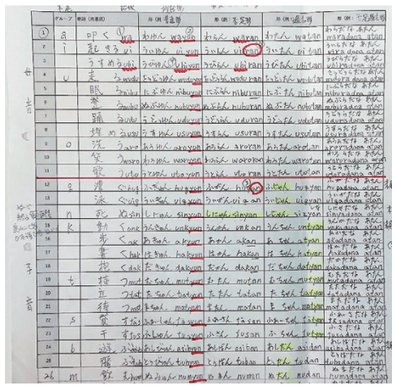
「言語復興の港」プロジェクトでは、沖永良部島、多良間島、竹富島、与那国島それぞれに伝わる昔話を題材とした4冊の絵本を出版しました。琉球のことばで語られる昔話を絵本の形で保存し、同時に、絵本を楽しみながら島のことばを聞く機会を増やし学ぶことを目指した取り組みです。絵本の制作・出版に当たり、クラウドファンディングで多くの方からご支援いただきました。
https://readyfor.jp/projects/minato

▲多良間島に伝わる昔話『カンナマルクールクの神』と竹富島に伝わる昔話『星砂の話』の地域のことば絵本（絵：山本史、ひつじ書房）。朗読音声と詳しいことばの解説付き。

地域の人々による、ことばの保存・継承へ

奄美群島沖永良部島において「しまむに」（島のことば）の公民館講座を2019年ごろから行っています。琉球列島で話されていることばは、島ごと、集落ごとに異なり、それらを少ない研究者で記録することは不可能です。ことばが消滅していくスピードにも追い付きません。地域の中でことばを記録する人を育成することが、この講座の一番の目的です。講座をきっかけに、自主的に記録活動を始めたり、言語学の研究手法で分析したり、方言の継承について議論したりする人たちが増えています。こうした地域の人々の主体的な活動があれば、ことばを保存・継承していくことが十分に可能です。

▲沖永良部島での公民館講座の様子



▶公民館講座の受講者による後（ご）接（せつ）方言の動詞活用の調査と分析（沖良子さん作成）

日本語と琉球諸語のルーツをひもとく

日本語はいつごろ、どこから来たのでしょうか？
日本列島に入ってきた言語は、
どのような変遷をたどったのでしょうか？
また、それらはどのようにして分かるのでしょうか？
日本語の歴史を研究しているトマ・ペラールさんと
五十嵐陽介さんに聞きました。

トマ・ペラール
東アジア言語研究所
（フランス国立科学研究センター）研究員

五十嵐陽介
国立国語研究所 教授

日本語と琉球諸語は同じ起源を持つ

――お二人は、なぜ日本語の歴史の研究を始められたのですか。

ペラール：もともと日本の伝統文化に興味があって、人類学や民俗学をやろうと思っていました。途中で、言語の研究の面白さに気付いたのです。そして、日本語がどのように変化してきたのか、今の日本のことばがどのように成立したかについて研究を始めました。すると、日本語の歴史を研究するには琉球諸語のデータが絶対に必要だと分かり、琉球諸語の研究を始めたのです。五十嵐さんと初めて会ったのは、フィールドワークで訪れた宮古島でしたね。五十嵐さんは、まだ日本語の歴史については手を付けていなかったと思います。日本語の歴史の研究に深く入り込むようになったのは、僕の影響があるかもしれませんね。それは大変うれしいことです。

五十嵐：言語の音声に興味があり、学生のときはロシア語、その後は日本語の諸方言の音声の研究をしていました。ある学会で、後に国語研の所長を務める田窪行則先生から「宮古島の池間方言の研究をしているのだが、アクセントがどうにも分からない。君は耳が良さそうだから、一緒に来てくれないか」と声をかけられたのです。それで宮古島を訪れたときに、宿が同じだったトマさんと知り合ったのでしたね。トマさんと琉球諸語の話をしていると、必ず歴史の話が出てきます。そして池間方言のアクセントを理解するのにも歴史を知らないといけないことに気付き、歴史言語学に触れてみたら面白かったのです。

――琉球方言ではなく、琉球諸語というのですね。

ペラール：琉球列島で話されていることばは、長い間、日本語の方言と見なされてきました。しかし日本語と通じないほど違うため、現在では日本語とは異なる言語とみなされています。また、琉球列島の中でも島や地域ごとに互いに通じないことばがあることから、まとめて「琉球諸語」と呼んでいます。

――日本語の歴史を知るために、なぜ琉球諸語の研究が必要なのでしょうか。

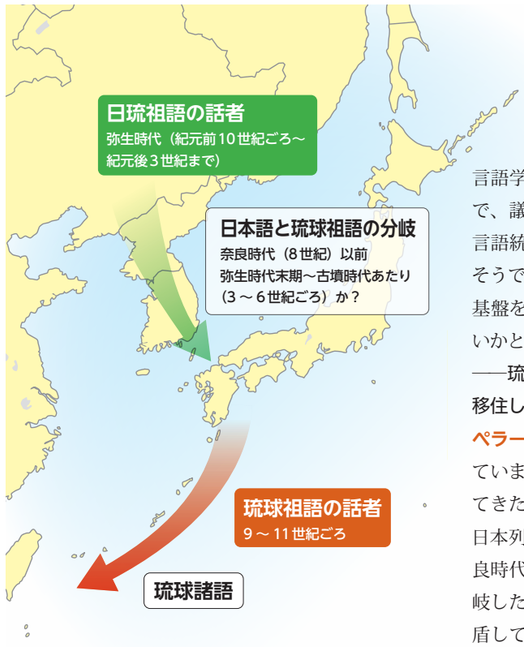
五十嵐：初めて宮古島を訪れたとき、食堂で隣に座ったおばあさん２人が話をしていたのですが、内容をまったく理解できませんでした。しかし、琉球諸語の話者に「これは何と言いますか」と1つずつ聞いていくと、意味と音の両方で日本語の単語と類似するものがたくさんあるのです。例えば、琉球諸語で「パ」「ペ」「ポ」で始まる単語は、日本語では「は」「へ」「ほ」で始まるというように、対応に規則性が見られます。アクセントにも規則的な対応があります。それらは比較言語学の教科書に書いてあるとおりの対応関係で、琉球諸語と日本語は同じ起源を持っている同系の言語であると判断できます。琉球諸語と日本語の共通祖先の言語を「日琉祖語」といいます。

ペラール：同じ系統の言語を比較することで、それらが分かれる以前の共通祖先の言語を推定できます。同系の言語がそれぞれパズルのピースを持っていて、それらを並べて共通祖先の言語の体系を再現していく、というイメージです。これを言語学では「再建」といいます。日本語の歴史を知るには日琉祖語がどういう言語であったかを再建することが不可欠で、それには琉球諸語を研究してパズルのピースを集める必要があるのです。

日本語と琉球祖語の分岐は？

――琉球諸語につながる言語と日本語は、いつ分岐したのでしょうか。

ペラール：奈良時代（8世紀）以前ということは、研究者の意見が一致しています。最も古い日本語の文献は奈良時代のものです。文献から、奈良時代にどのような音の区別があったかが分かります。さらに、文献にある言語の状態からそれ以前の言語の状態を理論的に推定する「内的再建」という歴史比較言語学の手法を使い、奈良時代以前には音の区別がもう少し多かったことが推定されています。奈良時代の文献ではすでに失われてしまったがそれ以前には存在していたと考えられる音の区別が、琉球諸語には見られます。ということは、琉球諸語の共通祖先に当たる「琉球祖語」と日本語が分かれたのは、日本語でその音の区別が失われる前、つまり奈良時代以前と考えられるのです。



――分岐時期の細かい推定は難しいのでしょうか。

ペラール：琉球諸語と日本語は、稲に関連した単語を共有しています。稲作が伝わる以前に分岐していたらそうならないので、稲作が伝わった弥生時代（紀元前10世紀ごろ～紀元後3世紀まで）以降と考えるのが自然です。ここからは研究を進めているところで、まだ私の個人的な推測なのですが、弥生時代末期から古墳時代あたり（3～6世紀ごろ）に分岐したのではないかと考えています。各地にクニと呼ばれる政治的なまとまりが生まれるなど、日本列島の社会が大きく変わった時代です。社会の変動は言語の変動にも影響すると予想されます。

五十嵐：弥生時代末期から古墳時代という分岐年代は、うなずけます。しかし、言語学的な証拠は何も得られていないので、議論が必要でしょう。トマさんは、言語統計学的に分析する方法を開発中だそうですね。それを適用したら言語学に基盤を置いた年代推定ができるのではないかと期待しています。

――琉球祖語の話者は、いつ琉球列島へ移住したのでしょうか。

ペラール：9～11世紀ごろだと考えられています。琉球諸語には日本語から入ってきた多くの漢語が見られます。漢語が日本列島に入ってきて普及したのは、奈良時代以降です。琉球祖語が日本語と分岐したのは奈良時代以前ということと矛盾しているのでは？と思うかもしれません。琉球祖語の話者は九州にしばらく暮らしていて、中国語から漢語を借用した日本語から、その漢語を借用して琉球列島に移住したと考えられます。言語として分岐した後、時間がたってから、その話し手が移動したのです。9～11世紀ごろというのは、琉球列島でグスク文化が発展し、狩猟採集文化から農耕文化へと変化した時代です。琉球祖語の話者が琉球列島に移住してグスク文化の担い手になり、そこで話されていた言語を置き換えていったと考えられています。

日本語の歴史をめぐる課題

――日琉祖語の話者は、いつ日本列島へ来たのでしょうか。

ペラール：弥生時代から日琉祖語の話者が日本列島に移住し、弥生文化の担い手になったと考えられます。考古学、遺伝学の研究から総合的に見て、日琉祖語の話者は東北アジア、中国北部、朝鮮半島あたりから来たと考えられます。古代の朝鮮半島の歴史を記録した文献に日本語と音と意味がとても似た地名が見られることから、日本語と系統関係にあった言語が古代の朝鮮半島で話されていた可能性が高いです。それは大陸倭語（わご）とも呼ばれますが、その後、別の言語に置き換えられて消滅しています。

――日本語の歴史の研究には、どのような課題がありますか。

五十嵐：日本語の系統に関しては、荒唐無稽な説が飛び交う時代が長く続いていました。しかし、現在話されている言語で日本語と同系であることが現時点で確認できる言語は、琉球諸語だけであることが、比較言語学から明らかになっています。しかし最近また日本語の系統について新しい説が出され、議論が沸き起こっています。これは、日琉祖語をしっかりした根拠をもって再建して、日琉祖語とはこういう言語なんだというところまで提示することができていないからだと思うのです。日本語の諸方言、琉球諸語の厳密な比較を通じて、日琉祖語の再建をしっかりやる必要があります。

ペラール：そのためにはデータが命です。特に、本土の諸方言、琉球諸語のデータが不可欠です。それらは消滅の危機にあるので、急がなければなりません。

――日本語の歴史を研究する面白さとは？ また、今後どのように研究を進めていこうとお考えですか。

五十嵐：多様性ですかね。なので、飽きることがない。私は、比較言語学、音声学、方言、どの分野でも「五十嵐に聞け」と言われるような一番の存在ではありません。しかしだからこそ、いろいろな分野からの話を総合して、1つの話をつくることができるかもしれない、できたらいいな、と思っています。

ペラール：日本語は多様性がありつつ、日本語と琉球諸語の間には比較言語学の教科書どおりの規則性が見られるので、理論的な研究もやりやすい、という面白さがあります。今まで海外と日本国内の研究をきちんと把握し、両方の研究者との間の架け橋となれるように努めてきました。これからも日琉諸語の歴史比較的研究という分野が世界レベルで発展するようにしたいと思います。

エッセイ

現在進行形のアイヌ語

中川 裕

なかがわ・ひろし。1955年生まれ。千葉大学名誉教授。専門はアイヌ語学・アイヌ文学。著書に『アイヌ語千歳方言辞典』（1995、草風館）、『アイヌの物語世界』（1997、平凡社）、『語り合うことばの力』（2010、岩波書店）、『ニューエクスプレス アイヌ語』（2013、白水社）、『アイヌ文化で読み解く「ゴールデンカムイ」』（2019、集英社）などがある。

野田サトルさんの漫画『ゴールデンカムイ』は、2014年から集英社『週刊ヤングジャンプ』誌上で連載が開始され2022年に完結。中身は一言で言えば20世紀初頭の北海道を舞台にした冒険活劇漫画ですが、綿密な時代考証の上で展開する奇想天外なストーリーと、抜群の画力・演出力によって、コミックス全31巻累計で2500万部を超える大ヒット作品になり、アニメ化もされ、実写化も着々と進んでいます。

そしてこの漫画は、魅力的で個性的なアイヌを主要人物として多数登場させ、彼らの文化や歴史を真っ向から物語の中に組み込んだことで、現実のアイヌを取り巻く社会的状況にも影響を及ぼすほどの作品となりました。これまでアイヌという存在にまったく興味のなかった人たちに、関心を持たせるきっかけをつくったのです。

また2020年には国立アイヌ民族博物館を含む民族共生象徴空間、愛称「ウポポイ*1」が北海道白老（しらおい）町に開設されました。古くからこのあたりはポロトコタン*2の名前で知られ、アイヌの伝統文化を紹介する北海道観光の中心地の一つでした。その地に新たに国立の施設が建てられたことで、全国からアイヌ文化に関心のある人たちが集まるようになりました。

こうした流れに伴い、アイヌ関連の著作物も画期的に増えています。特に、すでに絶版になっているような書籍が文庫化といった形で再刊されることが多くなり、アイヌ関連の話題がメディアで取り上げられることも日常的になってきました。

ただその中にあって、アイヌ自身のアイヌ語を巡る現在の活動はあまり注目を集めているとは言えません。

はっきり言って、アイヌ語の母語話者と見なせる人はもういません。母語話者を基準にする限り、アイヌ語はすでに「死語」というべき状況にあります。話者のいない言語に現在の活動もなにもないものだ、と思うのは当然でしょう。

しかし、アイヌ語を後世に残していこうという活動は、むしろ盛んになってきています。残すというのは、音声アーカイブをつくったり、辞書を編纂したりすることではありません。それは単に維持・復興のための資源を整備するだけのことであり、言葉を残すというのは、今生きている人がそれを使っていくことです。

*1 **ウポポイ**
アイヌ語で「（おおぜいで）歌うこと」の意味

*2 **ポロトコタン**
アイヌ語で「大きな湖の集落」の意味

漫画『ゴールデンカムイ』のアイヌ語監修を務めた中川裕氏によるアイヌ文化の入門書
絵：野田サトル／集英社

公益財団法人アイヌ民族文化財団ホームページで「アイヌ語動画講座」公開中！

このほかに紙人形劇、アイヌ語で動物紹介など、さまざまなコーナーがあります。

（出典：公益財団法人アイヌ民族文化財団「アイヌ語動画講座」）

アイヌ語自然講座

アイヌ語料理講座

ウポポイのアイヌ語サイン紹介

そうした活動の一つの例は、ウポポイの運営団体である公益財団法人アイヌ民族文化財団のホームページ（https://www.ff-ainu.or.jp/）にある「アイヌ語動画講座」というコーナーで見ることができます。これは2020年から始まった企画で、出演者はほぼ全員が、アイヌとしてのアイデンティティを持っている人か、ウポポイを含めた関連施設で働いている人。それぞれが好きなテーマで、とにかくアイヌ語を使って何かをするという動画を作成して公開しています。

実は私はその講座の企画編集委員の一人で、担当しているのは「自然講座」「料理講座」「ウポポイのアイヌ語サイン紹介」の三つです。

「自然講座」は、苫小牧（とまこまい）市の作田悟さんを講師として、樹木や山菜の見分け方や利用法、あるいは昔の生活などを教えてもらうコーナーです。進行役はウポポイ職員の川上将史さん。

「料理講座」は様似（さまに）町のアイヌ文化伝承活動の牽引者である熊谷カネさんに、様似地方の伝統料理を教えてもらうというコーナーで、進行役は熊谷さんの兄の孫に当たる工芸家の岡本朋也さん。熊谷さんと岡本さんは親戚同士ですが、この動画で初めてアイヌ文化に関わる活動を一緒に行ったということです。

ウポポイではほとんどの施設や設備にアイヌ語の名前がついていますが、それを紹介するのが「アイヌ語サイン紹介」。解説するのは苫小牧市出身でアイヌ民族博物館学芸員のヤンチャキさん。ウポポイの職員には全員、ポンレ「小さい名」と呼ばれるアイヌ語のあだ名がつけられていて、ヤンチャキは彼女のポンレです。

私は企画編集委員を名乗っていますが、やることはただ「作田さんになんかやってもらおう」とか「熊谷さんに料理をつくってもらおう」というようなことを提案するだけで、あとは全部出演者任せ。それでもちゃんとした動画が出来上がるのは、演者がみんななんとかして自分たちの先祖の言葉や文化をほかの人たちに伝えていこうという、強い熱意があるからです。

このほかにも動画講座には、紙人形劇や寸劇、子供たちとのアイヌ語学習や、ウポポイ職員へのアイヌ語インタビューなどいろいろなコーナーがあって、おおぜいの人がそれぞれのやり方でアイヌ語を使った活動を試みています。

アイヌ語の母語話者はすでにいません。しかし、このようにいろいろな人がアイヌ語を使う活動を日々行っていく限り、アイヌ語は消滅しないでしょう。アイヌ語は現在進行形なのです。

このアイヌ民族文化財団のホームページには、動画講座以外にもさまざまなコーナーがあり、一般にはあまり注目されていませんが、世界で一番アイヌ文化関連のコンテンツが充実したサイトですので、関心のある方は一度ご覧になっていただくとよいと思います。

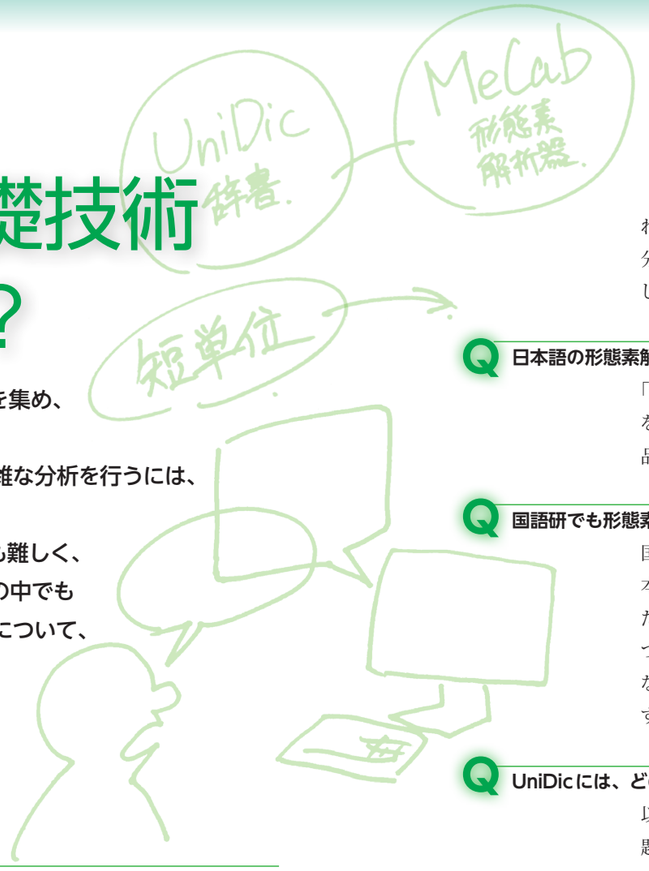
特集

コンピュータと人間の言語

自然言語処理の基礎技術「形態素解析」とは？

小木曽智信
国立国語研究所 研究系 教授

ことばの研究では、たくさんのことばを集め、それをさまざまな観点から分析します。多種多様なことばを大量に蓄積し、複雑な分析を行うには、コンピュータが不可欠です。しかし、ことばをコンピュータで扱うのはとても難しく、さまざまな技術が必要になります。その中でも基礎となる技術である「形態素解析（けいたいそかいせき）」について、小木曽智信教授に聞きました。



Q 小木曽さんは、どのような研究をされているのですか？

専門は日本語学と自然言語処理です。

Q 自然言語処理とは？

自然言語処理とは、私たちが普段使っていることばをコンピュータで扱うための技術のことです。コンピュータで用いるプログラミング言語に対して、人が用いる言語を自然言語と呼びます。日本語の自然言語処理にはさまざまな技術が用いられますが、中でも基礎となる技術が「形態素解析」です。

Q 形態素解析とは、どういう技術ですか？

形態素解析とは、文を単語に区切り、品詞や読みなどの情報を付ける技術をいいます。言語を研究するとき、文・単語・文字・音韻など、さまざまな単位で分析しますが、それらの中で基本になる単位が、単語です。

例えば、あるテキストの中の動詞Aについて、これがどのような言葉か調べたいとしましょう。そのためには、動詞Aが出てくる回数だけでなく、対象とするテキストに単語がいくつあって、どんな単語がどれだけ一緒に使われているかを知る必要があります。それらが分かって初めて、動詞Aを全体の中に位置付けて、統計的な分析ができるのです。

従って形態素解析は、言語研究において最初に必要な基礎的な技術です。しかし日本語は、単語に区切るということが、とても難しいのです。

Q 具体的にはどういう難しさがあるのでしょうか？

「私は国立国語研究所に勤めています」という文は、何語に分けられるでしょうか？

私 は 国立国語研究所 に 勤め て い ます

このように区切ると、8語です。

でも、**「国立国語研究所」**と長いものを1単語としてよいのか？と思う人もいるでしょう。

国立 国語研究所
国立 国語 研究所
国立 国語 研究 所

と区切ることもできます。

英語では、単語がスペースで区切られ分かち書きされるので、「これが1単語だ」と分かります。日本語は分かち書きされないので、1単語が明確ではないのです。しかし正しく区切っておかないと、「京都」と検索したら「東京都」、「カレー」と検索したら「エスカレーター」が出てしまったり、係り受け解析や構文解析が正しくできなかったりして、言語研究に使えません。

Q 日本語の形態素解析をどのように行うのですか？

「形態素解析器」と呼ばれるコンピュータプログラムを使います。文を入力すると、単語に分けて、さらに品詞や活用形などを判別した結果が出力されます。形態素解析器は、形態素解析用の辞書と組み合わせて使います。形態素解析器も解析用の辞書も、さまざまなものが開発されています。

Q 国語研でも形態素解析用の辞書などを開発しているのでしょうか？

国語研では2000年ごろから、1億語からなる「現代日本語書き言葉均衡コーパス（BCCWJ）」の構築に向けた準備を始めました。コーパスとは、ことばを大量かつ体系的に収集し、研究用の情報を付与してさまざまな検索ができるようにした、ことばのデータベースです。BCCWJのために、形態素解析を精度よく言語研究に適した形で行えるように、新しい辞書を国語研が中心となって開発しました。それがUniDic（ユニディック）です。

UniDicは、MeCab（めかぶ）という形態素解析器で利用できるようになっています。MeCabは、工藤拓氏によって開発されたソフトウェアで、高速かつ高精度であることから、現在最もよく利用されている形態素解析器です。

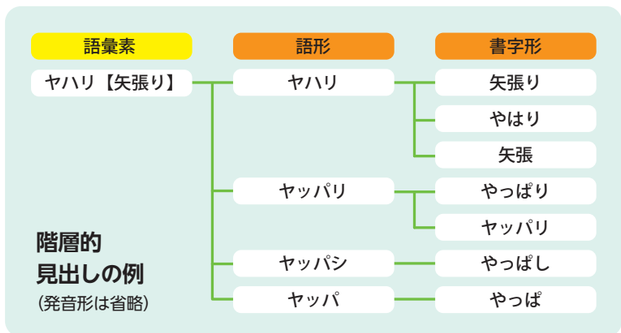
Q UniDicには、どのような特長があるのでしょうか？

以前からあった形態素解析用の辞書には、いくつか問題がありました。その一つが、単語の区切り方です。「国立国語研究所」の例で示したように、1単語を長く区切ることも、短く区切ることもできます。それまでの形態素解析用の辞書では、単語の区切り方が統一されていなかったため、統計的な分析が正確にできませんでした。

UniDicの特長の一つは、単語の区切り方を、国語研で決めた「短単位」というルールで統一していることです。短単位は、区切りの基準が分かりやすく、揺れが少なくなります。そのため、これを基盤としてさまざまな研究を行うことができるのです。

UniDicのもう一つの特長は、見出しが階層化されていることです。

Q 見出しが階層化されているとは？



UniDicでは見出しを、語彙素、語形、書字形、発音形と4つのレベルの階層構造にしています。皆さんが使う辞書の見出しに相当するのが、語彙素です。「ヤハリ」という単語は、「ヤッパリ」「ヤッパシ」「ヤッパ」と語形が揺れたり、「矢張り」「やはり」「矢張」と書字形が揺れたりします。以前の辞書は、複数ある語形、書字形を別々に扱っていました。これではそれぞれ別の見出しになるので、一括した検索や集計ができません。

UniDicでは語形や書字形の揺れにかかわらず、同一の見出しとしてまとめられているので、目的に応じて利用することが可能です。

Q UniDicを使って形態素解析を行うには、専門知識が必要ですか？

MeCabもUniDicも公開されているので、パソコンにインストールすれば誰でも形態素解析ができます。しかし、黒い画面にコマンドを打って操作するので、慣れていない人にはハードルが高いかもしれません。

そこで、もっと簡単に使えるように開発したのが、「Web茶まめ」です。Web茶まめは、インストールが不要でオンラインで使え、専門知識がなくても形態素解析ができます。

Q Web茶まめの使い方を教えてください。

正岡子規の俳句「**柿くへば鐘が鳴るなり法隆寺**」を形態素解析してみましょう。

Web茶まめの使い方

① 「Web 茶まめ」のサイトにアクセス
https://chamame.ninjal.ac.jp/

② 解析したいテキストを入力
テキストファイルをアップロードして解析することもできます。

柿くへば鐘が鳴るなり法隆寺

③ 使用する辞書を選択
この例では「近代文語」を選択します。

④ 出力したい項目を選択
一般的な項目が初期設定されています。

⑤ 出力形式を選択

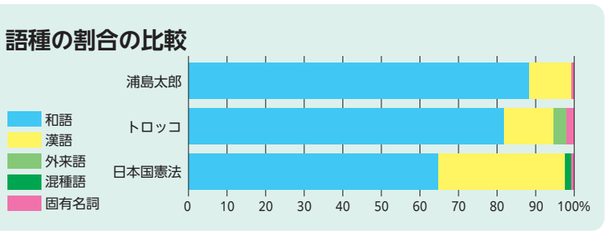
⑥ 「解析する」ボタンをクリック

⑦ 解析結果

辞書	文境界書字形（＝表層形）	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	語形(基本形)
近代文語B	柿	柿	カキ	名詞-普通名詞-一般			カキ	カキ	和	柿	カキ
近代文語I	くへ	食う	クウ	動詞-一般	文語四段-ハ行	已然形-一般	クエ	クヘ	和	くふ	クウ
近代文語I	ば	ば	バ	助詞-接続助詞			バ	バ	和	ば	バ
近代文語I	鐘	鐘	カネ	名詞-普通名詞-一般			カネ	カネ	和	鐘	カネ
近代文語I	が	が	ガ	助詞-格助詞			ガ	ガ	和	が	ガ
近代文語I	鳴る	鳴る	ナル	動詞-一般	文語四段-ラ行	連体形-一般	ナル	ナル	和	鳴る	ナル
近代文語I	なり	なり-断定	ナリ	助動詞	文語助動詞-ナリ-断定	終止形-一般	ナリ	ナリ	和	なり	ナリ
近代文語I	法隆	法隆	ホウリュウ	名詞-固有名詞-一般			ホーリュー	ホウリュウ	固	法隆	ホウリュウ
近代文語I	寺	寺	ジ	接尾辞-名詞的-一般			ジ	ジ	漢	寺	ジ

Q おすすめの解析を教えてください。

和語・漢語・外来語・固有名詞という語種の割合に注目すると、そのテキストの特徴が見えてきて面白いですよ。語種を自動で判別できることも、UniDicの特長です。



文部省唱歌の『浦島太郎』、芥川龍之介『トロッコ』の冒頭部分、日本国憲法 前文を、それぞれWeb茶まめで形態素解析を行い、語種の割合を比べると、左のようになりました。解析結果はExcel形式で出力できるので、グラフもつくれます。

好きなアーティストの曲の歌詞や、好きな作家の作品について、Web茶まめで形態素解析を行ってみては、いかがでしょうか。ほかのアーティストや作家のテキストと比べることで、気付いていなかった特徴が見えてくるかもしれません。

Q UniDicや形態素解析について、課題や今後の計画はありますか？

Web茶まめの「辞書選択」でお気付きかもしれませんが、UniDicは1種類ではありません。日本語といっても時代・地域によって違うので、それぞれの時代・地域に合わせた形態素解析用の辞書が必要です。古文用を少しずつ開発してきて、現在UniDicは13種類あり、奈良時代から現代の文章まで解析できるようになっています。関西方言用も間もなく公開できる予定です。ほかの方言用の辞書の開発にも着手しています。現代書き言葉、現代話し言葉の辞書についても、新しい見出しの追加などメンテナンスが欠かせません。

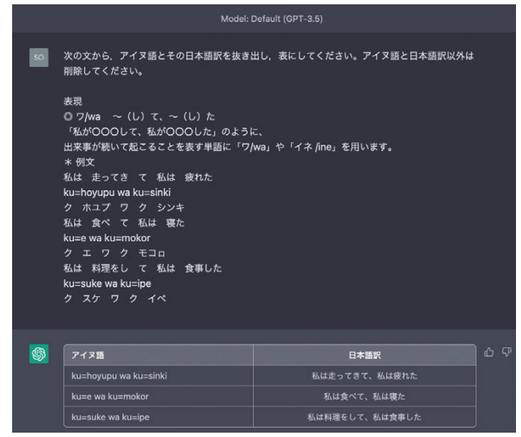
UniDicは、日本語研究だけでなく、文学などほかの分野や産業界でも広く使っていただいています。とてもうれしいですね。

形態素解析という技術は、日本語をコンピュータで処理し、分析するときに基礎となる技術なので、これからも重要な役割を果たしていくことでしょう。

ChatGPTを活用してコーパスを構築する

ことばをコンピュータで扱うといえば、ChatGPTが最近話題です。ChatGPTは、アメリカのOpenAI社が開発した対話型生成AIで、ブラウザやアプリで質問や指示を文章で入力すると、回答が返ってきます。専門知識がなくても手軽に使え、さまざまな分野や目的、場面に対応できる能力を持っています。コーパス構築にもChatGPTが使われ始めています。どう使う？ 利点は？ 宮川 創助教に聞きました。

テキストデータ出典：「令和４年度「アイヌ語ラジオ講座」テキスト Vol.3」（公益財団法人アイヌ民族文化財団）



宮川 創
国立国語研究所
研究系 助教

Q 宮川さんは、ChatGPTをよく使っているのですか？

常識的・倫理的に問題なさそうな範囲で、文章の要約や翻訳の参考に使ったり、また自分が書いた文章を校正したり文体を変えたり、さらにはプログラミングやデータの整理などに使ったりしてみて、その結果を参考にすることがあります。GPT-4という非常に優秀な次世代大規模言語モデルを利用できる有料版のChatGPT Plusを使っています。

Q ChatGPTをコーパス構築に活用しているそうですね。どのようにChatGPTを使うのですか？

コーパスとは、ラテン語のcorpus「体」が語源ですが、言語学では言語のテキストデータセットを指します。私は最近、自身の機械翻訳の研究用にアイヌ語の対訳テキストのデータを収集しました。使用したのは、国立アイヌ民族博物館のアイヌ語アーカイブや、ほか複数箇所のコーパス、アイヌ語の教科書です。

アイヌ語アーカイブなどの既存のコーパスは、テキストがデータ化され、アイヌ語とその日本語訳が構造化されているので、そのままコンピュータで扱うことができます。

アイヌ語の教科書は印刷物なので、まずOCR（光学的文字認識）にかけてPDF形式にして、そこからテキストをコピーして抽出します。しかし、そのデータをプレーンテキストとして別のソフトウェアに貼り付けると、アイヌ語と日本語訳の対応がバラバラになってしまいます。そのため、アイヌ語とそれに対する日本語訳というようにテキストデータを整理して、表などコンピュータで扱いやすい形式にする必要があります。その構造化にChatGPTを使いました。

ChatGPTで「次の文から、アイヌ語とその日本語訳を抜き出し、表にしてください。アイヌ語と日本語訳以外は削除してください」と指示をして、その下にアイヌ語の教科書をOCRにかけて抽出したテキストデータを入力します。すると、表になって返ってきます（右上写真）。これらのデータを使い、機械翻訳の学習用データセットを構築しました。

Q ChatGPTをコーパス構築に使うことには、どういう利点があるのでしょうか？

この例では教科書をOCRにかけて抽出したテキストデータを使いましたが、私が研究しているコプト語では、対象が中世の写本など手書きの場合もあります。手書きテキスト認識（Handwritten Text Recognition：HTR）のソフトウェアには機械学習可能なもの（TranskribusやeScriptoriumなど）があって文字認識の精度を手軽に向上させることができるので、手書きの資料をテキストデータにすることは近年どんどん簡単になってきています。しかし、OCRやHTRで得られたテキストデータから原文と現代語訳のペアを抽出したり、分析しやすいように表などにしたりして構造化するには、これまでは手作業で膨大な時間がかかっていました。ChatGPTは、そうしたバラバラのデータも柔軟に構造化でき、コーパス構築の効率が大幅に向上します。

アイヌ語の例を紹介しましたが、日本には消滅の危機にある言語・方言がたくさんあります。ChatGPTなどの最新技術を活用することで、コーパス構築を効率化していきたいと考えています。

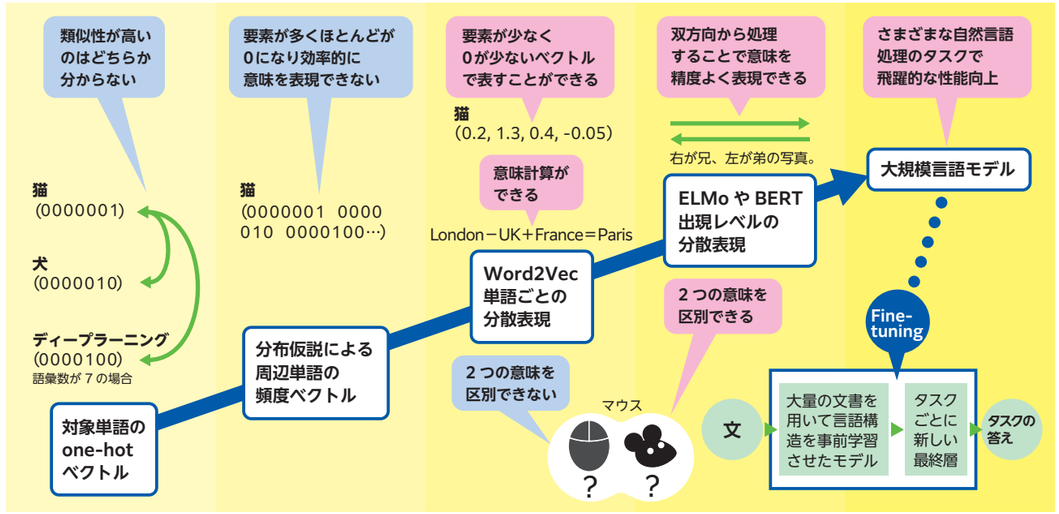
エッセイ

自然言語処理「言葉の意味を表す技術」のブレークスルー

古宮嘉那子

こみや · かなこ。東京農工大学大学院工学研究院先端情報科学部門准教授。東京農工大学卒業。同大学大学院で修士号および博士号（工学）を取得。東京工業大学博士研究員、東京農工大学特任助教、茨城大学講師を経て、2021年より現職。専門は自然言語処理。著書に『機械学習教本』（森北出版）、『文書分類からはじめる自然言語処理入門』（科学情報出版）がある。

自然言語処理による意味の表現形式の変遷



ChatGPT が登場してから１年余り。AI が質問に流ちょうな文章で答えてくれることが当たり前になりました。このようなことが可能になったのは、自然言語処理、特に、言葉の意味をコンピュータ上に表現する技術が急速に進歩してきたからにほかなりません。

自然言語処理とは、AI 研究の一分野で、日常私たちが使っている言語をコンピュータで扱う研究分野です。この分野において、ディープラーニングによる技術革新が起きたのは 2013 年から 2018 年のほんの 5、6 年のことでした。これは、言葉の意味を表す技術のブレークスルーだったと言えます。

●

自然言語処理では、主に言語学における分布意味論の「分布仮説」によって単語の意味を取り扱っています。分布仮説は、簡単に言えば「文脈の似た単語・句・文などは意味が似ている」ということ。ある単語を表す際、「one-hot ベクトル」と呼ばれる方法が用いられます。これは、語彙数分の要素を持つベクトル（数値の配列）を用意し、その単語に該当する要素だけを 1、その他の全ての要素を 0 とすることで単語を表したものです。ところが、one-hot ベクトルだけでは、単語の意味的な類似性を直接的には表現できません。例えば、「猫」と「犬」の意味的な類似性は「猫」と「ディープラーニング」の類似性より高そうですが、ベクトルの要素として単に「同じ単語であるかどうか」で考えると、共に「互いに異なっている」ということまでしか表せません。

そこで、分布仮説の出番です。意味を表したい単語の周辺の単語の頻度情報を「文脈」としてとらえるために、例えば周囲の one-hot ベクトルをいくつか連結したものを１つのベクトルとして単語を表すことにします。しかし、ベクトルの要素数がとても大きくなるわりに、周辺文脈に出現した単語以外の要素の値は 0 になるため、ほとんどの要素が 0 になってしまい、効率的に意味を表現できないという問題があります。

●

この問題に解を与えたのが、2013 年に Mikolov らの提唱した「Word2Vec」です。Word2Vec は、0 が多く疎であるベクトルの要素数（次元数）を圧縮する工夫を加えたものです。この際、ベクトルの要素数はもともとの語彙数よりはるかに小さくなります（語彙数は万単位ですが、Word2Vec のベクトルの次元数は通常たかだか 200 から 300 程度です）。

ここにディープラーニングの技術が用いられました。ディープラーニングは、次元数の大きい情報を圧縮することが非常に得意です。このように単語を低次元で 0 の要素が少ない密なベクトルで表すことを、「単語の分散表現」と呼びます。さらに、Word2Vec による分散表現には、言語学における「加法構成性」があるという特徴があります。これは、例えば「king − man + woman = queen」などの意味計算が可能だということです。

Word2Vec に代表される、語ごとの分散表現では、単語の見た目が同じであれば、文脈が異なっても同じベクトルとして表されます。これで困るのは、複数の意味を持つ単語です。例えば、「マウス」はコンピュータ・デバイスであると同時に動物の名前でもあります。でも、Word2Vec ではこれらを同時に１つのベクトルとして表すため、２つの意味を区別できません。

●

2018 年に Peters らによって提唱された「ELMo」（Embeddings from Language Models）は、それを可能にしました。見た目が同じ単語でも、文脈が異なれば別の意味ベクトルとして表せるようになったのです。

さらに 2018 年に、ChatGPT の前身である初代 GPT（Generative Pre-trained Transformer）が発表されました。これで、Transformer という高速計算が可能なネットワーク構造を持つモデルで事前に大量の文書を用いて言語構造を学習させておき、タスクごとにモデルの Fine-tuning（微調整）を行って利用する「大規模言語モデル」（Large Language Model：LLM）の枠組みが確立しました。現在の自然言語処理では、この大規模言語モデルを利用する手法が主流です。

さらに、その年の秋に発表された「BERT」（Bidirectional Encoder Representation from Transformer）は、文脈を前方向からも後ろ方向からも処理する双方向の Transformer の利用と言語モデルの工夫で、さまざまな自然言語処理のタスクで飛躍的な性能向上を成し遂げました。また、BERT による分散表現は、日本語学や言語学の観点からも有益です。

●

筆者は学生時代から文中の単語を辞書の意味項目に分類する研究を行ってきましたが、上記のような一連の技術の発展に伴って精度が格段に上がってきています。例えば筆者の研究室では、日本語の BERT を使って「現代日本語書き言葉均衡コーパス（BCCWJ）」に出現する語に『分類語彙表－増補改訂版』の意味番号を振りました。文章のタイプによりますが 81.0 ～ 93.8% の正解率が得られています*。

以上のように、大規模言語モデルの発展により、自然言語処理技術は一定の円熟を見せ、応用的にも学術的にも非常に有用であることが明らかになり、研究者がモデルを自作する時代は終わりを迎えつつあります。筆者は、今こそ日本語学や言語学などの文系学問と、大規模なデータに立脚する自然言語処理とが、共に発展していくときだと考えています。これからの自然言語処理の発展に大いに期待しています。

＊　浅田宗磨 · 古宮嘉那子 · 浅原正幸（2024）「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号悉皆付与」『言語処理学会第 30 回年次大会発表論文集』

インタビュー

新しいことを！

前川喜久雄　国立国語研究所 所長

――2023年4月、所長に就任されました。まず、これまでどのような研究をされてきたのかを教えてください。

専門は、音声学と言語資源です。音声学は、人が話しているときに、どのように音をつくり出していて、その音がどのような性格を持ち、どのように聞き取られるかを研究します。最近は、医療用のMRI（核磁気共鳴画像）装置で唇や舌や喉などが動く様子を撮影したリアルタイムMRI動画を用いた研究に力を入れています。音声学の研究は学生時代からなので40年以上になり、言語資源は20年ほど前からです。

――言語資源とは？

ことばの研究のために体系的に収集したデータのことです。研究者一人一人が自力でことばを集めて分析するのが、従来のやり方です。しかし一人の力では限りがあります。そこで、ことばを体系的かつ大規模に集め、コンピュータの力を借りて複雑な検索ができるようにしたコーパスがつくられるようになりました。コーパスは典型的な言語資源です。ただし、誰でも使えるように公開されているものでなければ言語資源とは言えない、というのが私の考えです。「日本語話し言葉コーパス」や「現代日本語書き言葉均衡コーパス」などの開発に携わり、最近では「リアルタイムMRI調音運動データベース」を公開しました。

――国語研にはどのようなミッションがあると、お考えですか。

国語研は新しいことをやっていかなければいけません。では新しいこととは何か？ 20年前は言語資源をつくって公開することが新しいことでした。それを続け、世界的にも充実した巨大なデータが得られています。今やらなければいけない新しいことは、巨大なデータを料理する方法の開拓です。巨大なデータを扱うには、情報科学の力を借りなければなりません。特に必要なのがモデリングの技術です。言語のモデル化は昔から行われていますが、頭の中で考えるものがほとんどでした。そうではなく、データに基づいてモデル化し、シミュレーションを行ってその結果をデータと照合し、モデルの正しさを検証できる。そういうモデリングの技術がないと、量に負けてデータに振り回されてしまいます。巨大なデータに対応できる情報科学と結び付いた、新しい言語学をつくらなければいけないのです。

――そのために必要なことは？

アメリカでは言語学とコンピュータサイエンス両方の学位を持つ人がいて、言語学研究を引っ張っています。日本ではまだ少数です。国語研は2023年春から総合研究大学院大学に日本語言語科学コースを開設し、日本語をデータに基づいて客観的・定量的に分析できる次世代の研究者の養成を目指しています。

――国語研が直面している課題はありますか。その解決策は？

日本の財政が厳しく研究所の運営が難しい中、いかに生き残っていくかを考えなければいけません。それには、研究者に限らず、ことばを使う人々にとって、なくてはならない研究所であることが不可欠です。皆さんは、スマートフォンなどで音声入力を使っているでしょう。音声認識の研究が始まったのは1950年代ですが、使い物にならないとずっと言われていました。そうした中、音声のデータベースに基づいて新しい音声処理・認識の技術を開発することを目指してつくられたのが「日本語話し言葉コーパス」であり、それによって認識精度が向上したのです。また国語研が開発した一連のコーパスを検索できる専門家向けの「中納言」というウェブサイトは、4万人以上のユーザーに年200万回以上利用されています。一般向けに「現代日本語書き言葉均衡コーパス」のデータを公開している「少納言」も、年60万回以上利用していただいています。

私たちの研究が皆さんの生活に直接利便性をもたらすことはなかなか難しいのですが、どういう研究をしているかを伝え、またコーパスを気軽に使っていただくことで、まずは国語研を知っていただきたいと思っています。また、研究の下支えとなる言語資源などの整備を継続することで、なくてはならない研究所だと、研究者から評価してもらえることを目指します。

――所長になることが決まったとき、どう思われましたか。

実は、組織のマネジメントには大学生のときから興味があり、本を読んだりしていました。国語研で室長になったときには、研究室運営のノウハウを教えてもらおうと、国内の研究室をいくつか訪問しました。一番大事なことは何かと尋ねると、答えはみんな同じで、人だと。私も同意します。マネジメントについてこっそり勉強してきたことや、音声学研究や言語資源の開発で実験的な手法や情報科学を取り入れてきた経験と知見を、活かしていければと思っています。

インタビュー

今も昔も「集めて、比べる」

松本 曜　国立国語研究所 副所長

――どのような研究をしているのか教えてください。

専門は語彙の意味論で、日本語の単語の意味と文法的な性質を研究しています。人やものが移動する動画を示してその状況をどのように表現するか、20 ほどの言語について調べるプロジェクトを進行中です。日本語、英語などのほか、ネパールのネワール語やエチオピアのシダーマ語など一般にはほとんど知られていない言語も含まれます。それぞれの言語でどのように表現するかを比較することで、日本語の特徴が分かってきます。

単語ではなく動画を使って出来事から出発するというアプローチが特徴です。出来事から出発すると、この言語にはこういう表現がない、と気付くことができます。単語から出発すると、その言語にない単語は最初から除外されて、研究できません。動画を使う研究は増えてきていますが、移動事象について 20 もの言語を対象にしたプロジェクトは世界最大級でしょう。

――日本語にはどういう特徴があるのですか。

誰かが階段を駆け上がった場合、こちらに駆け上がって来たのか、それ以外の方向に駆け上がって行ったのか、習慣的には区別しないで表現する言語もあります。日本語は「来た」「行った」と区別して表現したがる言語だといわれてきました。でも、そのように区別して表現したがる傾向が日本語より強い言語もあることが分かってきました。言語には、それぞれ違いがあります。ですが私は、一見違うのに実はよく似ている、というところに関心があり、違いの中に共通性を見つけようとしています。

――言語研究の道に進んだきっかけは？

大学受験直前まで理系志望でした。しかしこのままの成績では志望校には受からないとなり、文系に転向しました。英語の成績は良かったので英語の配点が高い大学を受験し、英語学科に進んだのです。高校時代、英和辞典を読むのが好きでした。この単語とこの単語は意味がどう違うのかな、この辞書ではこう書いてあるけれどもこっちの辞書ではこう書いてあるな、と休み時間に見比べていました。それが本業になりました。

――英語学のご出身なのですね。

私に限らず、英語学から日本語の研究に入る人は多いですよ。自分の中では、英語の研究と日本語の研究は分かれていません。日本語を研究してはいるけれども日本語だけを見ているのではないというのが、私の研究の特徴と言えるかもしれません。諸言語との比較の中で日本語を見ています。

――2022年4月から副所長になられました。

国際連携が担当の一つです。国語研では以前から国際出版に力を入れており、ドイツの De Gruyter Mouton 社から出版されている全12巻のハンドブックシリーズが 2024年中に完成する予定です。日本語で書かれることの多かった日本語研究の成果を、海外でも読みやすい英語で出版することは、日本語研究の普及と進展のために重要です。前川喜久雄所長は、海外の研究所との連携を進める計画も立てています。研究者間で共同研究はしていますが、研究所全体として連携することで初めてできる研究もありますから、実現に向けて取り組んでいきます。

――2023年4月に開設された総合研究大学院大学日本語言語科学コースのコース長もされています。このコースの特徴は？

学際性と先端性です。国語研には多様なバックグラウンドを持つ人が集まって幅広い分野の研究が行われています。また国語研は、日本語研究をリードする役割を果たしてきており、実験系の研究にも力を入れています。そうした国語研の研究者たちが指導する日本語言語科学コースから、次世代の言語研究を担う人材、特に次のコーパスの設計を担う人、実験系の研究をリードしていく人が出てきてほしいと思っています。

――さまざまな言語に接する中で、好きな言語はありますか。

新しい言語ですね。初めて接する言語の話者に会うと、「これをどう表現しますか？」と聞くんです。面白くて、もっと知りたくなります。世界に言語は 5,000 以上ありますから、尽きません。

――趣味は？

バードウォッチングです。双眼鏡をのぞき、「これは何という仲間の何という鳥だ」と判断していくのは、楽しいですよ。研究では、「この動詞のこの用法はこの意味の用法だ」とカテゴリゼーションをします。鳥は飛んでいってしまうのですぐ判断しないといけないけれども、動詞の研究はじっと考えていられます。そういう違いはありますが、やっていることは同じです。そういえば、子どものころは昆虫採集が好きでした。昆虫採集もバードウォッチングも言語研究も、集めて、比べる。今も昔も、研究でも趣味でも、同じことをやっていますね。

特集

現代日本語書き言葉均衡コーパス

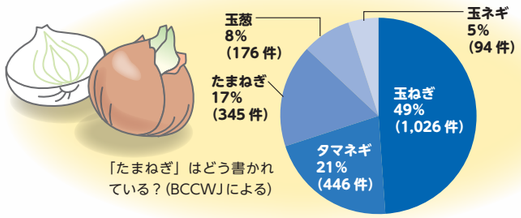
BCCWJ開発秘話

現代日本語の書き言葉の全体像を把握する――そのために国立国語研究所（国語研）が構築し、2011年より公開したのが、「現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese：BCCWJ）」です。約1億語からなる、日本語書き言葉の全体をバランスよく反映した2024年現在唯一のコーパス。その開発は試行と挑戦の連続でした。

コーパスとは？

コーパスとは、実際に使われた言葉を大量かつ体系的に集め、品詞情報など研究用の情報を付加してさまざまな検索ができるようにした、言葉のデータベースです。

野菜の「たまねぎ」は、どう書かれている？「○○的」という言葉には、どのようなものがある？　コーパスを使うと、そうした言葉の使われ方を調べることができます。



「たまねぎ」はどう書かれている？（BCCWJによる）

	出現数トップ 10	出現数 10 の例	出現数 1 の例
1	具体的（10,547）	アイドル的	BGM 的
2	基本的（9,965）	スター的	アウトドア的
3	積極的（7,585）	タイミング的	スーパーマン的
4	一般的（6,395）	あたし的	あたくし的
5	社会的（5,988）	味的	当たり前的
6	比較的（4,141）	色彩的	あら探し的
7	経済的（4,049）	観光的	おぬし的
8	個人的（3,416）	推論的	漢字的
9	総合的（3,343）	立場的	惣菜的
10	効果的（3,297）	実感的	俗人的

「○○的」という言葉にはどのようなものがある？（BCCWJ の元データによる「／○○／的」の検索結果）

世界各地でコーパスが次々に誕生

言葉を研究するには、言葉を集めて分析する必要があります。研究者はそれぞれの目的や関心に従って言葉を集めますが、個人では集めることができる量や範囲は限られます。そこで、多様な目的に利用できて言語研究の共通基盤となるように、言葉を大量かつ体系的に集めたコーパスを、大学や国が中心となってつくり始めました。

世界初の大規模なコーパスは、イギリスの「Survey of English Usage Corpus（SEU）」です。1959年から話し言葉と書き言葉それぞれ50万語の収集を始め、紙のカードに言葉を書き取って整理する方法でつくられました。

1964年にはアメリカで、100万語の書き言葉を収集した「Brown Corpus」が完成しました。コンピュータで使えるようにした、世界初の電子コーパスです。その後、世界各地でコーパスが開発され、1994年にはイギリスで話し言葉と書き言葉を合わせて1億語を集めた「British National Corpus（BNC）」が完成しました。

1億語の日本語書き言葉コーパスをつくろう！

文庫本1,700冊分に相当

国語研では1950年代から、雑誌や新聞などを対象に、言葉の使われ方や頻度を調べる語彙調査を行ってきました。それらは紙のカードを用いたものでした。2004年には国語研初の電子コーパスとして「日本語話し言葉コーパス」が完成しました。しかし、日本語の書き言葉の全体をバランスよく反映したコーパスが、まだありませんでした。

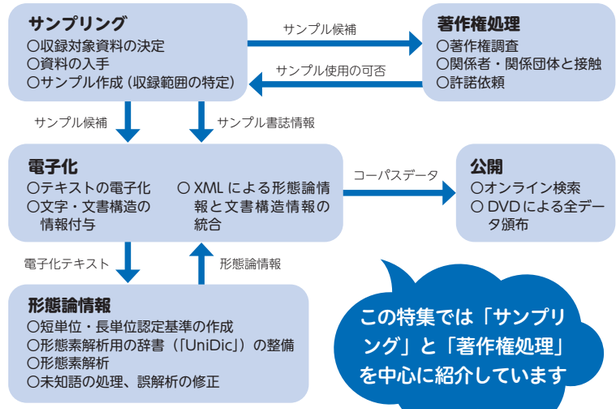
そこで、BNCの日本語版を目指し、「1億語の書き言葉コーパスをつくろう！」というプロジェクトが文部科学省科学研究費補助金（特定領域）の助成を受けて2006年に始まりました。

目指したのは現代日本語の書き言葉の縮図

新規にコーパスをつくる場合、どういう性質のコーパスとするのか、その設計方針が重要です。多くの議論を経て、次の4点を念頭に置いて設計することにしました。

①多様な書き言葉をバランスよく反映したコーパス
②幅広い目的に供するコーパス
③公開可能なコーパス
④先に公開した「日本語話し言葉コーパス」の解析単位との互換性を保持

コーパス構築の流れ



この特集では「サンプリング」と「著作権処理」を中心に紹介しています

1億語をどう選ぶ？

書き言葉にはさまざまなものがありますが、活字になって刊行されたものを対象とすることにしました。具体的には、書籍・雑誌・新聞です。では、どのようにサンプリングすれば、日本語の書き言葉を代表するサンプルを収集できるのか。それが大きな課題でした。

そして導き出した方法が、「2001～2005年に出版された全ての書籍・雑誌・新聞に含まれる総文字数を推計し、その比率をもとに各媒体からサンプリングする言語量を設定する」というものです。推計する対象として、書籍は国立国会図書館の蔵書目録にあるもの、雑誌は日本雑誌協会の加盟出版社発行のもの、新聞は全国紙・ブロック紙・有力地方紙としました。

書籍　31万7117冊・7491万1520ページ
雑誌　5万5779冊・1041万4955ページ
新聞　4万9625冊・119万8189ページ
※ 2001～2005年に出版された数

均衡とは……
テキストのジャンルや収録する際の手続きに「偏りがない」という意味。価値判断を伴うような代表性や、全てのジャンルについて同じ量だけ収録されているという意味ではない。

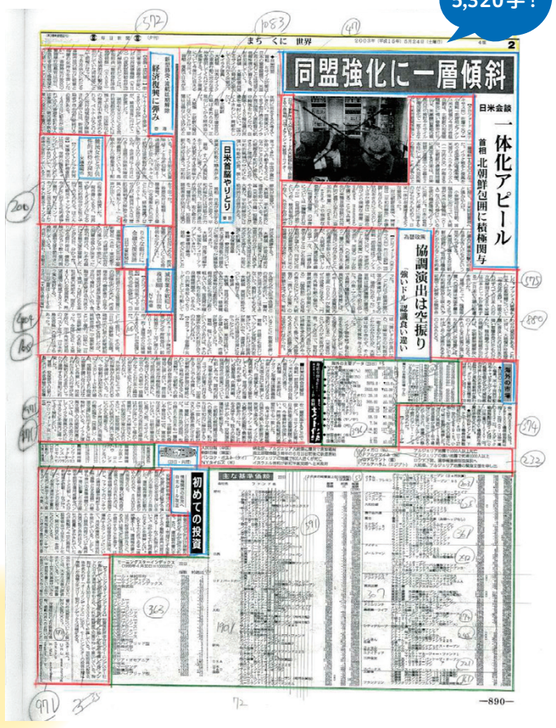
文字数をひたすら数える！

5年間に出版された書籍・雑誌・新聞に含まれる総文字数をどのように推計したと思いますか？

新聞については、全国紙4紙の朝夕刊の8冊・211ページを実測しました。右の写真は、実際に文字数を数えた新聞です。エリアを区切り、1文字1文字、ひたすら数えていきます。エリアごとの文字数が鉛筆で書き込まれています。このページの文字数は5,320字でした。

書籍については、日本十進分類法（NDC）の分類・判型ごとにランダムに選び出した227冊・1,135ページの文字数を、やはりひたすら数えました。雑誌は、NDCの分類・判型ごとにランダムに選び出した53冊・265ページを実測しました。

5,320字！



文字数を実測した新聞（毎日新聞2003年5月24日夕刊）

書き言葉の実態を把握できているのか？

5年間に出版された総文字数の推計をもとに、書籍・雑誌・新聞それぞれの種類・分類の構成比を算出しました（右上の表）。しかしこの比率でサンプリングする方法では、書き言葉が生み出された実態を把握することはできても、その言葉がどれくらい流通しているかは分かりません。また書き言葉には、書籍・雑誌・新聞以外にも多くの種類があります。

そこで、書籍については、読まれているという実態を反映するよう、図書館の蔵書やベストセラーを加えました。その他のさまざまな資料も加え、性質の異なる３つのサブコーパスで構成することにしました（右図）。

	層	推計文字数	比率		層	推計文字数	比率
書籍	0. 総記	1,636,414,548	2.50%	雑誌	総合	7,421,447,806	11.34%
	1. 哲学	2,597,610,813	3.97%		教育	877,875,592	1.34%
	2. 歴史	4,301,204,340	6.57%		政治	456,459,405	0.70%
	3. 社会科学	12,408,321,943	18.95%		産業	110,640,958	0.17%
	4. 自然科学	5,069,594,034	7.74%		工業	1,468,293,360	2.24%
	5. 工学	4,615,929,967	7.05%		医療	180,964,513	0.28%
	6. 産業	2,196,387,437	3.35%				16.06%
	7. 芸術	3,258,432,447	4.98%	新聞	全国紙	2,417,622,461	3.69%
	8. 言語	888,800,128	1.36%		ブロック紙	1,296,592,154	1.98%
	9. 文学	9,341,275,486	14.27%		地方紙	2,701,855,499	4.13%
	n. 分類なし	2,225,954,208	3.40%				9.80%
			74.14%				

現代日本語書き言葉均衡コーパス（BCCWJ）

出版サブコーパス
生み出された書き言葉を幅広く収集
書籍・雑誌・新聞（2001～2005年）
約3,500万語

図書館サブコーパス
流通している書き言葉を幅広く収集
東京都内13自治体以上の公立図書館で共通して所蔵しているもの（1986～2005年）
約3,000万語

特定目的サブコーパス
上の２つでは捉えにくいもので、現代日本語書き言葉の実態を把握するのに重要な資料を収集
白書・教科書・広報紙・ベストセラー・Yahoo!知恵袋・Yahoo!ブログ・韻文・法律・国会会議録（1976～2005年）
約3,500万語

原本を入手するため図書館・書店・古書店を巡る

古書店巡りの達人！？
図書館や新刊書店で入手できない資料は、古書店を次から次へと巡って探しました。*

＊BCCWJの構築に携わったスタッフのコメントを紹介します。

構成が決まったら、サンプルを抽出する対象をコンピュータでランダムに選び、原本を入手します。図書館で借りたり、比較的安価なものは購入したりしました。館外への持ち出しが禁止されているものは、許可をいただいてスキャナを持ち込みスキャンしました。新聞は縮刷版を入手し、縮刷版がない場合は国会図書館のマイクロフィッシュから取得しました。

入手した原本は、出版サブコーパスと図書館サブコーパスの書籍と雑誌だけでも3万1000冊以上！　図書館の貸出申請・借り出し・返却、書店への発注・納品確認など、大がかりで複雑な作業でした。

どの部分をサンプリングする？

資料の原本やスキャンデータを入手できたら、サンプリングを行います。資料ごとに、何ページのどの位置からサンプリングするか、ランダムに指定します。

１つの資料から、長さの異なる２種類のサンプルを取得します。「固定長サンプル」は基準点を始点として1,000文字目までの範囲を、「可変長サンプル」は基準点を含む章や節などの言語的なまとまり（ただし上限１万字）を抽出します。

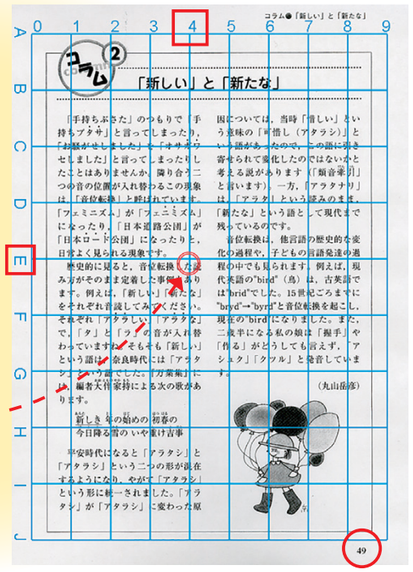
サンプル取得数
1　　SampleID：PB58_00482

NDLsize
21cm

優先順位	対象頁	有効	乱数1	乱数2
一位	49		4E	2H
二位	25		3C	6A
三位	78		4J	0H
四位	20		9J	2B

サンプル台帳の指示
「49ページの座標 4E」

マス目を印刷した透明なシートを対象ページに当て、座標により指定された交点に最も近接している文字を「サンプル抽出基準点」とする。この例では「た」。これを「キメマスシートでキメ点を決める」と言っていた。シートは判型に合わせ、さまざまなサイズを用意していた。



サンプル台帳とサンプル抽出基準点の例

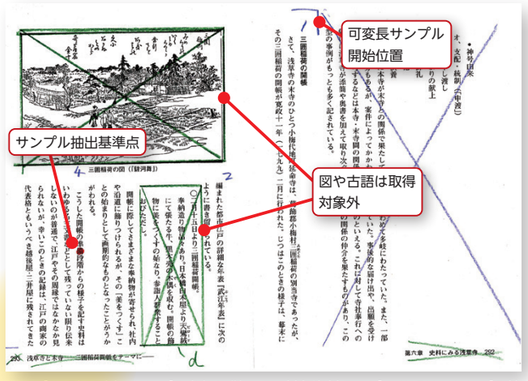
作業者泣かせの“改行なし”
可変長サンプルは、セリフ以外に改行がない文は、上限の１万字まで抽出することになっています。私が特に泣かされたのは、ロシア文学です。改行なしがずーっと続くんです。トルストイも、ドストエフスキーも。翻訳文も現代日本語として収録対象で、その翻訳文が素晴らしい文章なのですが、読みふけってしまっては文字数をカウントできません。ぐっとこらえて電卓をたたいていました。

この場合はどうする？の連続

サンプル抽出基準点はランダムに指定されるため、サンプル抽出基準点に文字がない場合もあります。その場合は、サンプル抽出基準点をサンプル台帳の乱数２、３……と変えていきます。それでも抽出できない場合は、対象ページの優先順位を下げていきます。

では、古語や外国語の文章だったら？　数式の場合は？　写真の場合は？　さまざまなケースが発生します。「現代日本語で書かれた表現」が対象ですから、ひとかたまりの古語、外国語、数式や、写真、図は、原則取得対象外です。

サンプル間・作業者間で揺れが生じないように、作業規則と判断基準を細かく設定し、マニュアル化しました。そうすることで、斉一な手順で均質なサンプリングを実施しました。



サンプル取得の例。サンプル台帳に従って、対象資料のコピーを取り、サンプル抽出基準点、固定長サンプルの範囲、可変長サンプル抽出開始位置、取得対象外などを色鉛筆で記入していく。この指示用紙に基づいて電子化作業が行われる。

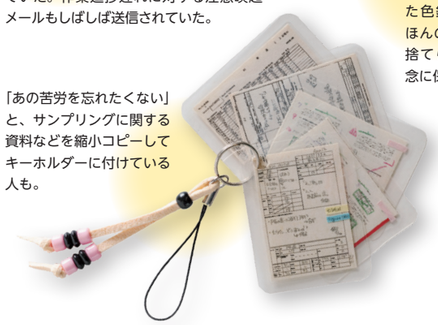
ランダムに指定される資料の誌面はさまざま。作業規則と判断基準に沿って１件ずつ手作業でサンプルを取得していく。

真夏のコピー地獄！
暑い時期の大量コピーは、かなりの重労働でした。というのも、当時の国語研の建物の窓には、ブラインドがほとんど付いていなかったのです。しかもコピー機は窓際にありました。真夏の昼の時間帯は強い日差しが照りつけ、「そろそろ限界……」「いや、ここまで終わらせたい」という気持ちのせめぎ合いでした。今、窓という窓にブラインドが付けられているのは、この作業がきっかけでは？と勝手に考えています。

国会図書館書庫内でのサンプリング作業

To Doリストや注意事項が書かれた紙。情報共有のためこうしたメモが、サンプリングの作業部屋の掲示板にたくさん貼られていた。作業進捗遅れに対する注意喚起メールもしばしば送信されていた。

サンプル取得に使った色鉛筆。これでもほんの一部。「とても捨てられない」と記念に保管している。



「あの苦労を忘れたくない」と、サンプリングに関する資料などを縮小コピーしてキーホルダーに付けている人も。

サンプリングの注意点が書かれた「魂の付箋セット」。「捨てるのがしのびない。後世まで遺していきましょう!!」とある。

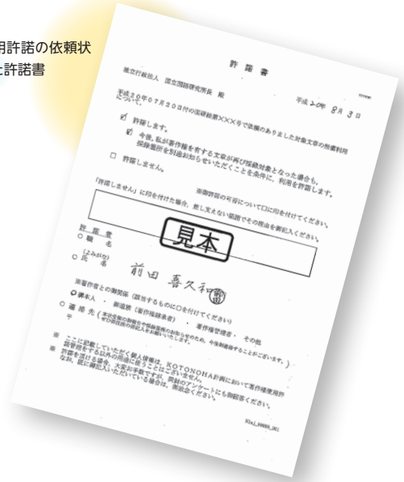
サンプリングに使用した購入資料は現在も保管されている。

著作権者から許諾をいただく

構築したコーパスを公開することは、重要な設計方針の1つでした。そのためには、収録するサンプルの著作権者に利用許諾を得る必要があります。この著作権処理に、とても苦労しました。

著作権処理が必要なサンプルは、書籍のみで2万4421件ありました。それを1件1件、地道に処理していきました。まず、サンプルの著作権者を特定します。そして連絡先を調べ、利用許諾の依頼状を送ります。連絡先が分からず、依頼状を送ることができない場合もあります。特に雑誌の場合、連絡ができない割合が書籍よりも高くなりました。

それでも根気強く連絡先を調べ、書籍については約90%に当たる2万1986サンプルの著作権者に連絡を取ることができました。



著作権者に利用許諾の依頼状と共に送付した許諾書

日本へのお見舞い

アメリカ在住の原著者に、最初に許諾をお願いしたときには断られてしまいました。ですが東日本大震災の後、日本へのお見舞いですと言って許諾をもらうことができました。

ようやく連絡が取れた作家さんは……

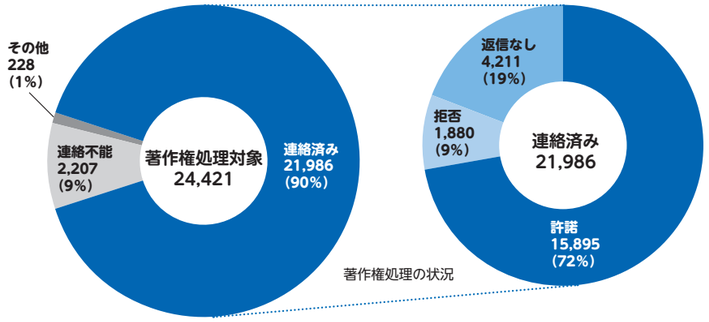
私が参加したのはプロジェクトの後期だったため、「連絡先不明」に分類された作家さんがたくさんいて、一人一人リサーチをやり直しました。そうした中、ある中国の作家さんと、幸運にも直接メールでコンタクトを取ることができました。とても丁寧なお返事と許諾もいただけて、やれやれ良かったと思っていたら……。しばらくして、ノーベル賞のニュースにその作家さんのお名前が！ びっくりしました。2012年のノーベル文学賞を受賞した莫言氏です。

利用拒否になると……

とてもありがたいことに、多くのサンプルの著作権者から利用の許諾をいただけました。しかし、利用拒否の回答が来る場合もあります。

サンプリングと著作権処理は同時に並行して行っていました。利用拒否の回答が来ると、すでにサンプリングが終わっていても、そのサンプルは利用できません。対象をランダムに選んだリストから次に優先順位が高いものの原本を入手し、サンプリングをやり直さなければならず、作業計画の見直しが必要になる場合もありました。

例えば、書籍については、3つのサブコーパス合わせて1,681冊が利用不可になりました。雑誌の著作権を有する出版社に利用を拒否された例などもありました。

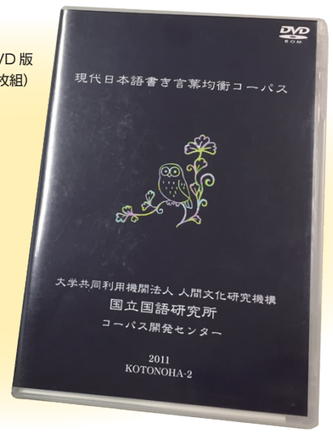


著作権処理の状況

1億語を超える日本語書き言葉のコーパス完成！

著作権処理が済んだサンプルは、電子テキスト化します。その電子テキストを用いて形態素解析を行い、品詞など研究に必要な情報を付与します。この特集では詳細を紹介できませんが、**電子化や形態論情報付与も、試行と挑戦の連続でした。**そして、全てのデータをパッケージ化。

こうして2006年から5年の歳月をかけて構築を進めてきた「現代日本語書き言葉均衡コーパス（BCCWJ）」が完成し、2011年より全てのデータを収録したDVD版の頒布（有償）とオンライン公開を開始しました。現在、Web上では、オンライン検索ツール「小納言」「中納言」「NINJAL-LWP」にて利用可能です。

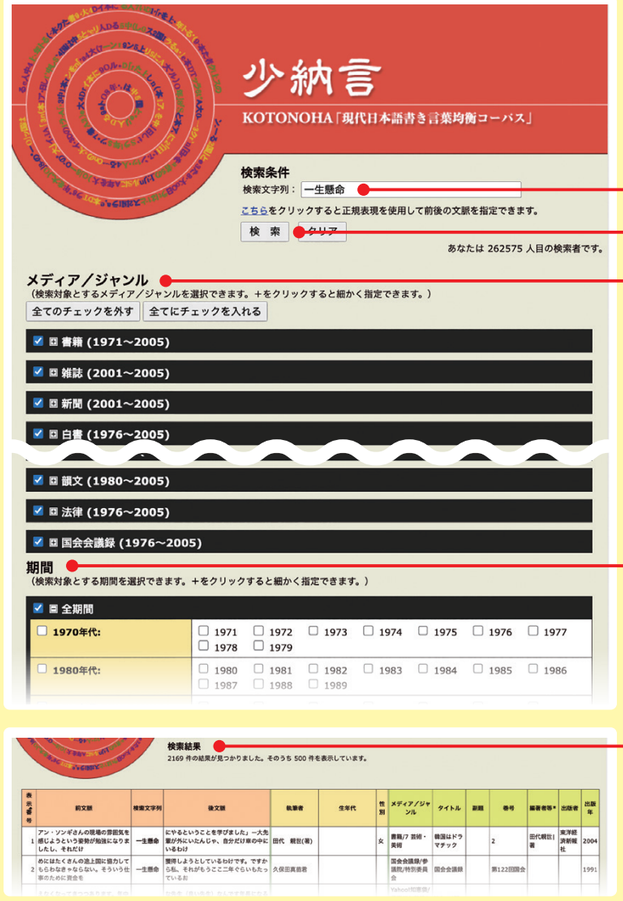


BCCWJ-DVD版
(DVD-R 4枚組)

BCCWJを使ってみませんか

国語研では、「少納言」というオンライン検索ツールを公開しています。利用条件に同意すれば、誰でも無料でBCCWJの全文検索をすることができます。「少納言」の基本的な使い方を紹介しましょう。

https://shonagon.ninjal.ac.jp/



①**「検索文字列」の窓に調べたい語句を入力します。**例えば、「一生懸命」を検索してみます。
入力した文字列がそのまま検索されるので、その他の表記形を検索したい場合は、入力文字を変えて検索します。例えば、「一生けん命」「一生けんめい」「いっしょうけんめい」「一所懸命」「一所けんめい」「いっしょけんめい」などと入力します。

②**目的に応じて、検索対象の「メディア／ジャンル」や、刊行年代の「期間」を指定することができます。**

③**「検索」をクリックします。**
前後の文脈を指定して検索したい場合は、検索ボタンの上の「こちら」をクリックして文字列を入力します。

④**検索結果が表示されます。**検索文字列の前文脈と後文脈それぞれ40字程度と、出典情報を確認することができます。検索結果のダウンロードはできません。
「一生懸命」は2,169件見つかりました。検索結果が500件以上になった場合は、その中からランダムに選んだ500件が表示されます。
検索対象の「メディア／ジャンル」や「期間」を指定することで、検索結果を絞って表示させることが可能です。「一生懸命」の場合、「メディア／ジャンル」で「雑誌」だけにすれば検索結果は76件となり、全ての用例を確認できます。

BCCWJの拡充が計画されています

BCCWJは、日本語研究をはじめ、日本語教育や国語教育、国語政策、辞書編集、自然言語処理など、さまざまな用途で使われています。BCCWJを用いた研究業績は2023年7月時点で1,784件でした。「少納言」は年60万回以上利用されています。

BCCWJの完成から10年以上たちました。BCCWJは、2011年の公開以降、更新がされていません。言葉は変化するものであり、またIT化により言葉が変化する速度は増していると言われることもあります。媒体の種類や文字数の比率も変わっていることでしょう。現代日本語の書き言葉の全体像を把握するためには、定期的なデータの追加・更新が必要です。

文化庁の施策の1つとして、令和6～10年度（2024～2028年度）にBCCWJを拡充する計画があります。2006年から2025年までの20年分の日本語書き言葉のデータを追加し、現在の1億語規模から2億語規模になる予定です。

書籍紹介

現代日本語における句読点の研究

研究概観と使用傾向の定量的分析

岩崎拓也
ココ出版
2023 年 2 月

本書は、日本語の句読点について、これまでの研究と使用実態を明らかにすることを目的としてまとめられた一冊です。句読点研究については、日本語学・国語教育・日本語教育における研究を網羅的にレビューして、その変遷をまとめています。句読点の使用実態については、日本語母語話者と日本語学習者の 2 つの側面から検討しています。日本語母語話者の句読点の打ち方については、接続詞直後の読点の使用や、括弧と句点の組み合わせの特徴などを分析・考察しています。また、日本語学習者の句読点使用については、句読点の多寡や助詞直後の読点の打ち方、日本語の習熟度と句読点使用の関係、母語による影響などを、定量的な分析と考察から明らかにしています。本書は、日本語の句読点についての研究を志す人に必携の一冊となっています。

▶ 岩崎拓也（筑波大学）

現代日本語における否定的評価を表すとりたて詞の研究

井戸美里
くろしお出版
2023 年 2 月

本書は、日本語の否定的評価を表すとりたて詞の意味的、形態・統語的現象について、記述的な一般化を行ったものです。日本語には、「洋子は電車で化粧ナンカしている」のナンカのように、話者の否定的な評価を「含み」として表示する働きを持つ、とりたて詞と呼ばれる特徴的な語群が存在します。先行研究では、とりたて詞は主となる命題と「含み」となる命題の論理的関係を表すものと考えられ、話者の評価のような主観的要素とは切り離して研究されてきました。一方本書では、むしろ話者の評価のような主観的要素こそが本質的に、とりたて詞の言語現象に意味的、形態・統語的に関与していることを、多様な現象記述を通して示しています。

▶ 井戸美里

方言地理学の視界

小林 隆、大西拓一郎、篠崎晃一〈編〉
勉誠出版
2023 年 5 月

方言が空間的な言語のバリエーションである以上、方言の研究では地理的視点を欠かすことができません。同時に、方言は言語として、語彙、文法、言語行動などの諸側面を持ち、それぞれが空間上での多様性を有します。また、日本の方言学は、方言地図を中心に地理的視点を追究するための資料を多くつくってきました。それらの諸分野で、先駆的かつ多くの優れた業績を残した、国立国語研究所名誉所員の佐藤亮一氏が、2020年11月に亡くなりました。方言の将来まで見越した佐藤氏による多方面の研究をたどりつつ、考え方の基盤を検証することを通して、方言地理学の進むべき方向を照らす論集を 24 名の気鋭の執筆陣により編みました。佐藤氏の略歴ならびに著作一覧も掲載しています。

▶ 大西拓一郎

日本語研究者がやさしく教える「きちんと伝わる」文章の授業

石黒 圭〈編〉、井伊菜穂子、市江 愛、井上雄太、本多由美子〈著〉
日本実業出版社
2023 年 2 月

本書は「文章を書くのが苦手」「もっと上手に文章を書きたい」と思う人に向けた本です。アイデアを文章にするのは難しいものです。本書は、1章「考えを形にする」、2章「変化をつける」、3章「中身を改める」、4章「相手を思いやる」の 4 つの章から成り、頭で考えたことを文章にするまでのプロセスが示されています。本書の大きな特徴は、考えたことを文章の形にするだけでなく、内容を吟味し質を高め、読み手への配慮を加えて「きちんと伝わる」文章にするまでの流れが理解できることです。また、書くことにおいて読み手の存在は欠かせません。本書では読み手との関係性によって内容や書き方も変わることが例をもって示されています。本文が先生と生徒の会話形式で始まるなど、読みやすさも意識されています。文章指導を行う教師にも参考になる一冊です。

▶ 本多由美子

こちらも注目

コミュ力は「副詞」で決まる
石黒 圭
光文社
2023 年 4 月

国立国語研究所の研究者が関わった書籍を紹介します（2023年1～6月発行）

書籍紹介

ていねいな文章大全

日本語の「伝わらない」を解決する108のヒント

石黒 圭

ダイヤモンド社

2023年9月

文章を書くのは得意ですか？　そう問われて、得意ですと即答できる人は限られます。しかし、私たちの仕事は言葉でできています。学生であっても、論文・レポートなど文章を書く機会は少なくなく、職場でも学校でも、私たちは書いた文章で評価されます。字が汚い人がきれいな字を書くことは難しいのですが、字をていねいに読みやすく書くことならできます。同様に、文章が苦手な人が名文を書くことは難しいのですが、文章をていねいに読みやすく書くことならできます。本書は、日本語研究者の立場から、文章をていねいに書くときの勘どころをまとめたものです。ご自身の文章の確認・校正に、ぜひご活用ください。

▶ 石黒 圭

一般言語学から見た 日本語の語形成と音韻構造

窪薗晴夫

くろしお出版

2023年10月

本書は、1995年に出版した『語形成と音韻構造』（くろしお出版）の続編として、過去30余年間の私の研究を、国内外における研究動向を踏まえてまとめたものです。音韻現象の普遍性や記述の一般性が求められる中、本書は東京方言の複合語アクセント（第2章）、「紅白歌合戦」のような音韻的に一語化しない複合語（第3章）、「NHK」などのアルファベット頭文字語（第4章）、「インスタ」（グラム）などの短縮語（第5章）、「ゴジラ」などの混成語やポケモンの命名（第6章）、「マンマ」などの赤ちゃん言葉や「サイヤ人」などの逆さ言葉（第7章）といったさまざまな語形成過程を、一般言語学の視点から考察しました。

▶ 窪薗晴夫

はちうえのハイビスカス（左） 洋服のおしゃべり（右）

松村雪枝〈文〉、横山晶子〈言語監修・解説〉、Hara Alina〈英語翻訳〉

言語復興の港、奈良芸術短期大学

2023年10月

小島光貴〈絵〉　田中茉央〈絵〉

奄美群島沖永良部島の作家、松村雪枝さんの連作物語「てぃんがま」から2編を、奈良芸術短期大学の小島光貴さん、田中茉央さんのイラストで絵本化しました。しまむに（沖永良部方言）・日本語・英語の3言語で、巻末に方言の解説が付いています。『はちうえのハイビスカス』は死期における生き方を、『洋服のおしゃべり』は意思を持ち表現する力を描いた作品です。絵本を通じてしまむにに興味を持つ人が増え、危機言語を継承する一助になることを願っています。

▶ 横山晶子

方言はなぜ存在するのか

ことばの変化と地理空間

大西拓一郎

大修館書店

2023年11月

方言という、ことばの地理的異なりを理解するためには、違いを生み出すもととなる言語変化を解明することが必要です。語彙の基本変化である「類音牽引」「同音衝突」「民間語源」「混淆」を「有縁化」という概念で統一的に説明しました。このような語彙が示す方言分布と、文法の分布は、特性が異なります。そのことを地図とグラフを使って明確化し、「方言分布の基本法則」を導き出しました。以上を基盤に学史上の「方言周圏論」と「方言区画論」の対立問題への視点を示し、言語地理学の進むべき指針を明らかにしました。図版を多用することで、一般の読者の方にも理解しやすい内容になっています。

▶ 大西拓一郎

コーパスによる日本語史研究 近世編

岡部嘉幸、橋本行洋、小木曽智信〈編〉

ひつじ書房

2023年12月

本書は「日本語歴史コーパス」（CHJ）の構築と活用を行ってきた国語研「通時コーパス」プロジェクトによるシリーズの1冊です。「CHJ 江戸時代編」を活用して行われた研究と、その設計や利用方法の解説から成ります。既刊の『中古・中世編』『近代編』の間に入るもので、3冊を時代順に並べると表紙や背表紙が1つにつながるデザインになっています。この表紙のように、近世語資料は日本語の歴史をつなげて見る際に必須のものですが、従来は紙の総索引すら整備が進んでおらず、コーパスが切望されていました。今、そのコーパスと利用ガイドとなる本書が世に出たことで、日本語の通時的研究がいっそう進展することに期待します。

▶ 小木曽智信

日语学术写作与研究方法

石黒圭〈监修〉、杨秀娥、费晓东、董芸、田佳月〈编〉

外语教学与研究出版社

2023年9月

知れば知るほど好きになる ことばのひみつ

柏野和佳子〈監修〉

高橋書店

2023年11月

国立国語研究所の研究者が関わった書籍を紹介します（2023年7～12月発行）

研究室訪問

本の匂いの中で

高田智和　国立国語研究所 研究系 教授

「本に埋もれている」と聞いた研究室を訪問――書棚に収まり切らない本が、いくつものワゴンを埋め尽くし、テーブルにも積み重なっています。「留学生が訪ねてきたとき、本の匂いがする、と言っていました。自分では意識したことはないのですが、それが私の研究室の特徴なのかな」と高田智和さん。

そして受け取った名刺の1枚を見て、「つち（土）ではなく、さむらい（士）ですね」とひと言。名刺には「吉」の字が。「漢字の形を研究しているので、異体字がある漢字には自然と目が行くのです。大学時代、国語学の『漢字字體の研究』という演習で、漢字の形が時代や地域で変わっていくことを知り、面白いなと思ったのが始まりです」

対象とする意味の漢字を1字ずつ、さまざまな文献から探して集め、時代や地域ごとに整理して比較し、漢字の形がどのように変わってきたか、またその理由を探っていくのですが、「以前は漢字のデータベースづくりが大変でした」と高田さん。「古い文献の漢字はコンピュータでは表示できないものばかりでした。デジタルカメラで文献を撮影し、その画像から対象の漢字を1字ずつ切り抜いて画像としてデータベースに保存する、という作業を延々とやっていました。あまりにも大変なので、コンピュータで漢字を扱うための文字コードの拡張にも取り組むようになりました」

▼文献調査では原稿用紙と鉛筆と巻き尺（水色）が必需品

必要に迫られて仕方なく？「いいえ。大学4年生のとき、JIS コード表の拡張開発が行われていて、拡張候補の原案にないが必要そうな漢字を提案しなさい、という課題が出ました。私は『𠮟（カク）』を提案しました。『三国志』に出てくる人名ではこの漢字だけ入っていなかったのです。第4水準に採用され、謝辞に私の名前も載っています。少しは社会の役に立てたという達成感があり、興味を持ちました」

2005年、国語研へ。電子政府を支える基盤整備として戸籍などで使われている漢字を調査しコンピュータで扱うための漢字情報データベースを構築するプロジェクトに参加。その後も、文字表記と漢字情報処理の研究を続けてきました。「同じ漢字でも、手書きと印刷では形が違います。明朝体とゴシック体でも形が違います。でも私たちは、同じ漢字と見なしていますよね。形が違っていても、同じ漢字と見なすことができる許容範囲はどこまでなのか、ということにも関心があります」

高田さんの研究対象は文献だけではありません。「15年ぶりに街路の看板の調査に行きました。注目しているのは地名の表記です。地名は生きていて、画数の多い漢字をひらがなにしたり、別の漢字に置き換えたり、表記が変わります。そこに住んでいる人からも話を聞き、15年でどのように変化したか、またその理由を調べようとしています」

とはいえ、一番楽しいときについて尋ねると「実物の文献を触っているとき」と。「写真印刷の性能が良くなり複製本から分かることも増えました。でも実物を触るとその質感などから、書かれた当時のこと、字がどう書かれたか、代々の所蔵者に大事にされていたかどうかも分かります。記号としての言語ではないところも感じたいのです」

趣味は？「部屋の窓から鳥を見ることかな。夏、木の葉が茂っているときに、姿は見えないけれども『いるぞ』と感じるのが好き。実物の文献に触れながら見えないものを感じたいというのと、通じるものがあるのかもしれません」

編集後記

『ことばの波止場』vol.13の特集は「日本で話されていることば その危機とルーツ」「コンピュータと人間の言語」「BCCWJ開発秘話」の3つです。表紙の写真は、BCCWJ（現代日本語書き言葉均衡コーパス）のサンプル取得に使った色鉛筆です。私は小学生のころから、1つ目の特集にも取り上げられた日本列島のことばの多様性に興味がありました。記事の編集を担当しているときには、このような研究の面白さを読者の方々に正確に伝えることの難しさと楽しさと素晴らしさとを、いつも感じていました。（五十嵐陽介）

ことばの波止場 vol.13
2024年3月29日発行

編集　国立国語研究所広報室ことばの波止場編集部会
（柏野和佳子（部会長）、朝日祥之、五十嵐陽介、井上文子、窪田悠介、中川奈津子、福永由佳、横山晶子）

発行　大学共同利用機関法人 人間文化研究機構 国立国語研究所
〒190-8561　東京都立川市緑町10-2
Tel　0570-08-8595
https://www.ninjal.ac.jp/

編集協力　フォトンクリエイト（鈴木志乃）
撮影　Studio CAC（吉田 号）
デザイン　デザインコンビビア（山田純一）

『ことばの波止場』はウェブサイト「ことば研究館」でもご覧になれます。